ISSN 1847-3938

mipro 2013

organizer upro

36th international convention

May 20-24, 2013, Opatija - Adriatic Coast, Croatia

mipro - path to knowledge and innovation

Lampadem tradere

mipro proceedings

My profession.
My organization.
My IEEE.

Discover the benefits of IEEE membership.

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

Join today
www.ieee.org/join

IEEE

MIPRO 2013

36th International Convention

May 20 - 24, 2013
Opatija, Croatia

Proceedings

Conferences:

Microelectronics, Electronics and Electronic Technology /MEET

Distributed Computing and Visualization /DC VIS

Telecommunications & Information /CTI

Computers in Education /CE

Computers in Technical Systems /CTS

Intelligent Systems /CIS

Information Systems Security /ISS

Business Intelligence Systems /miproBIS

Digital Economy - 10th ALADIN /DE

Government, Local Government, Public Services /GLGPS

MIPRO Junior - Student Papers /SP

Edited by:
Petar Biljanović

International Program Committee

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

organized by

MIPRO Croatian Society

technical cosponsorship

IEEE Region 8

under the auspices of

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

patrons

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

sponsors

HEP - Croatian Electricity Company Zagreb, Croatia
Ericsson Nikola Tesla Zagreb, Croatia
T-Croatian Telecom Zagreb, Croatia
Končar-Electrical Industries Zagreb, Croatia
InfoDom Zagreb, Croatia
Transmitters and Communications Company Zagreb
VIPnet Zagreb, Croatia
Nokia Siemens Networks
HROTE - Croatian Energy Market Operator Zagreb, Croatia
Microsoft Croatia Zagreb, Croatia
King-ICT Zagreb, Croatia
Storm Computers Zagreb, Croatia
Supra Net Zagreb, Croatia
Altpro Zagreb, Croatia
Croz Zagreb, Croatia
Mjerne tehnologije Zagreb, Croatia
Origo Rijeka, Croatia
IN2 Zagreb, Croatia
Adnet Zagreb, Croatia
CS Computer Systems Zagreb, Croatia
ib-proCADD Ljubljana, Slovenia
Selmet Zagreb, Croatia
Vidi-TO Zagreb, Croatia
Nomen Rijeka, Croatia
Systec Automatizacija Labin, Croatia

All papers are published in their original form

For Publisher:

Petar Biljanović

Publisher:

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

Printed by:

GRAFIK, Rijeka

ISBN 978-953-233-074-8

Copyright © 2013 by MIPRO

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

CONTENTS

LIST OF PAPER REVIEWERS.....XXVII

LIST OF AUTORS.....XXIX

FOREWORD.....XXXV

MICROELECTRONICS, ELECTRONICS AND ELECTRONIC TECHNOLOGY

INVITED PAPER 3

Min-Max Circuits as a Problem Solving Tool(s) 5
D. Kovačević

PAPERS 15

Tuning Electronic Properties of Molybdenum Disulfide by a Substitution in Metal Sublattice...17
V. Fedorov, N. Naumov, A. Lavrov, M. Tarasenko, S. Artemkina, A. Romanenko, M. Medvedev

Alternating Current Electrochemical Anodisation of Silicon on Insulator Layers.....19
M. Ivanda, V. Đerek, M. Balarin, O. Gamulin, M. Kosović, S. Musić, M. Ristić, Z. Crnjak Orel

Porous Silicon Prepared from p^- and p^{++} Bulk Silicon.....24
V. Đerek, R. Baltić, M. Marcijuš, M. Ristić, Z. Crnjak Orel, M. Ivanda

Some Physical Problems in the Preparation and Analysis of the Heavily Boron and Phosphorous Doped Polysilicon Thin Films.....28
M. Očko, S. Žonja, M. Ivanda

Luminescence, Raman and Synchrotron XPS Study of Amorphous Ge_2S_3 Based Films.....34
V. Mitsa, M. Ivanda, O. Gamulin, R. Holomb, O. Kondrat, N. Popovych, G. Lovas, S. Petreckiy, N. Tsud, V. Matolín, K. Prince

Light Coupling Between the Whispering Gallery Modes of a Coated Microsphere and a Tapered Fiber.....40
D. Ristić, M. Mazzola, A. Chiappini, P. Féron, A. Rasoloniaina, G. Cibiel, M. Ivanda, G. Righini, M. Ferrari

Eu^{3+}/ Yb^{3+} - Activated Nanocrystals of Lanthanide Phosphates 44
A. Lukowiak, A. Bednarkiewicz, R. Wiglusz, W. Strek, L. Kepinski, L. Krajczyk, D. Ristić, A. Chiasera, M. Ferrari

Numerical Study and Performance Analysis of Carbone Nanotube Field Effect Transistors 49
R. Djamil, B. Salima, L. Kheireddine

New Silicon Reach-Through Avalanche Photodiodes with Enhanced Sensitivity in the DUV/UV Wavelength Range....54
D. Grubišić, A. Shah

175,000 Device-Hours Operation of AlGaN/GaN HEMTs on Diamond at 200°C Channel Temperature....61
D. Babić, Q. Diduck, C. Khandavalli, D. Francis, F. Faili, F. Ejeckam

Low Power CMOS Sub-threshold Circuits....66
B. Dokić, A. Pajkanović

Low-Power and High-Speed 4-2 Compressor....72
A. Najafi, B. Mazloom-Nezhad, A. Najafi

High-Order Inductorless Elliptic Filter with Reduced Number of Capacitors Using Signal-Flow Graphs....76
Đ. Glavinić, D. Jurišić

Engineering Education: Where We Are and Where We Go from Here....82
B. Pejčinović

Concept for a SIL3 Middleware Encapsulating Safety-related Aspects of Applications for an 8051-based SIL3 Multi-Core System-on-Chip....88
M. Schreiber, E. Delic, A. Hayek, J. Börcsök

Validation of a SIL3 Middleware for Safety-related System-on-Chips....92
E. Delic, M. Schreiber, A. Hayek, J. Börcsök

Weighted Least-Squares Design of Sharpened CIC Filters....98
G. Molnar, M. Glavinić Pecotić, M. Vučić

Some Improvements of the Gaussian Elimination Method for Solving Simultaneous Linear Equations....103
S. Delić, Ž. Jurić

LED Powered Identification Tag - Energy Harvesting and Power Consumption Optimization 109
G. Horvat, D. Vinko, T. Švedek

Self Powered Current Acquisition System with Wireless Data Transfer....115
V. Valchev, A. Marinov, E. Dimitrova

Investigation of an Intelligent Power Supply System for an Electric Heater Furnace with Silicon Carbide Heaters....120
G. Nikolov, B. Dimitrov, V. Valchev

Končar Smart Metering - Innovative Solutions for Well Known Problems....126
D. Cmuk, D. Cihlar, M. Šimunović

Micropositioning Mechatronics System Based on FPGA Architecture....132
E. Kamenar, S. Zelenika

On-water Measurement and Analysis System for Estimating Kayak Paddlers' Performance...138
G. Vadai, G. Makan, Z. Gingl, R. Mingesz, J. Mellár, T. Szépe, A. Csamangó

Random Interleaver Design in Turbo Coding for Digital Magnetic Recording Channel....144
S. Stancescu, C. Sandu

Interpretation of Pulmonary Function Test Results in Relation to Asthma Classification Using Integrated Software Suite **147**
A. Badnjević, D. Koruga, M. Cifrek, H. Smith, T. Bego

DISTRIBUTED COMPUTING AND VISUALIZATION

PAPERS **155**

DISTRIBUTED COMPUTING

High Productivity Processing - Engaging in Big Data around Distributed Computing **159**
M. Riedel, M. Memon, A. Memon, G. Fiameni, C. Cacciari, T. Lippert

Scaling the Performance and Cost While Scaling the Load and Resources in the Cloud **165**
M. Simjanoska, S. Ristov, G. Velkoski, M. Gusev

Semidefinite optimization of High Performance Linpack on Heterogeneous Cluster **171**
D. Tomić, L. Gjenero, E. Imamagić

Distributed Virtual Test Bed: an Approach to Intergation of CAE Systems in UNICORE Grid Environment **177**
G. Radchenko, E. Hudyakova

Message Oriented Framework with Low Overhead for Efficient High-Performance Monte Carlo Simulations **183**
E. Atanassov, T. Gurov, A. Karaivanova

WRF4SG: A Scientific Gateway for the Weather Research and Forecasting Model **186**
C. Blanco, A. Cofiño, V. Fernández-Quiruelas

Generation of the Scrambled Halton Sequence Using Accelerators **191**
E. Atanassov, M. Durchova

Intel vs AMD: Matrix Multiplication Performance **196**
N. Anchev, M. Gusev, S. Ristov, B. Atanasovski

Web Service for Separating Components in the Exponential Decay Process **202**
A. Martinović, D. Arbula, Ž. Jeričević

Performance Drawbacks for Matrix Multiplication Using Set Associative Cache in GPU Devices **207**
L. Djinevski, S. Arsenovski, S. Ristov, M. Gusev

Defending the Need for a New Global Software Approach: a Literature Review **213**
M. Trkman, S. Vrhovec, D. Vavpotič, M. Krisper

Parallelization of NAPOM Implementation **217**
M. Depolli, G. Kosec, J. Ugovšek, V. Malačič

Faster Solution of Large, Over-Determined, Dense Linear Systems **222**
Ž. Jeričević, I. Kožar

Virtual Computers and Virtual Data Storage **226**
A. Šimec, O. Staničić

AdriaScience Gateway: Application Specific Gateway for Advanced Meteorological Predictions on Croatian Distributed Computing Infrastructures.....231
D. Davidović, T. Lipić, K. Skala

Handling Web Service Interfaces.....236
J. Matković, K. Fertalj

Loosely or Tightly Coupled Affinity for Matrix - Vector Multiplication.....242
G. Velkoski, S. Ristov, M. Gusev

Computation and Analysis of Sobol Coefficients for Air Pollution Concentrations over the Territory of Bulgaria.....248
E. Atanassov, S. Ivanovska

Algorithms for Computational Procedure Acceleration for Systems Differential Equations in Matlab 252
H. Nenov, B. Dimitrov, A. Marinov

Support for Data-intensive Computing with CloudMan 257
Y. Kowsar, E. Afgan

IaaS Type Cloud Infrastructure Assessment and Monitoring 263
M. Kozlovszky, M. Törőcsik, T. Schubert, V. Póserné

Energy Efficiency in Large-Scale Distributed Computing Systems 267
R. Trobec, M. Depolli, K. Skala, T. Lipić

3D Visual Analytics for Quality Control in Engineering.....272
M. Klemenčić, K. Skala

VISUALIZATION SYSTEMS

Remote Visualization of 3D Graphical Models.....279
D. Gorgan, O. Capatana

Fast Network Communities Visualization on Massively Parallel GPU Architecture 285
M. Mržek, J. Blažič

Realization of Natural Interaction over the Micro Soot Device Model 291
P. Mrazović, M. Pilipović, M. Volarević, Ž. Mihajlović

Accelerating Mean Shift Image Segmentation with IFGT on Massively Parallel GPU.....295
J. Sirotković, H. Dujmić, V. Papić

ARIS MashZone KPI Visualisation of Simulated Routing Protocol Operations for Smart Dust Environments.....302
K. Hajdarević, A. Civa

Physical Simulation of 3D Dynamical Surfaces on Graphic Clusters 308
A. Sabou, D. Gorgan

Road Line Detection based on Optical Correlator.....314
T. Harasthy, J. Turán, Ľ. Ovseník

BIOMEDICAL ENGINEERING

The Specific Approach for Establishment of Innovation Hospitals in Gorenjska Region 319
U. Stanič, S. Dolinšek, A. Cerkvenik Škafar, I. Grmek-Košnik, J. Stanič, G. Cerinšek

Multifunctional Protective Clothing System: Development vs. Functionality.....325
D. Zavec Pavlinić, A. Oder, V. Grm, U. Stanič

Functional Textiles Have Important Role on the Field of Human Life.....330
D. Zavec Pavlinić

I-TEHMED Biomedical Technology Platform - Structure and Results.....335
J. Stanič, J. Jelenc, U. Stanič

Multi-Functional Wireless Body Sensor - Analysis of Autonomy.....340
K. Bregar, V. Avbelj

Personal Health Monitoring with Android Based Mobile Devices 344
M. Kozlovszky, L. Bartalis, B. Jókai, J. Ferenczi, P. Bogdanov, Z. Meixner, L. Németh, K. Karóczkai

Signal Processing Methods for ST Variability Assessment in ECG 349
A. Rashkovska, V.Avbelj

Coarse Grain Molecular Dynamics Study of Voids Present in the Membrane of a Lipid Vesicle.....353
J. Pavlič

Low-Frequency Ultrasound *in vitro*: Experimental System and Ultrasound-Induced Changes of Cell Morphology.....358
J. Jelenc, J. Jelenc, D. Miklavčič, A. Maček Lebar

Using Hadoop MapReduce in a Multicluster Environment 363
I. Tomašić, A. Rashkovska, M. Depolli

Approaches to DNA *de novo* Assembly 369
I. Sović, K.Skala, M.Šikić

Ski Injury Triggers of Tibial Plateau Compression Fracture.....378
R. Trobec, I. Belehar, J. Polajnar, M. Veselko

A Model of Selective Stimulation and ENG Recording in the Human Left Vagus Nerve.....383
P. Pečlin, J. Rozman

TELECOMMUNICATIONS AND INFORMATION

INVITED PAPERS 389

New Challenges and Perspectives on Understanding and Managing the End User Experience: Presenting European Research Efforts on the Quality of Experience of Multimedia Systems...391
L. Skorin-Kapov

ICT i povećanje socijalne uključenosti osoba sa složenim komunikacijskim potrebama.....393
Ž. Car

PAPERS 395

SERVICE QUALITY

PSQM - Platform for Service Quality Management 399
O. Jukić, I. Heđi

A Possibility of Estimation of IPTV QoS Parameters by Using DSL Transceivers Physical Layer Parameters 404
N. Škaljo, A. Begović, N. Behlilović

An Impact of Access Network Physical Impairments on IPTV QoE in DSL Enviroment 408
N. Goran, M. Hadžialić, A. Begović

Metodologija vrednovanja kvalitete usluge e-Poziva u sklopu pilot projekta HeERO.....412
P. Britvić, M. Rimac, K. Vidović

Utjecaj kvalitete usluga i zadovoljstva korisnika na lojalnost korisnika usluga 416
H. Kuč, A. Kuč

Istraživanje zadovoljstva korisnika multidimenzionalnih usluga na tržištu telekomunikacija Bosne i Hercegovine 422
N. Karahmetović

INNOVATIONS AND BUSINESS IMPROVEMENTS

A Blend of Semi-Open Innovation & Open Sourceness: Accelerating Innovation within a Large Telecommunications Company 431
M. Bervanakis, S. Dešić

SEO in e-Commerce: Balancing between White and Black Hat Methods.....437
S. Duk, D. Bjelobrk, M. Čarapina

Izazovi razvoja odgovarajućih poslovnih procesa neophodnih za migraciju telekomunikacijskih usluga.....443
K. Musa, D. Markulin

Povezanost društveno odgovornog poslovanja i poslovnih performansi.....448
N. Pivačić, N. Kusić

WEB APPLICATIONS & M2M COMMUNICATIONS

Web Application Functional Size Estimation Based on COSMIC Method and UWE Approach.....457
D. Čeke, M. Đurek, S. Kasapović

Značajke i mogućnosti web orijentiranih radnih sučelja 465
Ž. Širanović, N. Majstorović, T. Tolić

Abstraction and Semantics Support in M2M Communications 472
V. Čačković, Ž. Popović

Machine-to-Machine Communication as Key Enabler in Smart Metering Systems 477
M. Dujak, V. Parać, M. Đurasević, A. Herić

DEVELOPMENT METHODS

Improved Distributed Band-Preference Dynamic Spectrum Management for Upstream VDSL2 485
Y. Gao, Y. Li, G. Wang, M. Jin

Dictionary Based Optimization for Adaptive Compression Techniques.....491
A. Mokhtar Mansour, M. Fouad

Adoption of Agile Methodology in Software Development 496
D. Duka

Developments and Current Trends in Ethernet Technology 501
D. Valenčić, V. Lebinac, A. Skendžić

Pregled metoda ublažavanja utjecaja endogenih faktora na neuspješnost IT projekata.....507
N. Marušić

INFRASTRUCTURE DEVELOPMENT

Terastream Implementation of All IP New Architecture 515
B. Zaluški, B. Rajtar, H. Habjanić, M. Baranek, N. Šlibar, R. Petračić, T. Sukser

Merging Dynamic Address Autoconfiguration and Security Key Protocols in Mobile Ad Hoc Networks 519
M. Shurman, M. Al-Mistarihi, K. Darabkh

MAC Layer Back-off Algorithm for Ad hoc Networks 524
M. Shurman, M. Al-Mistarihi, Z. Alomari

Heterogeneous Networks in Theory and Practice 530
T. Ivanković

Information Systems Infrastructure Planning Based on Collected Data from Monitoring Tools.....534
I. Golub, B. Radojević

REGULATORY ASPECTS

Regulatory Aspects of Mobile Broadband Technologies 541
A. Džebo, H. Mutapčić

Role of Infrastructure Atlas in NGA Deployment.....546
S. Hadžović, S. Kovačević

Trendovi i mogućnosti primjene mobilnog širokopojasnog pristupa.....549
A. Krivošić, A. Džebo

Prenosivost telefonskih brojeva 555
A. Gabela, F. Gabela, J. Gabela

WIRELESS SENSOR NETWORKS

Effective Energy Consumption on Wireless Sensor Networks: Survey and Challenges 561
A. Boonsongsrikul, S. Kocijančič, S. Suppharangsan

An Optimal Algorithm Based on Gridding and Clustering for Decrease Energy Consumption in WSN.....566
A. Firuzbakht, A. Bouyer

Hierarchical Clustering Using Genetic Algorithm in Wireless Sensor Networks 571
M. Shurman, M. Al-Mistarihi, A. Mohammad, K. Darabkh, A. Ababnah

Efficient Wireless Signal Emitting Sensor Distribution within an Environment Modified for the Needs of the Aging Population 576
V. Prčić, M. Đurek, D. Šimunić

TeleAssistant - Assisted Living Platform Applied to Fall Detection.....582
M. Čavka, R. Krpetić, D. Vasić, V. Bilas

RADIO COMMUNICATIONS SYSTEMS

Properties of Hybrid FSO/RF Link with 60 GHz RF Backup Link.....589
M. Tatarko, Ľ. Ovsénik, J. Turán

Comparison of New Technologies for Data Exchange in the Maritime HF Frequency Band 592
S. Valčić, Z. Mrak, D. Kezić

A Modular Simulation Framework for Analyzing Coexistence Issues in Wireless Radio Networks in Close Spatial Proximity.....597
R. Neumeier, G. Ostermayer

On the Performance of DoA Estimation Algorithms in Cognitive Radio Networks:- A New Approach in Spectrum Sensing.....601
T. Dhope, D. Šimunić

Design and Implementation of SDR Based QPSK Modulator on FPGA.....607
T. Kazaz, M. Kulin, M. Hadžialić

Mogućnost korištenja bežičnih komunikacija za uspostavu "mesh" podatkovne mreže sa svim objektima pomorske signalizacije na Jadranu 613
I. Karin, J. Vrlić, Z. Belak, I. Ružić

ENVIRONMENTAL PROTECTION

Analiza rezultata mjerenja elektromagnetskih polja u mobilnim mrežama Hrvatskog Telekoma 621
S. Gros, R. Gajić

Kolocirani izvori visokofrekvencijskih elektromagnetskih zračenja.....627
B. Tanatarec

Analysis Methodologies of Assessment of Impact Teleworking on the Decrease Carbon Emissions of Physical Mobility.....631
M. Berković, A. Kosovac, A. Džebo

COMPUTERS IN EDUCATION

INVITED PAPER **641**

Tablet Learning Environment in iSchool Vežica **643**
V. Nikolić, L. Zima Krnelić, D. Kolarić, M. Nikolić

PAPERS **647**

Are We Trendy? **649**
H. Jaakkola, B. Brumen, J. Henno, J. Mäkelä

Concept Maps in Computer-Assisted Knowledge Assessment **658**
J. Petrović, B. Jeren, P. Pale

Tracking the Roots of Plagiarism - Results of the Initial International Survey **663**
D. Lukač, M. Čičin-Šain

Considerations on the Evolution of Attention Paid to Education in Romania Compared to EU, in the 1998-2010 Period **667**
M. Zaharia, D. Enachescu

Synergy of e-Science and Libraries as a Support to Higher Education Institutions in Croatia **673**
R. Vrana

Modes of Use and Preservation of Digital Information Resources as a Support to the Learning Process at the Faculty of Humanities and Social Sciences in Zagreb **679**
R. Vrana

Towards a Framework for Usability Evaluation of Educational Artifacts Created with Web 2.0 Applications: A Pilot Study **685**
T. Orehovački, N. Žajdela Hrustek

Parameter Identification of Transfer Functions Using MATLAB **691**
M. Fruk, G. Vujisić, T. Špoljarić

Web Interactive Learning - Case Study: Faculty of Economics of Josip Juraj Strossmayer University in Osijek **697**
M. Lamza-Maronić, J. Glavaš

The Role of Computerisation and Automation of School Libraries Management in Teaching Information Literacy on the Example of Curriculum Unit Online Library Catalogues **701**
D. Vincek

Design and Development of School Assets Management System **707**
A. Vladimirou, A. Kokkinaki

Data Mining in Hybrid Learning: Possibility to Predict the Final Exam Result **711**
J. Gamulin, O. Gamulin, D.Kermek

Students' Attitudes to the Use of Open Source Technology in Education **717**
K. Pavlina, B. Petrović

More Efficient Training of Teachers of Informatics Using Modern Information-Communication Technology in Croatian Education and Teacher Training Agency 720
E. Ban, T. Bronzin

Student's Level of ICT Literacy by Study Groups on Faculty of Philosophy, University of Mostar 725
A. Ivanković, S. Špiranec, D. Miljko

E_Learning vs. Multimedia (Ivica & Marica - Hansel & Gretel) 730
K. Bedi, N. Žajdela Hrustek, T. Vrbanec

ICT Learning Experience in the Multilingual and Multicultural Environment of ERASMUS IP TAARLoG Project 736
J. Žufić, H. Cvilinder

Incorporating Social Network Workshops in Elementary School Curriculum 742
M. Draženović

Mathematical Modelling, Computers and GeoGebra in University and College Mathematics Education 745
T. Sekulić, Đ. Takači

Application of Gantt Charts in the Educational Process 751
A. Bednjanec, M. Filipović Tretinjak

Interaktivnost u učionicama u kontekstu novih obrazovnih tehnologija i alata 756
K. Blažeka

PC and Microcontrollers Applications in the Laboratory Exercises of the Electrical Engineering 760
O. Tasić, V. Tasić, D. Brodić, V. Despotović, M. Pavlov, D. Milivojević

On-line e-Learning Platform Supporting Education and Practice of Accountants in EU Space 765
D. Paľová, M. Vejačka

Choosing Free Library Software: Experiences of the Faculty of Humanities and Social Studies in Zagreb 771
T. Čonč

Development of Information Competencies via Process Modelling and ICT Integration 778
L. Révészová

Research Assistance Library Service - In between Formal and Informal Learning 784
B. Marijanović, S. Šimara, J. Lisek

Gaming in the Adolescent's Population - Pilot Research 789
J. Žufić, D. Kiralj

Compact Flexible Router Based Remote Experimenting System 795
T. Szepe, S. Toth, Z. Gingl

Web 2.0 Tools in Croatian Higher Education: an Overview 800
M. Čarapina, D. Bjelobrk, S. Duk

Using Microsoft Project for Project Management in Non-Governmental Organisations 805
D. Kažović, D. Valenčić

Heuristic Evaluation in the Human Computer Interaction Course 809
D. Bošković, N. Borovina

Motherboard and User Experience 813
J. Tomljanović, T. Turina, E. Krelja-Kurelović

QR Codes in Library - Does Anyone Use Them? 819
I. Jelić, D. Vrkić

Systems of Suport for Persons with Complex Communication Needs 824
Z. Stančić, J. Škrinjar, Ž. Car, I. Lovrek, V. Vlahović Štetić, J. Pibernik

Framework for Web Application Domain Knowledge Extraction 829
I. Rožanc

The Influence of Social Networks on Student's Evaluation and Results.....835
A. Skendžić, B. Kovačić, D. Valenčić

Will Universities and University Teachers Become Extinct in Our Bright Online Future?.....840
J. Henno, H. Jaakkola, J. Mäkelä, B. Brumen

Cloud Computing in Education and Student's Needs.....850
E. Krelja Kurelović, S. Rako, J. Tomljanović

Mobile Phone in the Classroom 856
Ž. Požgaj, V. Bosilj Vukšić

Future-Oriented Education with a Modern Flexible Production Unit.....861
C. Ungermanns, W. Werth

An Approach to the Evaluation of Embedded Engineering Study Programs 866
I. Kaštelan, M. Barak, V. Sruk, M. Anastassova, M. Temerinac

BYOD: a Challenge for the Future Digital Generation 872
I. Pogarčić, M. Gligora Marković, V. Davidović

ICT Didactics: The New Study Discipline is Needed.....877
M. Krašna, B. Bedrač

Computer Science Curriculum 2013 - An Example 883
V. Jovanović

Determination of the Torque-Speed Characteristic of Induction Motor in Electric Machinery Education 889
R. Buinac, V. Tomljenović

Relevance and Realization of Communication between Information and Communication Technology Employers and Educational Sector 894
M. Rotim

The Influence of Situational Factors on Accepting e-Learning Technology by University Teachers 900
S. Babić

The Next Generation of University Information Systems - TTI Version for XXI Century 906
B. Misnevs, N. Fila

Preparedness of Teachers and Learning Materials for e-Learning.....911
N. Lazić, J. Klindžić, M. Odak

The Strategic Oriented Information Science Education in the Function of Gaining Strategic Knowledge.....916
M. Čičin-Šain, S. Vukmirović, M. Čičin-Šain

E-learning Controversy in Practical Application.....922
T. Bratina, M. Duh, M. Krašna

COB: Online Educational Tool for Collaborative Ontology Building 928
M. Jovanov, M. Gusev, D. Martinovikj

Object-oriented Programming for Graphical Modeling on the Example of Logistics Network Design 934
S. Vukmirović, D. Kovač, I. Mežnarić

Calculating Duration and Convexity of Bonds Using Excel.....940
Z. Prohaska, B. Olgić Draženović, V. Šarić

PaintArt - An Interactive Computer Software for Education of Visual Arts.....944
Z. Prohaska, I. Uroda

Laboratorij za emulaciju računalnih mreža i primjena u praktičnoj nastavi.....948
A. Jakupović, E. Pirić

Otvoreni edukacijski *online p*rogram u funkciji kvalitetnije dostupnosti službenih informacija.....954
R. Pekorari, T. Horvat

Primjena računalnih igara u obrazovanju.....958
M. Mirković

Istraživanje o učestalosti korištenja računala i Interneta kod učenika Osnovne škole Centar u Puli.....964
B. Kontošić, A. Kontošić

Elektronski udžbenik u osnovnim i srednjim školama 970
J. Đurić, V. Jelkić, A. Nikitović

Istraživanje utjecaja rada za računalom na zdravlje učenika Osnovne škole Centar u Puli.....976
B. Kontošić, A. Kontošić

OpenProj - Open Source softverski alat za upravljanje IT projektima.....981
D. Valenčić, D. Kažović, A. Radovan

Zaštita i sigurnost djece na internetu 986
M. Kosovec, M. Dumančić

Mogućnosti primjene informacijsko-komunikacijske tehnologije u primarnom obrazovanju ..991
M. Lukić, M. Dumančić

Facebook u obrazovanju **995**
D. Drenjanac, M. Dumančić

Usporedba stavova dvije generacije budućih učitelja prema odgoju i obrazovanju u budućnosti **999**
V. Galešev

Primjena web sadržaja i web alata na području filozofije znanosti **1003**
M. Sertić

Information Literacy Instruction: Self-Education through e-Tutorials **1009**
A. Papić

Zaštita podataka od plagiranja **1013**
S. Šimundić, D. Barbarić

Primjer uporabe dinamičkih struktura podataka u programskom jeziku Python **1017**
P. Brođanac, L. Budin, Z. Markučič, S. Perić

Programiranje u općoj, jezičnoj i klasičnog gimnaziji **1021**
M. Menciger

Poboljšanja u primjeni računalom vođenog demonstracijskog stola u nastavi **1025**
T. Horvat, Lj. Cvitaš

Metodologija izrade *E - learning* sadržaja u svrhu osposobljavanja za polaganje ECDL certifikata **1031**
S. Grgić, A. Lacković

Primjena testova s bodovanjem zasnovanim na sigurnosti ispitanika u ispravnost odgovora **1035**
B. Nožica, Ž. Širanović, M. Čarapina

Informacijske tehnologije u funkciji ekološkog obrazovanja učenika strukovnih škola **1040**
V. Matić

Primjena e-učenja u NATO okružju (e-NATO) **1046**
D. Možnik

Sigurnost i mobilnost za sve **1052**
I. Ružić

Motivacija učenika za e-učenje **1055**
A. Czwyk Marić

Obrazovanje za poduzetništvo pomoću računalne tehnologije u nastavi vježbeničke tvrtke **1060**
I. Ivančić

COMPUTERS IN TECHNICAL SYSTEMS

INVITED PAPER **1067**

New Wave of FPGA Prototyping for System-on-Chip Designs **1069**
H. Mlinarić

PAPERS 1071

Model-based Verification of the SIP Invite Scenario 1073
S. Matetić, M. Bagić Babac

Improved Automatic DTV/STB Menu Navigation for Functional Verification 1079
V. Marinković, B. Kordić, M. Popović, V. Peković

Reliability Sensitivity Analysis for Yahoo! Pipes Mashups 1083
G. Delač, M. Šilić, K. Vladimir

Requirement Bottlenecks in a Cloud Based SCADA System 1089
D. Sutić, B. Atlagić

Impact of Control Flow Blocks Granularity on Custom Processor Design Time 1095
D. Ivošević, V.Sruk

Numerical Relay Protection Coordination Using Simulation Software 1101
S. Nikolovski, M. Havranek, P. Marić

Engine Fault Detection for Piston Engine Aircraft 1106
D. Miljković

Active Noise Reduction for General Aviation Aircraft Using Conventional Headsets 1112
D. Miljković

Reducing Off-road Vehicle Seat Vibrations Using Pneumatic Active Suspension and Fuzzy Logic Based Controller 1118
A. Salihbegović, A. Salihbegović, M. Hadžiselimović, Z. Avdagić

Comparative Analysis Between Methodologies and Their Software Realizations Applied to Modeling and Simulation of Industrial Thermal Processes 1123
B. Dimitrov, H. Nenov, A. Marinov

Twenty Years of Monitoring and Control of Electricity Consumption in RTB Bor, Serbia 1128
V. Tasić, V. Despotović, D. Brodić, M. Pavlov, D. Milivojević

Remote Control of Model Vehicles Using Android Mobile Devices 1133
B. Vunderl, M. Žagar, D. Basch

Optimal Control of a Family House Heating System 1139
M. Vašak, A. Martinčević

Model Predictive Control of a Heating, Ventilation and Air Conditioning System 1145
M. Vašak, A. Starčić

Tuning of Model Predictive Controller for Wind Turbine Load Reduction 1151
G. Banjac, V. Spudić, M. Baotić

Control of the Grid Side Converter in a Wind Turbine 1157
T. Hadjina, M. Baotić, N. Perić

A Leader-Follower Approach to Formation Control of Multiple Non-Holonomic Mobile Robots 1163
T. Petrinić, I. Petrović

One Variant of Self-tuning DC Motor Adaptive Controller 1168
Ž. Jurić, H. Šiljak

An Experience on Manufacturing Execution System in the Field of Metals Industry 1174
J. Novak

Primjena zelenog računalstva u podatkovnom centru Rafinerije nafte Rijeka 1179
S. Horvat, B. Žeželj, Z. Grgić

Nadziranje i dijagnostika rada rafinerijskih postrojenja.....1185
I. Mohler, N. Bolf, R. Karlović, I. Šarlija

Korištenje zalihosne opreme poslužitelja u VHF DSC sustavu Plovputa 1190
J. Vrlić, N. Grgić, I. Ružić

Primjena Microsoft Windows Embedded operativnog sustava na ugrađenim računalnim platformama 1196
S. Braica, N. Grgić

Upotreba mini računala s Linux operativnim sustavom na lokacijama sa zahtjevnim uvjetima okoline 1201
N. Grgić, J. Vrlić, I. Ružić

INTELLIGENT SYSTEMS

INVITED PAPER 1209

Let the Machines Do. How Intelligent is Artificial Intelligence? 1211
H.-D. Burkhard

PAPERS 1217

Some Security Considerations on Crowd-Sourcing an Ontology 1219
M. Migliardi, E. Riccomagno

Bayesian Posterior Probability Classification of Colorectal Cancer Probed with Affymetrix Microarray Technology 1225
M. Simjanoska, A. Madevska Bogdanova, Z. Popeska

Feature Extraction From Electroencephalographic Records Using EEGFrame Framework..1231
A. Jović, L. Sućˇ, N. Bogunović

Technique for Indirect Analysis of Domestic Power Consumers Based on Power Pattern Recognition for Smart Energy Metering.....1237
N. Nikolaev, Y. Rangelov, V. Valchev, A. Marinov

Testing of Several Overlapping Optimization Methods for Bin-Packing Problem.....1241
T. Rolich, D. Domović, D. Grundler

Investigation of Coevolutionary Approach in Gene Regulatory Network Inference.....1247
D. Komlen, D. Jakobović

Evolution of Vehicle Routing Problem Heuristics with Genetic Programming 1254
M. Gulić, D. Jakobović

Multimedia Stimuli Databases Usage Patterns: A Survey Report.....1259
M. Horvat, S. Popović, K. Ćosić

Navigational Evaluation of Breadth First Search Spanning Trees.....1264
D. Helić, M. Strohmaier, W. Wójcik

Comparison of the Nodes of Knowledge Method with Other Graphical Methods for Knowledge Representation 1270
A. Jakupović, M. Pavlić, A. Meštrović, V. Jovanović

Identification of the Global Text Skew Based on the Convex Hulls 1275
D. Brodić, D. Milivojević, V. Tasić, Z. Milivojević

Indexing Positively Constructed Formulas 1281
A. Larionov, I. Terekhin

On a Method of Solution the Equality Problem for Positively Constructed Formulas 1285
I. Terekhin, A. Larionov

The Calculus of Positively Constructed Formulas, its Features, Strategies and Implementation.....1289
A. Larionov, A. Davydov, E. Cherkashin

Logical Programming and Data Mining as Engine for MDA Model Transformation Implementation.....1295
A. Sokolov, I. Terekhin, E. Cherkashin, S. Butakov, A. Larionov

Building Process of SCT Generators 1303
D. Radošević, I. Magdalenić, T. Orehovački

A Proportional Distribution of AUVs under Decentralized Control.....1309
N. Nagul

A P300-Based Brain-Computer Interface: Towards a Simpler Approach 1315
F. Melinščak, A.B. Jerbić, M. Cifrek

Utjecaj nedostajućih vrijednosti na rezultate dubinske analize podataka.....1319
M. Somek

From Digitisation Process to Terminological Digital Resources.....1323
S. Seljan, I. Dunđer, A. Gašpar

Comparative Study of Color Histograms as Global Feature for Image Retrieval 1329
V. Ljubović, H. Šupić

Intelligent Traffic Recommender System.....1334
Lj. Šerić, M. Jukić, M. Braović

Cloning Fuzzy Inference System Component Behavior with an Artificial Neural Network Component 1339
D. Bošković, Z. Avdagić

Principal Component Analysis of Fatigue Strength **1343**
Lj. Šerić, K. Korun Curić, M. Šušnjar

Agile Methods for Cloud Computing **1349**
S. Kalem, D. Đonko, D. Bošković

Generation Technique for Django MVC Web Framework Using the Stratego Transformation Language **1354**
D. Annenkov, E. Cherkashin

INFORMATION SYSTEMS SECURITY

PAPERS **1361**

Security Evaluation in Cloud **1363**
O. Mirković

Security Assessment of Virtual Machines in Open Source Clouds **1369**
A. Donevski, S. Ristov, M. Gusev

Security Analysis of Croatia's Receipt Registration and Verification System **1375**
S. Groš

Conceptual Information Modelling within the Contemporary Information Security Policies **1380**
A. Klaić, M. Golub

Influence of Established Information Security Governance and Infrastructure on Future Security Certifications **1386**
I. Sedinić, Z. Lovrić

Knowledge Based Authentication Requirements **1391**
K. Skračić, P. Pale, B. Jeren

A New Method for the Identification of Proactive Information Security Management System Metrics **1396**
K. Hajdarević, P. Allen

Cyber Crime and Violation of Copyright **1402**
B. Vukelić, K. Škaron

Pravna regulativa kompjuterskog kriminala s osvrtom na Bosnu i Hercegovinu **1406**
F. Gabela

BUSINESS INTELLIGENCE SYSTEMS

INVITED PAPER **1415**

The Generalized Physical Design Problem in Data Warehousing Environment: Towards a Generic Cost Model **1417**
L. Bellatreche, S. Breß, A. Kerkad, A. Boukorca, C. Salmi

PAPERS 1425

Analysis of World Bank Indicators for Countries with Banking Crises by Subgroup Discovery Induction 1427
D. Gamberger, D. Lučanin, T. Šmuc

Transformation of OWL Ontology Sources into Data Warehouse 1432
M. Gulić

Using Big Data and Sentiment Analysis in Product Evaluation 1438
L. Banić, A. Mihanović, M. Brakus

Model of the Business Intelligence System for Credit Risk Analysis 1444
T. Gazdić, Lj. Kašćelan

The Integral OLAP-Model of the Emergency Risk Estimation in the Case of Krasnoyarsk Region 1450
A. Korobko, T. Penkova, V. Nicheporchuk, A. Mihalev

Automatizacija generiranja prezentacijskog sloja skladišta 1456
S. Pavlek, M. Sorić

Upravljanje matičnim podacima u suradnji sa skladištem podataka 1462
A. Matić, M. Sekula, D. Udier

Suvremena rješenja poslovne inteligencije zasnovana na konceptu računarstva u oblaku 1468
I. Sekula, M. Franić

Končar MIS - snažan alat za mudra poduzeća 1474
A. Frančić, D. Ferenčak, D. Cmuk

DIGITAL ECONOMY – 10th Alpe Adria Danube Universities Initiative (ALADIN)

PAPERS 1481

Business Valuation in Oil & Gas Industry: New Challenges 1483
S. Brlečić Valčić, B. Crnković-Stumpf, J. Katunar

Komparativna analiza hrvatskih online novina koje se plaćaju 1489
O. Prlić, A. Lacković, F. Lončar

ePrivacy Rules and Data Processing in Users' Terminal Equipment: a Croatian Experience . 1495
N. Gumzej, S. Grgić

The Cost of Information Security Management in Offshore SMB ICT Companies 1502
S. Aksentijević, E. Tijan, D. Čišić

ICT Contribution to the Economic Development of Some SEE Countries in Transition 1507
M. Vidas-Bubanja

Optimizacija ICT kapaciteta po filozofiji modela JIT(UNV) 1513
Z. Buljubašić, I. Kapetanović Serdarević, N. Buljubašić

Analiza problema repova kod telekomunikacijskih operatera 1518
Z. Buljubašić, N. Salkić, N. Buljubašić

Od sustava e-poslovanja klastera do e-klastera 1523
M. Davidović

Electronic Commerce in Romania Features and Dynamics 1529
D. Enachescu, M. Zaharia

Web 2.0 Tools Support for SME Start-up Projects 1534
I. Ogrizek Biškupić, I. Vrečko, K. Jurina

e-Maintenance for Ship Electrical Propulsion Plants 1540
J. Ćelić, A. Cuculić

Implementing Responsive Web Design for Enhanced Web Presence 1544
S. Mohorovičić

GOVERNMENT, LOCAL GOVERNMENT, PUBLIC SERVICES

PAPERS 1551

Problemi prijenosa podataka iz prostornih planova u geoinformacijski sustav prostornog uređenja u županiji 1553
A. Butigan, M. Pavlić

Measuring Success of Implementation County Development Strategy by Using Balanced Scorecard Method 1559
R. Zdjelar

Effects of e-Government in Croatia 1565
J. Klasinc

E-Hrvatska, jedinstven informacijski sustav 1568
S. Šimundić, J. Šimić

Mogućnosti primjene računalstva u oblaku u tijelima lokalne uprave i samouprave 1574
Ž. Širanović, B. Nožica, D. Bjelobrk

Regulating IT Projects and Information Systems Audit in Public and State Administration of the Republic of Croatia 1580
B. Cerin, G. Vojković

MIPRO Junior - STUDENT PAPERS

PAPERS 1589

Upravljanje vozilom na električni pogon 1591
D. Damjanović, V. Kirinčić, S. Skok

The Development of Charging Stations for Electric Vehicles: A Solution or a Problem? 1599
I. Grdić, V. Kirinčić, S. Skok

Monitoring of Low Voltage Facilities Using Synchronized Phasor Measurements 1604
D. Bulić, R. Jančić, V. Kirinčić, S. Skok

Studentski projekt razvoja vozila na električni pogon kao dijela elektroenergetskog sustava budućnosti 1608
P. Šaša, V. Kirinčić, S. Skok

The Educational Model of the System Integrity Protection Scheme 1614
I. Sterpin, V. Kirinčić, S. Skok

Stochastic Simulation of the Smart Grid and Demand Response Implementations on a City-wide Scale 1620
B. Morvaj, B. Jurišić, N. Holjevac

Framework for Designing a Smart Grid Testbed.....1626
B. Jurišić, N. Holjevac, B. Morvaj

Optimal Scheduling of a CHP System with Energy Storage 1632
L. Majić, I. Krželj, M. Delimar

Poboljšanje performansi bežičnih mreža korištenjem relejnih postaja 1637
A. Caplić, Ž. Ilić

Integration of Advanced Additional Capabilites to a Modern IPTV Platform: a Technical Study 1643
G. Macut, N. Drljević

Upravljanje rehabilitacijskim okruženjem koristeći M2M tehnologiju.....1649
M. Gabrić, J. Novak

Sustav za podršku razvoju edukacijskih aplikacija zasnovanih na potpomognutoj komunikaciji 1653
Š. Brtan, D. Perić, B. Hemar, D. Tomić, H. Novak, Ž. Kompes, T. Vodopijak, M. Vuković, B. Zagajšek, Ž. Car

Benchmarking of Parallel Ray Tracing Using Different Load Granularity 1659
E. Mujagić, A. Demir, N. Omerbegović, E. Mujalo, M. Kozar, N.Nosović

A HTTP-Enabled Cryptographically Secure Hardware Random Number Generator.....1663
M. Božićević, A. Radovan

Parallelization of the Ant Colony Optimization for the Shortest Path Problem using OpenMP and CUDA 1668
M. Arnautović, M. Curić, E. Dolamić, N. Nosović

Protein Database Search Optimization Based on CUDA and MPI 1673
D. Pavlović, R. Vaser, M. Korpar, M. Šikić

Parallel Implementation of the Ant Colony System Algorithm for Automated Combinational Logic Circuits Design Using CUDA Architecture 1676
Dž. Bajgorić, A. Porović, T. Handanović, N. Nosović

Performance Analysis of the Parallelized Matrix Multiplication Using C# Threading.....1680
M. Delić, E. Omerović, A. Sokolović, N. Nosović

Comparison of a Sequential and a MapReduce Approach to Joining Large Datasets.....1684
M. Lalić, E. Memić, F. Kešan, E. Gondžić, N. Smajić, N. Nosović

Acceleration of DCT Transformation in JPEG Image Conversion.....1687
N. Frid, H. Mlinarić, J. Knezović

A Preliminary Study of Croatian Language Syllable Networks.....1691
K. Ban, I. Ivakić, A. Meštrović

Performance Comparison of Sequential and Parallel Execution of the Ant Colony Optimization Algorithm for Solving the Traveling Salesman Problem.....1696
E. Fejzagić, A. Oputić

Parallelization and Performance Analysis of the Simulated Annealing Algorithm for Graph Coloring Problem.....1701
L. Bećirspahić, A. Đulović, N. Nosović

Performance Analysis of the Traveling Salesman Problem Optimization Using Ant Colony Algorithm and OpenMP.....1705
A. Milanović, M. Duranović, N. Nosović

Closed-Loop System for Assisted Strength Exercising.....1709
T. Pozaić, M. Varga, D. Džaja, S. Žulj

LIST OF PAPER REVIEWERS

Abello Gamazo, A.	(Spain)
Aksentijević, S.	(Croatia)
Antolić, Ž.	(Croatia)
Ašenbrener, M.	(Croatia)
Avbelj, V.	(Slovenia)
Babić, D.	(Croatia)
Babić, S.	(Croatia)
Bača, M.	(Croatia)
Bagić Babac, M.	(Croatia)
Bala, P.	(Poland)
Ban, Ž.	(Croatia)
Banek, M.	(Croatia)
Baotić, M.	(Croatia)
Basch, D.	(Croatia)
Bazijanac, E.	(Croatia)
Bažant, A.	(Croatia)
Begušić, D.	(Croatia)
Benc, I.	(Croatia)
Berberović, S.	(Croatia)
Blašković, B.	(Croatia)
Bogunović, N.	(Croatia)
Bojić, I.	(Croatia)
Botički, I.	(Croatia)
Bregar, K.	(Slovenia)
Brkić, Lj.	(Croatia)
Budin, A.	(Croatia)
Budin, L.	(Croatia)
Budiselić, I.	(Croatia)
Buljan, M.	(Croatia)
Butavicius, M.	(Australia)
Butković, Ž.	(Croatia)
Cafuta, D.	(Croatia)
Capan, I.	(Croatia)
Capuder, T.	(Croatia)
Car, Ž.	(Croatia)
Cerinšek, G.	(Slovenia)
Cihlar, D.	(Croatia)
Crnković Stumpf, B.	(Croatia)
Čandrlić, S.	(Croatia)
Čavrak, I.	(Croatia)
Čičin-Šain, M.	(Croatia)
Čubrilo, M.	(Croatia)
Čupić, M.	(Croatia)
Dadić, M.	(Croatia)
Davidović, D.	(Croatia)
Delač, G.	(Croatia)
Delimar, M.	(Croatia)
Dešić, S.	(Croatia)
Dodig, I.	(Croatia)
Dolinšek, S.	(Slovenia)
Džapo, H.	(Croatia)
Feki, J.	(Tunisia)
Filjar, R.	(Croatia)
Fratrić, I.	(Croatia)
Fruk, M.	(Croatia)
Furdek, M.	(Croatia)
Gajić, R.	(Croatia)
Galinac Grbac, T.	(Croatia)
Gamberger, D.	(Croatia)
Garza, P.	(Italy)
Gebavi, H.	(Croatia)
Genova, J.	(Bulgaria)
Gerić, S.	(Croatia)
Glavaš, J.	(Croatia)
Gledec, G.	(Croatia)
Gojanović, D.	(USA)
Golfarelli, M.	(Italy)
Golub, M.	(Croatia)
Golubić, S.	(Croatia)
Groš, S.	(Croatia)
Grudenić, I.	(Croatia)
Gulić, M.	(Croatia)
Hadjina, N.	(Croatia)
Heimburger, A.	(Finland)
Henno, J.	(Estonia)
Hocenski, Ž.	(Croatia)
Horvat, G.	(Croatia)
Hrabar, S.	(Croatia)
Hrkać, T.	(Croatia)
Hudek, V.	(Croatia)
Huljenić, D.	(Croatia)
Humski, L.	(Croatia)
Hursa Šajatović, A.	(Croatia)
Hutinski, Ž,	(Croatia)
Ileš, Š.	(Croatia)
Ilić, D.	(Croatia)
Ipšić, I.	(Croatia)
Ivanda, M.	(Croatia)
Ivašić-Kos, M.	(Croatia)
Ivošević, D.	(Croatia)
Jaakkola, H.	(Finland)
Jakobović, D.	(Croatia)
Jarm, T.	(Slovenia)
Jelenković, L.	(Croatia)
Jevtić, D.	(Croatia)
Ježić, G.	(Croatia)
Ježić, Z.	(Croatia)
Jović, A.	(Croatia)
Juričić, V.	(Croatia)
Kalafatić, Z.	(Croatia)
Kalinić, H.	(Croatia)
Katona, M.	(Serbia)
Kecskeméti, G.	(Hungary)
Kirinić, V.	(Croatia)
Kišasondi, T.	(Croatia)
Knezović, J.	(Croatia)
Kolonić, F.	(Croatia)
Koričić, M.	(Croatia)
Kosec, G.	(Slovenia)
Kostanjčar, Z.	(Croatia)
Kovačić, B.	(Croatia)
Krajcar, S.	(Croatia)
Krašna, M.	(Slovenia)
Križanović, K.	(Croatia)
Krois, I.	(Croatia)
Kušek, M.	(Croatia)
Kuzle, I.	(Croatia)
Lacković, I.	(Croatia)

Lasić-Lazić, J. (Croatia)
Lončar, D. (Croatia)
Lučić, D. (Croatia)
Magdalenić, I. (Croatia)
Majnarić, T. (Croatia)
Malarić, K. (Croatia)
Maleković, M. (Croatia)
Mance, D. (Croatia)
Marciuš, M. (Croatia)
Marčetić, D. (Croatia)
Marinović, M. (Croatia)
Martinčić-Ipšić, S. (Croatia)
Mateljan, V. (Croatia)
Matković, K. (Croatia)
Mauher, M. (Croatia)
Mekterović, I. (Croatia)
Mihajlović, Ž. (Croatia)
Mikac, B. (Croatia)
Mikulandra, N. (Croatia)
Milašinović, B. (Croatia)
Milišić, P. (Croatia)
Mišković, N. (Croatia)
Mitsa, V. (Ukraine)
Mlinarić, H. (Croatia)
Mohorovičić, S. (Croatia)
Molnar, G. (Croatia)
Možek, M. (Slovenia)
Načinović, L. (Croatia)
Palomäki, J. (Finland)
Pavčević, M.O. (Croatia)
Pavlič, J. (Slovenia)
Pavlić, M. (Croatia)
Pečar-Ilić, J. (Croatia)
Pejčinović, B. (USA)
Pelin, D. (Croatia)
Perić, B. (Croatia)
Petošić, A. (Croatia)
Pintar, D. (Croatia)
Podobnik, V. (Croatia)
Popović, M. (Serbia)
Požgaj, Ž. (Croatia)
Rashkovska, A. (Slovenia)
Ribarić, S. (Croatia)
Ristov, S. (Croatia)
Schatten, M. (Croatia)
Seršić, D. (Croatia)
Skala, K. (Croatia)
Skorin-Kapov, L. (Croatia)
Skorin-Kapov, N. (Croatia)
Skuliber, I. (Croatia)
Sluganović, I. (Croatia)
Somek, B. (Croatia)
Sović, I. (Croatia)
Sovilj, S. (Croatia)
Sruk, V. (Croatia)
Stanič, U. (Slovenia)
Stojković, N. (Croatia)
Subašić, M. (Croatia)
Suligoj, T. (Croatia)
Szewczyk, R. (Poland)
Šarić, D. (Croatia)
Šegvić, S. (Croatia)
Šikić, M. (Croatia)
Šilić, M. (Croatia)
Šimić, Z. (Croatia)
Šišul, G. (Croatia)
Škvorc, D. (Croatia)
Šnajder, J. (Croatia)
Šterk, M. (Slovenia)
Šunde, V. (Croatia)
Švedek, T. (Croatia)
Švelec, D. (Croatia)
Tijan, E. (Croatia)
Tomašić, I. (Slovenia)
Tomić, D. (Croatia)
Tomiša, T. (Croatia)
Topić, D. (Croatia)
Topić, G. (Croatia)
Trobec, R. (Slovenia)
Trontelj, J. (Slovenia)
Tržec, K. (Croatia)
Valčić, M. (Croatia)
Varga, M. (Croatia)
Vidaček-Hainš, V. (Croatia)
Vojković, G. (Croatia)
Vranić, M. (Croatia)
Vrdoljak, B. (Croatia)
Vrsalović, D. (Croatia)
Vučić, D. (Croatia)
Vučić, M. (Croatia)
Vukić, Z. (Croatia)
Vukmirović, S. (Croatia)
Weber, M. (Croatia)
Werth, W. (Austria)
Zavec Pavlinić, D. (Slovenia)
Zazula, D. (Slovenia)
Zlatović, M. (Croatia)

AUTHOR INDEX

Ababnah, A. 577
Afgan, E. 263
Aksentijević, S. 1508
Allen, P. 1402
Al-Mistarihi, M. 525, 530, 577
Alomari, Z. 530
Anastassova, M. 872
Anchev, N. 202
Annenkov, D. 1360
Arbula, D. 208
Arnautović, M. 1674
Arsenovski, S. 213
Artemkina, S. 17
Atanasovski, B. 202
Atanassov, E. 189, 197, 254
Atlagić, B. 1095
Avbelj, V. 346, 355
Avdagić, Z. 1124, 1345
Babić, D. 61
Babić, S. 906
Badnjević, A. 153
Bagić Babac, M. 1079
Bajgorić, Dž. 1682
Balarin, M. 21
Baltić, R. 24
Ban, E. 726
Ban, K. 1697
Banić, L. 1444
Banjac, G. 1157
Baotić, M. 1157, 1163
Barak, M. 872
Baranek, M. 521
Barbarić, D. 1019
Bartalis, L. 350
Basch, D. 1139
Bećirspahić, L. 1707
Bedi, K. 736
Bednarkiewicz, A. 44
Bednjanec, A. 757
Bedrač, B. 883
Bego, T. 153
Begović, A. 410, 414
Behlilović, N. 410
Belak, Z. 619
Belehar, I. 384
Bellatreche, L. 1423
Berković, M. 637
Bervanakis, M. 437
Bilas, V. 588
Bjelobrk, D. 443, 806, 1580
Blanco, C. 192
Blažeka, K. 762
Blažič, J. 291
Bogdanov, P. 350
Bogunović, N. 1237
Bolf, N. 1191
Boonsongsrikul, A. 567
Börcsök, J. 94, 98
Borovina, N. 815
Bosilj Vukšić, V. 862
Bošković, D. 815, 1345, 1355
Boukorca, A. 1423
Bouyer, A. 572
Božićević, M. 1669
Braica, S. 1202
Brakus, M. 1444
Braović, M. 1340
Bratina, T. 928
Bregar, K. 346
Breß, S. 1423
Britvić, P. 418
Brlečić Valčić, S. 1489
Brodić, D. 766, 1281, 1134
Brođanac, P. 1023
Bronzin, T. 726
Brtan, Š. 1659
Brumen, B. 655, 846
Budin, L. 1023
Buinac, R. 895
Bulić, D. 1610
Buljubašić, N. 1519, 1524
Buljubašić, Z. 1519, 1524
Burkhard, H.-D. 1217
Butakov, S. 1301
Butigan, A. 1559
Cacciari, C. 165
Capatana, O. 285
Caplić, A. 1643
Car, Ž. 399, 830, 1659
Cerin, B. 1586
Cerinšek, G. 325
Cerkvenik Škafar, A. 325
Cherkashin, E. 1295, 1301, 1360
Chiappini, A. 40
Chiasera, A. 44
Cibiel, G. 40
Cifrek, M. 153, 1321
Cihlar, D. 132
Civa, A. 308
Cmuk, D. 132, 1480
Cofiño, A. 192
Crnković-Stumpf, B. 1489
Crnjak Orel, Z. 21, 24
Csamangó, A. 144
Cuculić, A. 1546
Curić, M. 1674
Cvilinder, H. 742
Cvitaš, Lj. 1031
Czwyk Marić, A. 1061
Čačković, V. 478
Čarapina, M. 806, 1041, 443
Čavka, M. 588
Čeke, D. 463
Čičin-Šain, M. 669, 922
Čičin-Šain, M. 922
Čišić, D. 1508
Čonč, T. 777
Ćelić, J. 1546

Ćosić, K. 1265
Damjanović, D. 1597
Darabkh, K. 525, 577
Davidović, D. 237
Davidović, M. 1529
Davidović, V. 878
Davydov, A. 1295
Delač, G. 1089
Delic, E. 94, 98
Delić, M. 1686
Delić, S. 109
Delimar, M. 1638
Demir, A. 1665
Depolli, M. 223, 273, 369
Despotović, V. 766, 1134
Dešić, S. 437
Dhope, T. 607
Diduck, Q. 61
Dimitrov, B. 126, 258, 1129
Dimitrova, E. 121
Djamil, R. 49
Djinevski, L. 213
Dokić, B. 66
Dolamić, E. 1674
Dolinšek, S. 325
Domović, D. 1247
Donevski, A. 1375
Draženović, M. 748
Drenjanac, D. 1001
Drljević, N. 1649
Duh, M. 928
Dujak, M. 483
Dujmić, H. 301
Duk, S. 443, 806
Duka, D. 502
Dumančić, M. 992, 997, 1001
Dunđer, I. 1329
Duranović, M. 1711
Durchova, M. 197
Džaja, D. 1715
Džebo, A. 547, 555, 637
Đerek, V. 21, 24
Đonko, D. 1355
Đulović, A. 1707
Đurasević, M. 483
Đurek, M. 463, 582
Đurić, J. 976
Ejeckam, F. 61
Enachescu, D. 673, 1535
Faili, F. 61
Fedorov, V. 17
Fejzagić, E. 1702
Ferenczi, J. 350
Ferenčak, D. 1480
Fernández-Quiruelas, V. 192
Féron, P. 40
Ferrari, M. 40, 44
Fertalj, K. 242
Fiameni, G. 165
Fila, N. 912
Filipović Tretinjak, M. 757
Firuzbakht, A. 572
Fouad, M. 497
Francis, D. 61
Frančić, A. 1480
Franić, M. 1474
Frid, N. 1693
Fruk, M. 697
Gabela, A. 561
Gabela, F. 561, 1412
Gabela, J. 561
Gabrić, M. 1655
Gajić, R. 627
Galešev, V. 1005
Gamberger, D. 1433
Gamulin, J. 717
Gamulin, O. 21, 34, 717
Gao, Y. 491
Gašpar, A. 1329
Gazdić, T. 1450
Gingl, Z. 144, 801
Gjenero, L. 177
Glavaš, J. 703
Glavinić Pecotić, M. 104
Glavinić, Đ. 76
Gligora Marković, M. 878
Golub, I. 540
Golub, M. 1386
Gondžić, E. 1690
Goran, N. 414
Gorgan, D. 285, 314
Grdić, I. 1605
Grgić, N. 1196, 1202, 1207
Grgić, S. 1037, 1501
Grgić, Z. 1185
Grm, V. 331
Grmek-Košnik, I. 325
Gros, S. 627
Groš, S. 1381
Grubišić, D. 54
Grundler, D. 1247
Gulić, M. 1260, 1438
Gumzej, N. 1501
Gurov, T. 189
Gusev, M. 171, 202, 213, 248, 934, 1375
Habjanić, H. 521
Hadjina, T. 1163
Hadžialić, M. 414, 613
Hadžiselimović, M. 1124
Hadžović, S. 552
Hajdarević, K. 308, 1402
Handanović, T. 1682
Harasthy, T. 320
Havranek, M. 1107
Hayek, A. 94, 98
Heđi, I. 405
Helić, D. 1270
Hemar, B. 1659
Henno, J. 655, 846
Herić, A. 483
Holomb, R. 34
Holjevac, N. 1626, 1632
Horvat, G. 115
Horvat, M. 1265
Horvat, S. 1185
Horvat, T. 960, 1031

Hudyakova, E. 183
Ilić, Ž. 1643
Imamagić, E. 177
Ivakić, I. 1697
Ivančić, I. 1066
Ivanda, M. 21, 24, 28, 34, 40
Ivanković, A. 731
Ivanković, T. 536
Ivanovska, S. 254
Ivošević, D. 1101
Jaakkola, H. 846
Jakobović, D. 1253, 1260
Jakupović, A. 954, 1276
Jančić, R. 1610
Jelenc, J. 341, 364
Jelenc, J. 364
Jelić, I. 825
Jelkić, V. 976
Jerbić, A.B. 1321
Jeren, B. 664, 1397
Jeričević, Ž. 208, 228
Jin, M. 491
Jókai, B. 350
Jovanov, M. 934
Jovanović, V. 889, 1276
Jović, A. 1237
Jukić, M. 1340
Jukić, O. 405
Jurić, Ž. 109, 1174
Jurina, K. 1540
Jurišić, B. 1626, 1632
Jurišić, D. 76
K. Korun Curić, 1349
Kalem, S. 1355
Kamenar, E. 138
Kapetanović Serdarević, I. 1519
Karahmetović, N. 428
Karaivanova, A. 189
Karin, I. 619
Karlović, R. 1191
Karóczkai, K. 350
Kasapović, S. 463
Kašćelan, Lj. 1450
Kaštelan, I. 872
Katunar, J. 1489
Kazaz, T. 613
Kažović, D. 811, 987
Kepinski, L. 44
Kerkad, A. 1423
Kermek, D. 717
Kešan, F. 1690
Kezić, D. 598
Khandavalli, C. 61
Kheireddine, L. 49
Kiralj, D. 795
Kirinčić, V. 1597, 1605, 1610, 1614, 1620
Klaić, A. 1386
Klasinc, J. 1571
Klemenčić, M. 278
Klindžić, J. 917
Knezović, J. 1693
Kocijančič, S. 567
Kokkinaki, A. 713
Kolarić, D. 649
Komlen, D. 1253
Kompes, Ž. 1659
Kondrat, O. 34
Kontošić, A. 970, 982
Kontošić, B. 970, 982
Kordić, B. 1085
Korobko, A. 1456
Korpar, M. 1679
Koruga, D. 153
Kosec, G. 223
Kosovac, A. 637
Kosovec, M. 992
Kosović, M. 21
Kovač, D. 940
Kovačević, D. 5
Kovačević, S. 552
Kovačić, B. 841
Kowsar, Y. 263
Kozar, M. 1665
Kozlovszky, M. 269, 350
Kožar, I. 228
Krajczyk, L. 44
Krašna, M. 883, 928
Krelja-Kurelović, E. 819, 856
Krisper, M. 219
Krivošić, A. 555
Krpetić, R. 588
Krželj, I. 1638
Kuč, A. 422
Kuč, H. 422
Kulin, M. 613
Kusić, N. 454
Lacković, A. 1037, 1495
Lalić, M. 1690
Lamza-Maronić, M. 703
Larionov, A. 1287, 1291, 1295, 1301
Lavrov, A. 17
Lazić, N. 917
Lebinac, V. 507
Li, Y. 491
Lipić, T. 237, 273
Lippert, T. 165
Lisek, J. 790
Lončar, F. 1495
Lovas, G. 34
Lovrek, I. 830
Lovrić, Z. 1392
Lučanin, D. 1433
Lukač, D. 669
Lukić, M. 997
Lukowiak, A. 44
Ljubović, V. 1335
Macut, G. 1649
Maček Lebar, A. 364
Madevska Bogdanova, A. 1231
Magdalenić, I. 1309
Majić, L. 1638
Majstorović, N. 471
Makan, G. 144
Mäkelä, J. 655, 846
Malačič, V. 223
Marcijuš, M. 24

Marić, P. 1107
Marijanović, B. 790
Marinković, V. 1085
Marinov, A. 121, 258, 1129, 1243
Markučič, Z. 1023
Markulin, D. 449
Martinčević, A. 1145
Martinović, A. 208
Martinovikj, D. 934
Marušić, N. 513
Matetić, S. 1079
Matić, A. 1468
Matić, V. 1046
Matković, J. 242
Matolín, V. 34
Mazloom-Nezhad, B. 72
Mazzola, M. 40
Medvedev, M. 17
Meixner, Z. 350
Melinščak, F. 1321
Mellár, J. 144
Memić, E. 1690
Memon, A. 165
Memon, M. 165
Menciger, M. 1027
Meštrović, A. 1276, 1697
Mežnarić, I. 940
Migliardi, M. 1225
Mihajlović, Ž. 297
Mihalev, A. 1456
Mihanović, A. 1444
Miklavčič, D. 364
Milanović, A. 1711
Milivojević, D. 766, 1134, 1281
Milivojević, Z. 1281
Miljko, D. 731
Miljković, D. 1112, 1118
Mingesz, R. 144
Mirković, M. 964
Mirković, O. 1369
Misnevs, B. 912
Mitsa, V. 34
Mlinarić, H. 1075, 1693
Mohammad, A. 577
Mohler, I. 1191
Mohorovičić, S. 1550
Mokhtar Mansour, A. 497
Molnar, G. 104
Morvaj, B. 1626, 1632
Možnik, D. 1052
Mrak, Z. 598
Mrazović, P. 297
Mržek, M. 291
Mujagić, E. 1665
Mujalo, E. 1665
Musa, K. 449
Musić, S. 21
Mutapčić, H. 547
Nagul, N. 1315
Najafi, A. 72
Najafi, A. 72
Naumov, N. 17
Németh, L. 350
Nenov, H. 258, 1129
Neumeier, R. 603
Nicheporchuk, V. 1456
Nikitović, A. 976
Nikolaev, N. 1243
Nikolić, M. 649
Jaakkola, H. 655
Nikolić, V. 649
Nikolov, G. 126
Nikolovski, S. 1107
Nosović, N. 1665, 1674, 1682, 1686, 1690, 1707, 1711
Novak, H. 1659
Novak, J. 1180, 1655
Nožica, B. 1041, 1580
Očko, M. 28
Odak, M. 917
Oder, A. 331
Ogrizek Biškupić, I. 1540
Olgić Draženović, B. 946
Omerbegović, N. 1665
Omerović, E. 1686
Oputić, A. 1702
Orehovački, T. 691, 1309
Ostermayer, G. 603
Ovsénik, Ľ. 320, 595
Pajkanović, A. 66
Pale, P. 664, 1397
Paľová, D. 771
Papić, A. 1015
Papić, V. 301
Parać, V. 483
Pavlek, S. 1462
Pavlič, J. 359
Pavlić, M. 1276, 1559
Pavlina, K. 723
Pavlov, M. 766, 1134
Pavlović, D. 1679
Pečlin, P. 389
Pejčinović, B. 82
Pekorari, R. 960
Peković, V. 1085
Penkova, T. 1456
Perić, D. 1659
Perić, N. 1163
Perić, S. 1023
Petračić, R. 521
Petreckiy, S. 34
Petrinić, T. 1169
Petrović, B. 723
Petrović, I. 1169
Petrović, J. 664
Pibernik, J. 830
Pilipović, M. 297
Pirić, E. 954
Pivačić, N. 454
Pogarčić, I. 878
Polajnar, J. 384
Popeska, Z. 1231
Popović, M. 1085
Popović, S. 1265
Popović, Ž. 478
Popovych, N. 34
Porović, A. 1682

Póserné, V. 269
Pozaić, T. 1715
Požgaj, Ž. 862
Prčić, V. 582
Prince, K. 34
Prlić, O. 1495
Prohaska, Z. 946
Prohaska, Z. 950
Radchenko, G. 183
Radojević, B. 540
Radošević, D. 1309
Radovan, A. 987, 1669
Rajtar, B. 521
Rako, S. 856
Rangelov, Y. 1243
Rashkovska, A. 355, 369
Rasoloniaina, A. 40
Révészová, L. 784
Riccomagno, E. 1225
Riedel, M. 165
Righini, G. 40
Rimac, M. 418
Ristić, D. 40, 44
Ristić, M. 21, 24
Ristov, S. 171, 202, 213, 248, 1375
Rolich, T. 1247
Romanenko, A. 17
Rotim, M. 900
Rozman, J. 389
Rožanc, I. 835
Ružić, I. 619, 1058, 1196, 1207
Sabou, A. 314
Salihbegović, A. 1124
Salihbegović, A. 1124
Salima, B. 49
Salkić, N. 1524
Salmi, C. 1423
Sandu, C. 150
Schreiber, M. 94, 98
Schubert, T. 269
Sedinić, I. 1392
Sekula, I. 1474
Sekula, M. 1468
Sekulić, T. 751
Seljan, S. 1329
Sertić, M. 1009
Shah, A. 54
Shurman, M. 525, 530, 577
Simjanoska, M. 171, 1231
Sirotković, J. 301
Skala, K. 237, 273, 278, 375
Skendžić, A. 507, 841
Skok, S. 1597, 1605, 1610, 1614, 1620
Skorin-Kapov, L. 397
Skračić, K. 1397
Smajić, N. 1690
Smith, H. 153
Sokolov, A. 1301
Sokolović, A. 1686
Somek, M. 1325
Sorić, M. 1462
Sović, I. 375
Spudić, V. 1157
Sruk, V. 872, 1101
Stancescu, S. 150
Stančić, Z. 830
Stanič, J. 325, 341
Stanič, U. 325, 331, 341
Staničić, O. 232
Starčić, A. 1151
Sterpin, I. 1620
Strek, W. 44
Strohmaier, M. 1270
Sućˊ, L. 1237
Sukser, T. 521
Suppharangsan, S. 567
Sutić, D. 1095
Szepe, T. 144, 801
Šarić, V. 946
Šarlija, I. 1191
Šaša, P. 1614
Šerić, Lj. 1340, 1349
Šikić, M. 375, 1679
Šilić, M. 1089
Šiljak, H. 1174
Šimara, S. 790
Šimec, A. 232
Šimić, J. 1574
Šimundić, S. 1019, 1574
Šimunić, D. 582, 607
Šimunović, M. 132
Širanović, Ž. 471, 1041, 1580
Škaljo, N. 410
Škaron, K. 1408
Škrinjar, J. 830
Šlibar, N. 521
Šmuc, T. 1433
Špiranec, S. 731
Špoljarić, T. 697
Šupić, H. 1335
Šušnjar, M. 1349
Švedek, T. 115
Takači, Đ. 751
Tanatarec, B. 633
Tarasenko, M. 17
Tasić, O. 766
Tasić, V. 766, 1281, 1134
Tatarko, M. 595
Temerinac, M. 872
Terekhin, I. 1287, 1291, 1301
Tijan, E. 1508
Tolić, T. 471
Tomašić, I. 369
Tomić, D. 177, 1659
Tomljanović, J. 819, 856
Tomljenović, V. 895
Törőcsik, M. 269
Toth, S. 801
Trkman, M. 219
Trobec, R. 273, 384
Tsud, N. 34
Turán, J. 320, 595
Turina, T. 819
Udier, D. 1468
Ugovšek, J. 223
Ungermanns, C. 867

Uroda, I. 950
Vadai, G. 144
Valchev, V. 121, 126, 1243
Valčić, S. 598
Valenčić, D. 811, 841, 987, 507
Varga, M. 1715
Vaser, R. 1679
Vasić, D. 588
Vašak, M. 1145, 1151
Vavpotič, D. 219
Vejačka, M. 771
Velkoski, G. 171, 248
Veselko, M. 384
Vidas-Bubanja, M. 1513
Vidović, K. 418
Vincek, D. 707
Vinko, D. 115
Vladimir, K. 1089
Vladimirou, A. 713
Vlahović Štetić, V. 830
Vodopijak, T. 1659
Vojković, G. 1586
Volarević, M. 297
Vrana, R. 679, 685
Vrbanec, T. 736
Vrečko, I. 1540
Vrhovec, S. 219
Vrkić, D. 825
Vrlić, J. 619, 1196, 1207
Vučić, M. 104
Vujisić, G. 697
Vukelić, B. 1408
Vukmirović, S. 922, 940
Vuković, M. 1659
Vunderl, B. 1139
Wang, G. 491
Werth, W. 867
Wiglusz, R. 44
Wójcik, W. 1270
Zagajšek, B. 1659
Zaharia, M. 673, 1535
Zaluški, B. 521
Zavec Pavlinić, D. 331, 336
Zdjelar, R. 1565
Zelenika, S. 138
Zima Krnelić, L. 649
Žagar, M. 1139
Žajdela Hrustek, N. 691, 736
Žeželj, B. 1185
Žonja, S. 28
Žufić, J. 742, 795
Žulj, S. 1715

FOREWORD

The *36th International ICT Convention MIPRO 2013* was held from 20st until 24th of May 2013 in Opatija, the Adriatic Coast, Croatia. The Convention consisted of ten conferences under the titles: *Microelectronics, Electronics and Electronic Technology (MEET), Distributed Computing and Visualization (DC VIS), Telecommunications & Information (CTI), Computers in Education (CE), Computers in Technical Systems (CTS), Intelligent Systems (CIS), Information Systems Security (ISS), Business Intelligence Systems (miproBIS), Digital Economy (DE), Government, Local Government, Public Services (GLGPS)*. A special conference was dedicated to works of the students: *MIPRO Junior-Student Papers (SP)*.

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

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

To all who helped organize the 36th International ICT Convention *MIPRO 2013* as well as editing of this Book of Proceedings we extend our heartfelt thanks.

Prof. Petar Biljanović, PhD
International Program Committee Chairman

MEET

International Conference on

MICROELECTRONICS, ELECTRONICS AND ELECTRONIC TECHNOLOGY

Steering Committee

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

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

INVITED PAPER

Min-Max Circuits as a Problem Solving Tool(s)

Darko Kovačević
Faculty of Maritime Studies Split, University of Split,
Zrinsko-Frankopanska 38, 21000 Split, Croatia
Phone: +385 21 380 762, E-mail: darko.kovacevic1@st.t-com.hr

Abstract – In numerous practical circuit applications, including waveshaping and nonlinear function generation, waveforms are intentionally clipped. That can be achieved in many different ways using diodes, transistors and integrated circuits. In this paper (invited lesson), we are going to show how dedicated min and max circuits can be used for mentioned purpose(s). Also, we are going to show how dedicated min and max circuits can be introduced as a problem solving tool in a soft computing, AI-modeling, etc.

I. INTRODUCTION

In numerous practical circuit applications, including wave-shaping and nonlinear transfer function generation waveforms are intentionally clipped. That can be achieved in many different ways [1]. In this article we are going to show how dedicated *min* and *max circuit* can be used for such purpose. On the other hand, our intention is to show how *min* and *max circuit* applications can be diverse and versatile in the very hardware design idea [2]. Mentioned *min* and *max circuits* can be used as a problem solving tool in many technical and non-technical areas; from electronics to applied mathematics, Petri nets, soft computing and AI systems [3], [4].

In the beginning we will show how *min* and *max circuits* can be used in the design of clipping devices. Then, clipping devices will be used for different purposes. For example, the transfer characteristic of a clipping device can be used in modeling an AI environment (syllogisms, membership functions, etc.)[5]. More complex devices and transfer characteristics, based on clipping circuits, will be created to represent agent(s) activities in a hardware model of newspaper distribution and sell process (NDSP) [4].

Also, we will show how *min* and *max circuits* can be used in every day electronic practice as rectifiers, active peak detectors and transfer function generators.

A simple EWB program will be used as a support hardware model tool, but note that all circuits shown in this presentation have been designed in a real hardware and that we are using them in student laboratory exercises at Faculty of Maritime Studies.

II. MIN-MAX CIRCUITS AND STRAIGHT LINE GENERATORS

In general, our approach is based on two presumptions:

- amplifier can represent a straight line in a phase space as it is shown in Fig. 1. and Fig. 2. [2].

- *min* and *max circuits* can be used to generate different transfer functions

Straight line generators can be realized in hardware environment on many different ways [4]. One model of such generator is shown in Fig. 1.

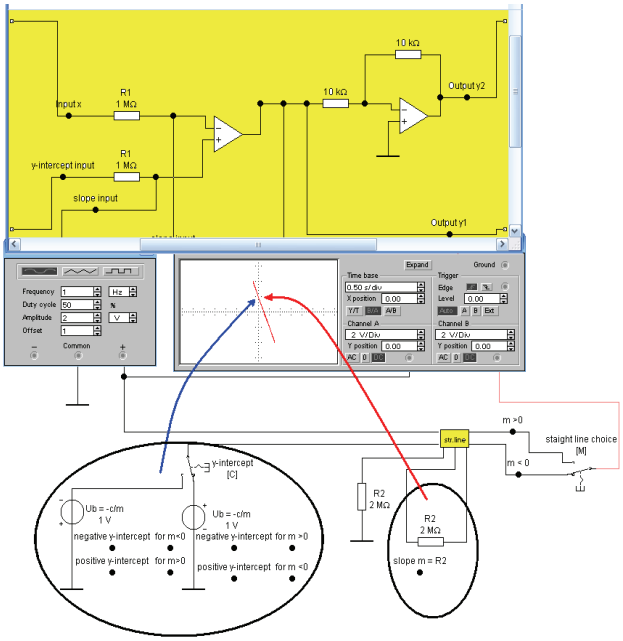


Figure 1. Straight line generator ($l \equiv y = mx + c$) in an *Op-Amp* technique

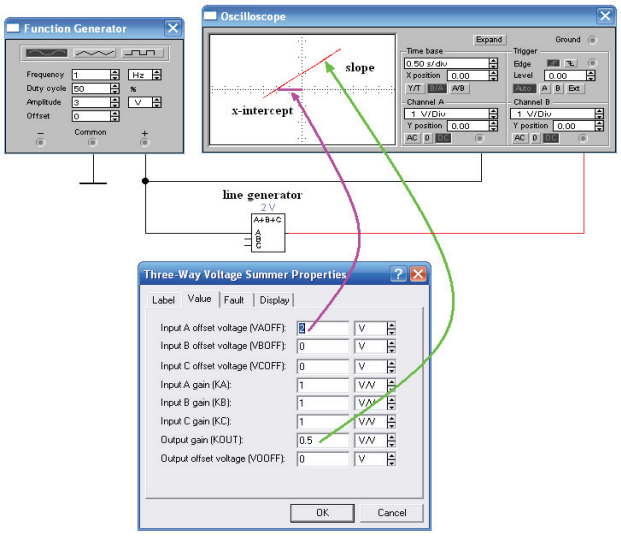


Figure 2. „Three-Way Voltage Summer“ generates straight line(s)

A variety of *min* and *max circuits* performing corresponding set operations (see Fig. 3.) can be found in literature [5], but in the most of our applications we will use our original circuits (see Fig. 4. and Fig. 5.) based on standard operational amplifiers [6].

A. Min and Max Circuits

Min and *max circuits* as a problem solving tool rely, on a mathematical theory of max (min) algebra [7], [8], [9], [10], and [11].

B. Min Circuit

Definition. Assume A and B are two fuzzy subsets of X. Their intersection is a fuzzy subset D of X, denoted $D = A \cap B$ (see Fig. 3. for illustration), such that for each $x \in X$

$$D(x) = \min[A(x), B(x)] = A(x) \cap B(x) \tag{1}$$

It is common practice in the soft computing (fuzzy logic) literature to use symbol $\wedge$ as the *min* operator. The operation *min* can be written in algebraic form that is convenient for hardware implementation [4]:

$$\min(a,b) = \left((a+b) - |a-b|\right)/2 \tag{2}$$

Using relation (2) we can design a *min circuit* in *Op-Amp* environment as it is shown in Fig. 4.

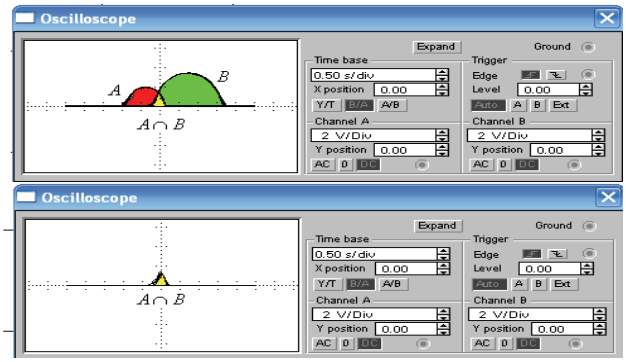


Figure 3. Fuzzy set intersection or fuzzy logic AND-function

C. Max circuit

Definition. Assume A and B are two fuzzy subsets of X. Their intersection is a fuzzy subset C of X, denoted $C = A \cap B$, such that for each $x \in X$

$$C(x) = \max[A(x), B(x)] = A(x) \cup B(x) \tag{3}$$

It is common practice in the soft computing (fuzzy logic) literature to use symbol $\vee$ as the *max* operator. The operation *max* can be written in algebraic form:

$$\max(a,b) = \left[(a+b) + |a-b|\right]/2 \tag{4}$$

Using relation (4) we can design a *max circuit* in *OP-Amp* environment as it is shown in Fig. 5.

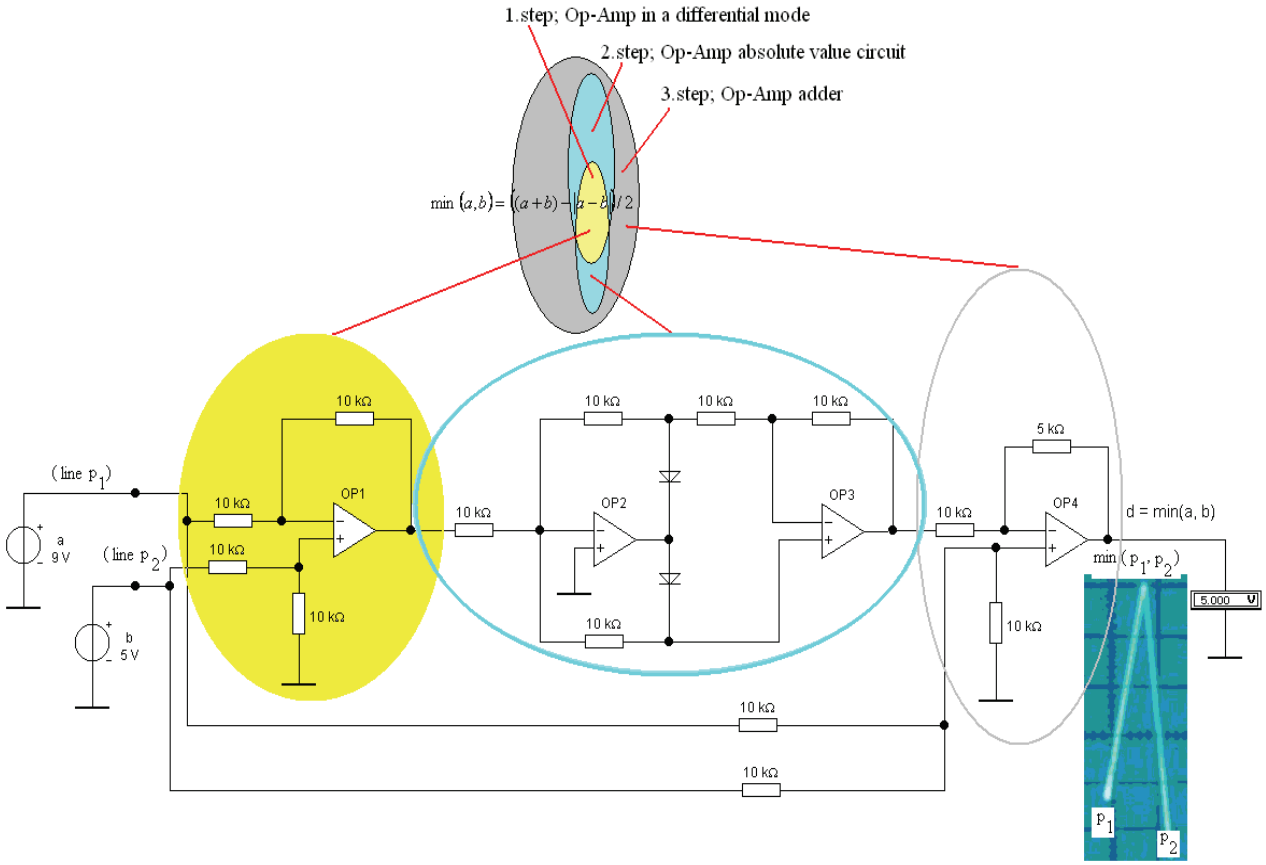


Figure 4. *Min circuit* in *Op-Amp* technique along with its „static“ and „dynamic“ responses to applied inputs. Note that *Op-Amps* (see Fig. 1.) were used to generate lines p_1 and p_2 to get “dynamic” response in a phase plain

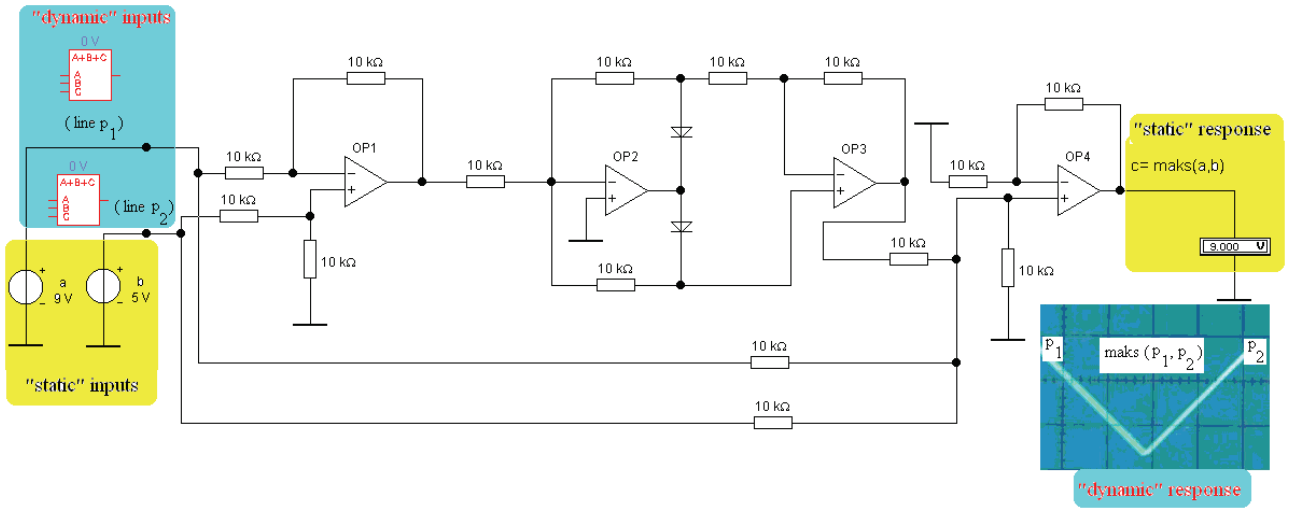


Figure 5. *Min circuit* in *Op-Amp* technique along with its „static“ and „dynamic“ responses to applied inputs. Note that EWB *Op-Amps* were used to generate lines p_1 and p_2 (“dynamic” response)

III. Min (Max) Circuit Operative Extensions

A “better” look over *min* (*max*) *circuit* designed *Op-Amp technique* gives us an opportunity to reconsider about additional facilities of that circuit; *min/max function exchange* can be easily established and new outputs give that circuit extra applicability (see Fig. 6.)

A. Positive and negative clipping

Clipping can be recognized as an undesirable result of overdriving an amplifier, i.e. any attempt to push an output voltage beyond the limits through which it can “swing” causes the tops and/or bottoms of a waveform to be “clipped” off. However, in numerous practical applications, including waveshaping and nonlinear function generation, waveforms are intentionally clipped [1]. Clipping as a notion can be introduced in the design of AI-algorithms [4].

Most often, the diodes along with reference batteries (voltage) are used to select for transmission a part of a waveform which lies above or below some reference voltage level. In some applications diodes are combined with operational amplifier(s) to obtain precision limiting circuits [1].

In this paper a somewhat different approach (a functional approach) to clipping actions is described. The approach is based on “standard” *min* and *max circuits* in “functional combination” with operational amplifiers that “play a geometric role” in the design of a desired transfer function [2] [3] [4].

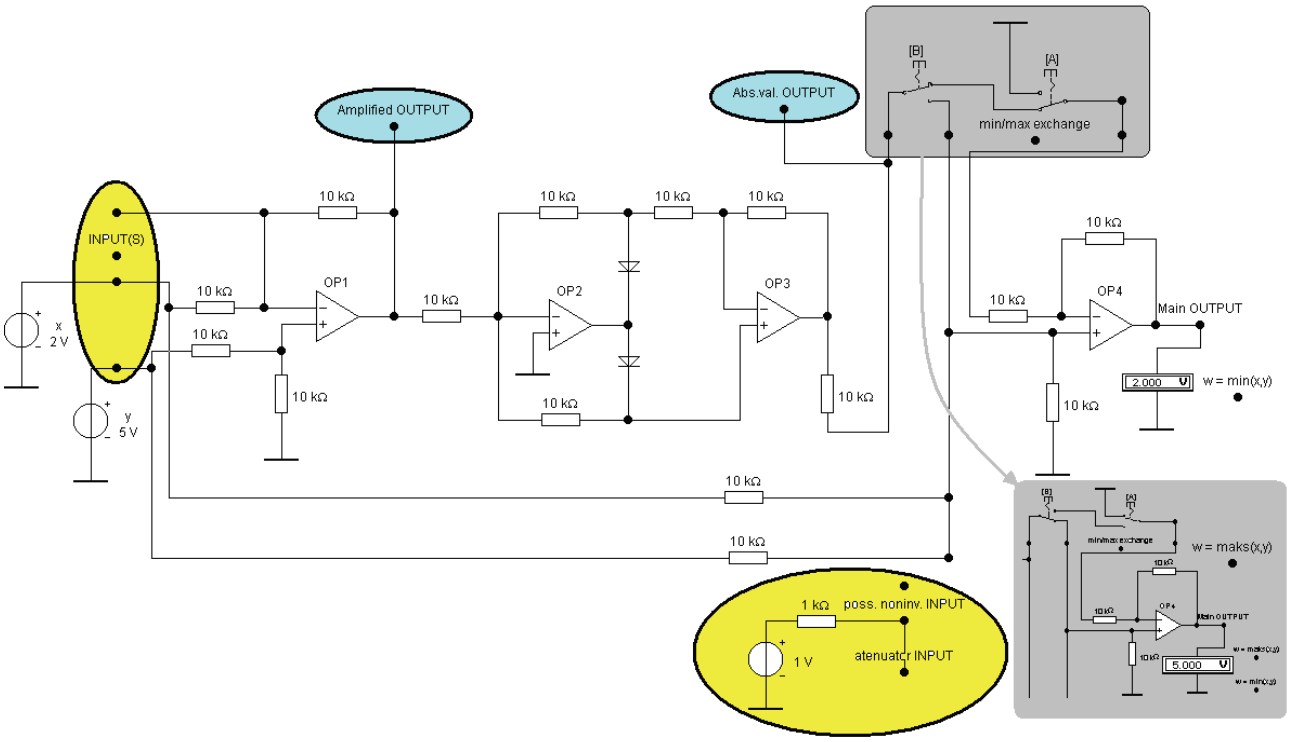


Figure 6. *Min* (*max*) *circuit* designed in *Op-Amp technique* can have additional features; easy switching between *min* and *max* function, difference absolute value output, etc.

Now, instead of using biased diodes, clipping can be accomplished by designing a desired transfer characteristic of a clipping device. *Min* and *max circuits* are used to combine linear transfer characteristics of operational amplifiers in a design of a nonlinear transfer characteristic of a certain clipping circuit. In that manner it is easy to design and mathematically describe a circuit transfer function of:

- positive clipping circuit
- negative clipping circuit
- positive and negative clipping circuit
- negative peak pass circuit
- positive peak pass circuit

If we take $E_B^- = 0$ then the transfer characteristic of a positive and negative clipping device can be used to represent a fuzzy set of "departures from zero".

B. Positive clipping

If a sinusoidal waveform is to be clipped on a certain positive level (see Fig. 7.), two straight lines (amplifiers) have to be combined in a phase plain along with *min circuit* function to design desired transfer function G^+. Horizontal line $l_1^+ \equiv v_0$ represents clipping level ($v_0 = +E_B$) while the line $l_2^+ \equiv v_o = m^+ v_i + c^+$ outside of clipping level, having a slope $m^+ = 1$ and $c^+ = 0$, represents a linear region of a device (unity gain region).

In that case the transfer characteristic of the clipping device can be formally represented as:

$$G^+ = \min(l_1^+, l_2^+) \tag{5}$$

It is obvious (see relation (5)) that lines l_1^+ and l_2^+ are "inputs" of a *min circuit*.

C. Negative clipping

If a sinusoidal waveform is to be clipped on a certain negative level (see Fig. 8.), two straight lines (amplifiers) have to be combined in a phase plain along with *max* circuit function. Horizontal line $l_1^- \equiv v_0$ represents clipping level ($v_o = -E_B$) while the line $l_2^- \equiv v_o = m^- v_i + c^-$ outside of clipping level, having a slope $m^- = 1$ and $c^- = 0$, represents a linear region of a device (unity gain region). Well specified (defined) lines (*Op-Amps* outputs) are applied to *min* and (or) *max* circuit inputs to achieve desired non-linear transfer characteristic.

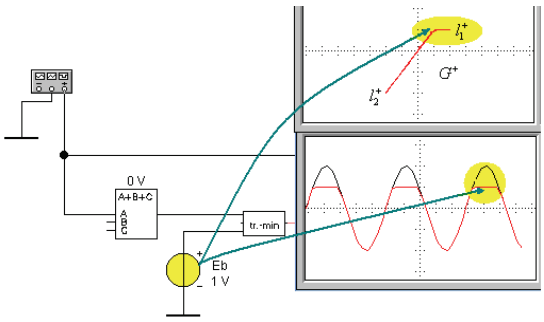


Figure 7. Positive clipping

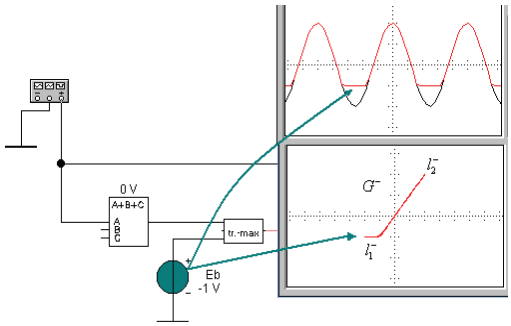


Figure 8. Negative clipping

The transfer characteristic of the clipping device can be now formally represented as:

$$G^- = \max(l_1^-, l_2^-) \tag{6}$$

In this case lines l_1^- and l_2^- are "inputs" of a *max circuit* (see relation (6)).

D. Positive and negative clipping

Positive and negative clipping device is a *min-max* combination (see Fig. 9.), as its transfer function *G* can be expressed as:

$$G = \min(E_b^+, \max(E_b^-, l)) \tag{6}$$

where *l* represents a straight line $u_{out} = m\, u_{in}$ with $m = 1$.

Clipping levels can be regarded as time varying variables and such conditions in a clipping device can be modeled as it is shown in Fig.10.

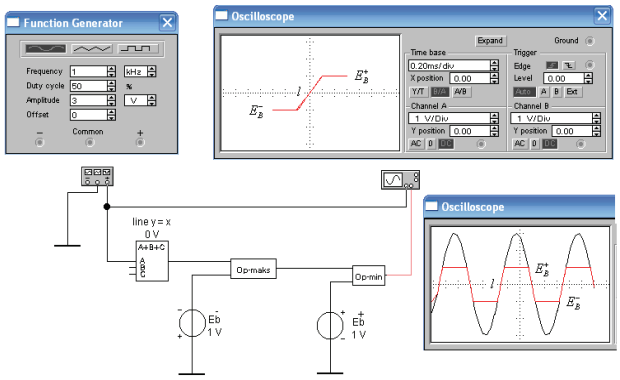


Figure 9. Positive and negative clipping device

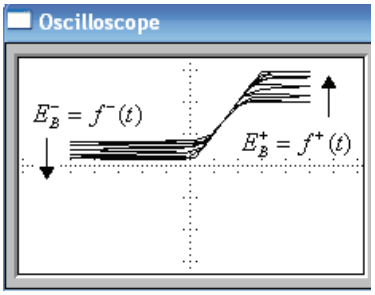


Figure 10. Positive and negative clipping device with time dependent (modulated) clipping levels. Clipping levels can be a subject of agent negotiations what can cause their changes in a certain time frame

IV. AI APPLICATION OF A CLIPPING DEVICE

Clipping devices showed before can be used in modeling so called *four-part multiagent agent diagram* shown in Fig.11. [4.]. Different transfer characteristics can represent different (re)actions of a *wholesaler (agent)* in the NDSP [4.].

Let say that the agent activities are explained by a set of rules:

If operator's order is too high, then impose a certain limitation to operator's current order
If operator's order is a zero order, then according to a current publisher wish,
impose a minimal order (so called "presence")to an agency

In this case agent activities can be modeled using a time dependent clipping device showed in Fig.9. One must know that *order limitations* as well as the "*presence*" are variable quantities, along a tourist season.

The transfer characteristic of a *wholesaler* (*agent*) can be expressed using relation (7):

$$G(k) = \min(n^W_{\max}(k), \max(n^W_{\min}(k), n^O(k))) \qquad (7)$$

where $n^W_{\max}(k)$ represents maximum permissible order for an agency, $n^W_{\min}(k)$ represents so called presence, i.e. minimal permanent order that must be distributed through an agency and $n^O(k)$ is a local operator order for a *k*-day. Note that *wholesaler* (*agent*), for some reasons can impose restrictions (changes) on operator's order(s) [4].

V. PEAK PASS DEVICES

Somehow different kind of clipping action is introduced in so called *peak pass devices* (see Fig.12. and Fig.13.). Instead of the positive or negative peaks being chopped off, the device follows ($v_{out} = v_{in}$) the input when signal is above or below a certain level. The transfer characteristics show that linear operation (input signal following) occurs only when certain signal levels are reached and that the output remains constant below those levels.

If we combine a straight line generator having $v_{out} = v_{in}$ with a battery ($+E_B$) along with a *max* circuit then we are going to realize a *positive peak pass device*. Transfer characteristic of the device is shown in Fig. 12.

When the output of a straight line generator ($v_{out} = v_{in}$) and a battery ($-E_B$) are applied to the inputs of a *min* circuit (see Fig. 13.) a *negative peak pass device* will result. Input and output waveforms along with a transfer characteristic of the device are shown in Fig. 13.

VI. HALF AND FULL WAVE RECTIFIERS

The design of a half-wave rectifier in *min-max-circuit technique* is quite simple (see Fig. 14.). A sinusoidal voltage $E\sin\omega t$ is applied to one input on of a *max circuit* while the other input of *max circuit* is grounded (switch *S* is open). Therefore, output voltage can be expressed as:

$$v_{out}(t) = \max(E\sin\omega t, 0)) \qquad (8)$$

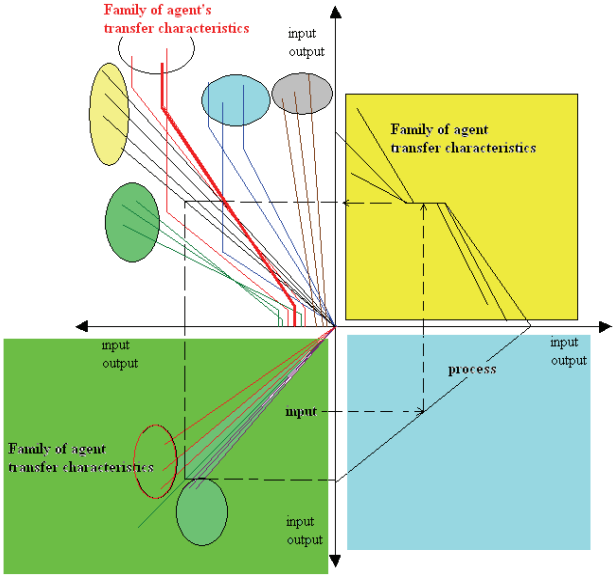


Figure 11. *Four-part multiagent diagram*; *wholesaler* (*agent*) controls order(s) on an operator [4]

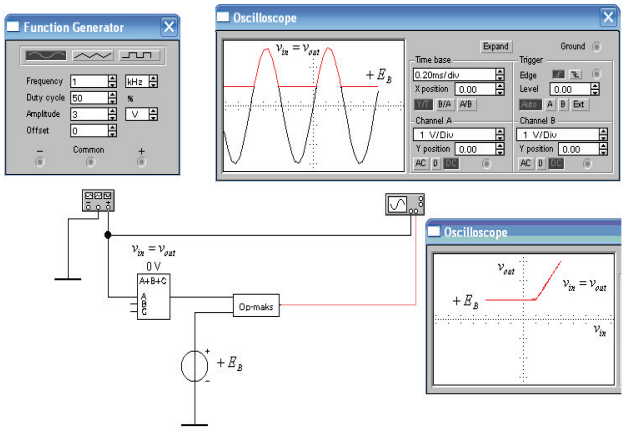


Figure 12. Positive peak pass device; $G = \max(+E_B, v_{in})$

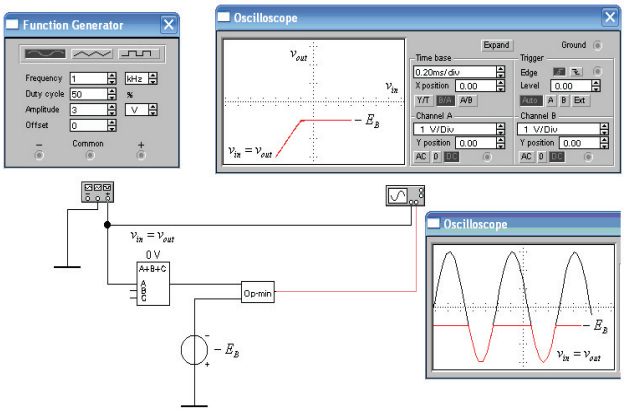


Figure 13. Negative peak pass device; $G = \min(-E_B, v_{in})$

A full-wave rectifier in *min-max-circuit technique* (see Fig. 14.) can be designed in many ways (see Fig. 14 and Fig. 15.)[4]. If we use a circuit shown in Fig. 14, then the voltage on its output can be expressed by relation:

$$v_{out}(t) = |E\sin\omega t| = \max(E\sin\omega t) - \min(E\sin\omega t) \qquad (9)$$

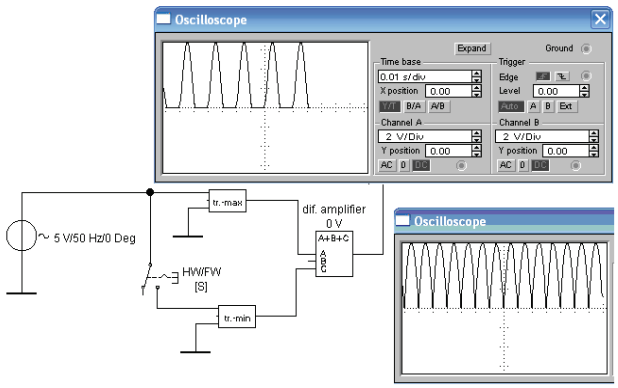


Figure 14. A combination of a half and a full wave rectifier in *min-max-circuit technique*

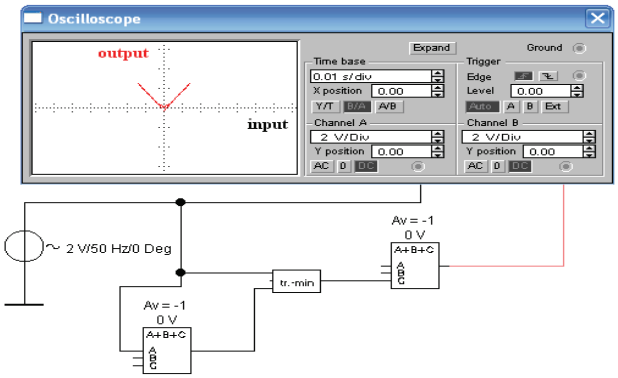


Figure 15. $v_{out} = -\min(v_{in}, -v_{in})$

A. D-operators

By modifying the *min*/*max* clipping circuits, we can obtain a class of transfer characteristics called *D-operators*, i.e. "ideal diode" operators (see Fig. 16.). For switching purposes the "ideal diode" is one way switch that is open when the imposed voltage is of one polarity and closed when the polarity is opposite.

The *D-operator* circuit is a voltage-to-voltage *min* (*max*) *circuit* that would have the same response as a circuit that used an "ideal diode" as a switching element (see Fig. 16.).

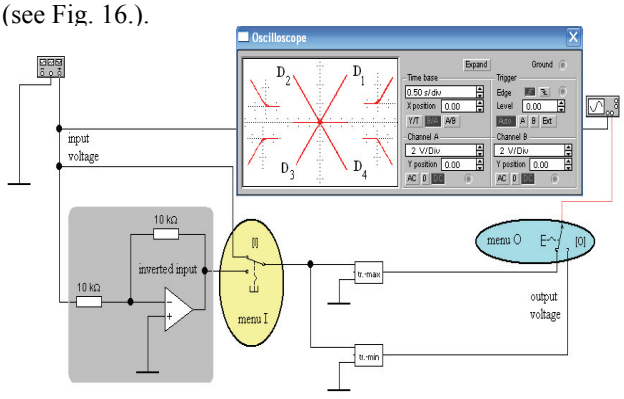


Figure 16. *D-operators* generator; Combing input and output *menu* it is possible to generate transfer characteristics of four precision rectifiers; $G_{D_1} = \max(0, v_{in})$, $G_{D_2} = \max(0, -v_{in})$, $G_{D_3} = \min(0, v_{in})$ and $G_{D_4} = \min(0, -v_{in})$

VII. MIN-MAX CIRCUITS AS A DEDICATED PROBLEM SOLVING TOOL

A. Solving (in)equalities

Let say that we want to find an electronic graphical solution for:

a) $|x+1| = 3$ and

b) $|x+1| \geq 3$

Let us graph $|x+1| = 3$. We will start with the plot of $|x|$ (see Fig. 17.).

The plot of $|x|$ have to be shifted to the left to make $|x+1|$, as it is shown in Fig. 18.

Now, the plot of $|x+1|$ has to be shifted down to make it $|x+1| - 3 = 0$ (see Fig. 19.) to get final solution, i.e. $x_1 = 2$ and $x_2 = -4$.

In the second case ($|x+1| \geq 3$) we need to add one *min circuit* to the output of the circuit shown in Fig. 19., to get the set of solutions, i.e. numbers *x* whose distance to -1 is at least 3 (see Fig. 20.).

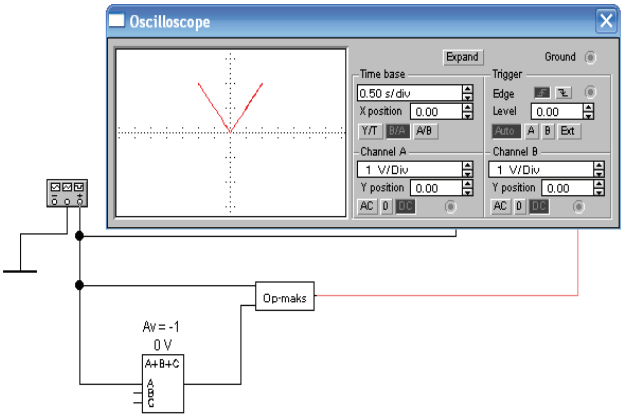


Figure 17. Plot of $|x|$ in a phase plain; $|x| = \max(x, -x)$

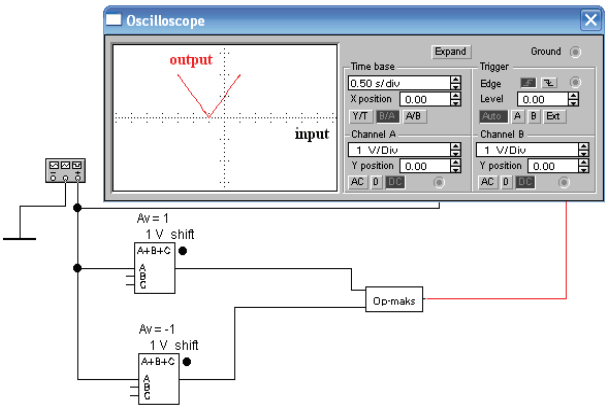


Figure 18. Plot of $|x|$ is shifted left to make the plot of $|x+1|$.

Two *Op-Amps* are need for that graph translation; $y = u_{out} = \max((x+1), -(x+1))$

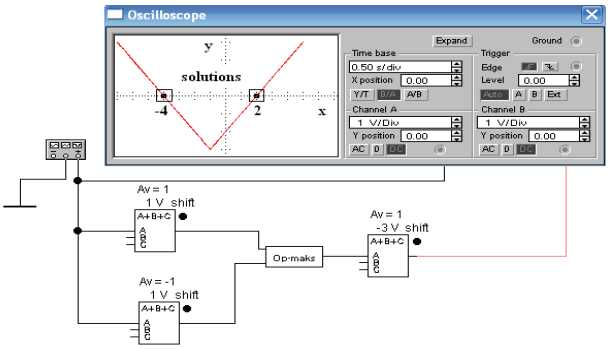


Fig. 19. One can see two solutions: $x_1 = u_{in} = -4$ or $x_2 = u_{in} = -4$

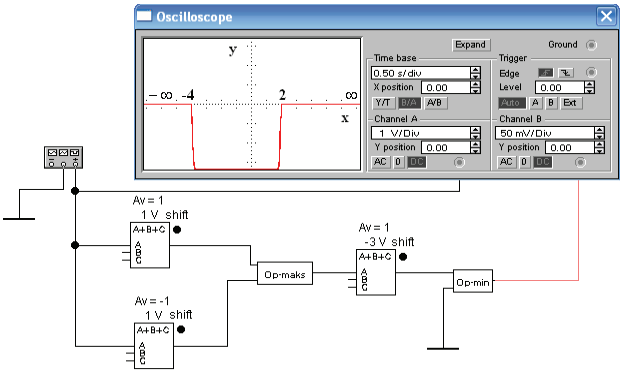


Figure 20. Electronic solution for the inequality $|x+1| \geq 3$;

$$(-\infty,-4] \cup [2,\infty)$$

B. *Trigonometry*

Trigonometric identity, relating circular functions of sine and cosine, i.e. $\cos^2 x + \sin^2 x = 1$ can be interpreted in hardware *min-max* environment as a line having a unity length as it is shown in Fig. 21.

C. *Dead zone*

In a *dead zone* operation (see Fig. 22.), the output is typically a linear function of the input, except for a band that is insensitive to the input. For example, input v_{in},

$$v_{out} = \max(m_2 v_{in} + c_2, \min(m_1 v_{in} + c_1, 0)) \qquad (10)$$

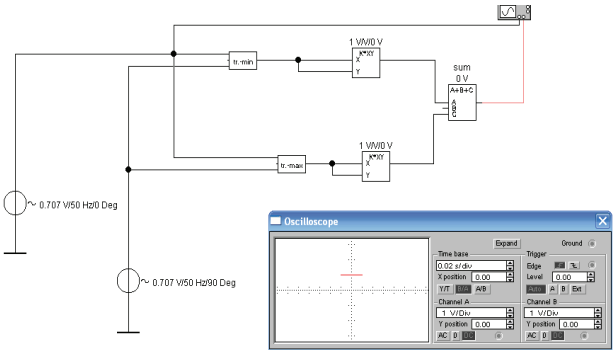


Figure 21. Basic trigonometric relation (identity) interpreted in *min-max* hardware environment; $[\min(\sin x, \cos x)]^2 + [\max(\sin x, \cos x)]^2 = 1$

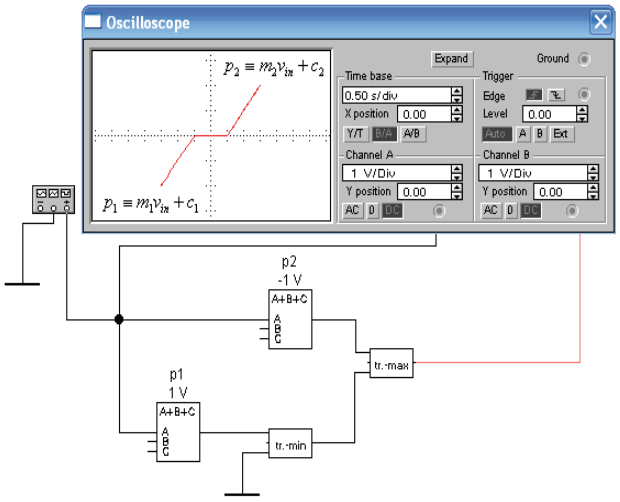


Figure 22. *Dead zone* for where $m_1 = m_2 = c_1 = 1$ and $c_2 = -1$

$$p_1 \equiv m_1 v_{in} + c_1 \quad p_2 \equiv m_2 v_{in} + c_2$$

D. *Membership function generator*

A fuzzy set may be represented by a membership function. This function gives the grade (degree) of membership within the set, of any element of the universe of discourse.

The membership function maps the elements of the universe on to numerical values in the interval [0,1]. Specially,

$$\mu_Z(x) : X \rightarrow [0,1]$$

where $\mu_Z(x)$ is the membership function of the fuzzy set *Z* in the universe in *X*.

Membership function can have distinct shapes. The most popular is a „tent function“. A „tent“ function can be generated using *min-max-circuit technique* as it is illustrated in Fig. 23.

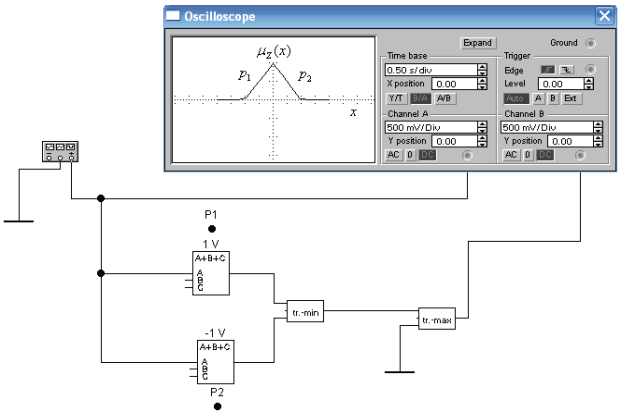


Figure 23. A „tent“ function (membership function) can be expressed as

$$\mu_Z(x) = maks(0, \min(p_1, p_2))$$

where: $p_1 \equiv \mu_Z(x) = x + 1$ and $p_2 \equiv \mu_Z(x) = -x + 1$

E. *Nonlinear transfer function synthesis*

Previously we showed so called *four-part multiagent diagram* representing NDSP [4]. The main role in that multiagent system plays a local operator. His (her) activities in that agent society can be briefly described by a simple rule base (see Fig. 24. for graphic illustration):

Rule No.1. If a current remnant is below the lower limit $R_{l.l.}(k)$, *then increase the current order*

Rule No.2. If a current remnant is between the lower limit $R_{l.l.}(k)$ *and the upper limit* $R_{u.l.}(k)$, *then do not change the current order*

Rule No.3. If a current remnant is over the upper limit $R_{u.l.}(k)$, *then increase the current order*

Rule No.4. If the current order has been changed, then change the current operator's transfer characteristic (see Fig. 18.)

Operator's reaction (*Rule No.1. or Rule No.3.*) is typically a linear function of the input what means that lines p^+ and p^- represent the reactive part of operator's transfer characteristic (OTC).

Hardware modeling starts with characteristic line generating (see Fig. 25a.). Then *max* and *min* circuits are added to synthesize OTC (see Fig. 25b.).

F. *Min/max circuits as theorem provers*

Absolute value is any maping $|u|$ from a domain D to the real numbers R satisfying four fundamental properties among which subadditivity $|u+v| \le |u| + |v|$ can be proven using *Op-Amp* technique (see Fig. 26.).

VIII. MIRKO POLJAK, OCJENA NA HRVATSKOM.DOCXMIN & MAX CIRCUITS AS ELEMENTS OF A COMPLEX GATE (AGENT COORDINATION DETECTOR)

One of the major problems that arise in cooperative activity in NDSP is that of inconsistencies between local operator (agent) and inspector [4]. Wholesaler agent must establish negotiation platform once when two simultaneous, equal or unequal orders are detected in the ordering process and in the case when both expected orders are missing.

In the hardware model of NDSP, an "exclusive OR gate" based on *min-max-Op-Amp technique* (see Fig. 27.) can be used for that purpose. The circuit will deliver zero output voltage as indication (trigger) that two non-zero orders are present simultaneously.

In the case when $n_A = n_B = 0$, the same coordination mechanism is going to be triggered again as the output of the gate $Y = 0$ (see gate dynamics shown in Fig. 28.).

Note than comparators used in the design of the "exclusive OR gate" are designed in *min-max-Op-Amp technique* as it is shown in Fig. 29. Three "input lines" are combined on *min-max functions* platform to design a complex circuit having "standard" comparator transfer characteristic (see Fig. 29.).

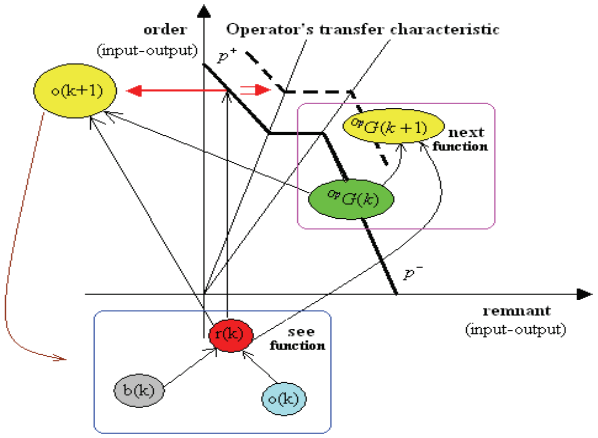
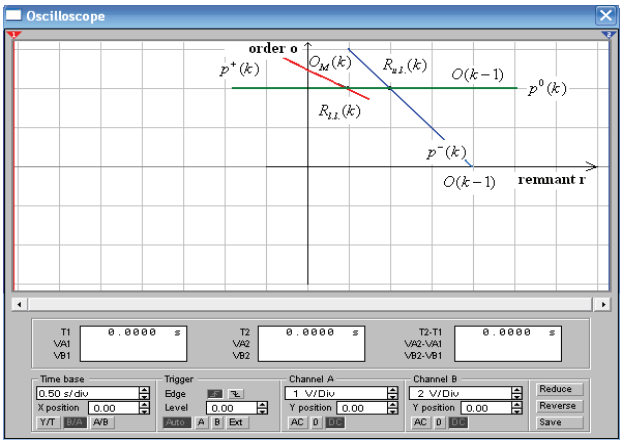
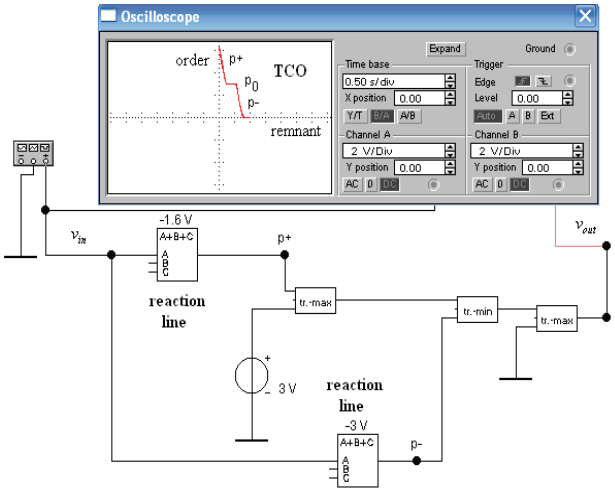


Figure 24. Two implications can be recognized in operator's activities: $r(k) = \max(o(k) - b(k), 0) \Rightarrow o(k+1)$, $o(k+1) \Rightarrow {}^{Op}G(k+1)$



a)"premises"(rules); control rule base $R_{base}(k)$



b)"conclusion"; transfer characteristic of an operator synthesizes existing control rule base

Figure 25. Hardware model of the control rule base is based on a single rule:
If a control rule base about distribution and sale of a newspaper exists then it is possible to transform that rule base in a nonlinear transfer characteristic of an operator ${}^{O}G(k)$, i.e.
${}^{O}G(k) = maks(0, \min(p^-(k), \max(p^0(k), p^+(k))))$

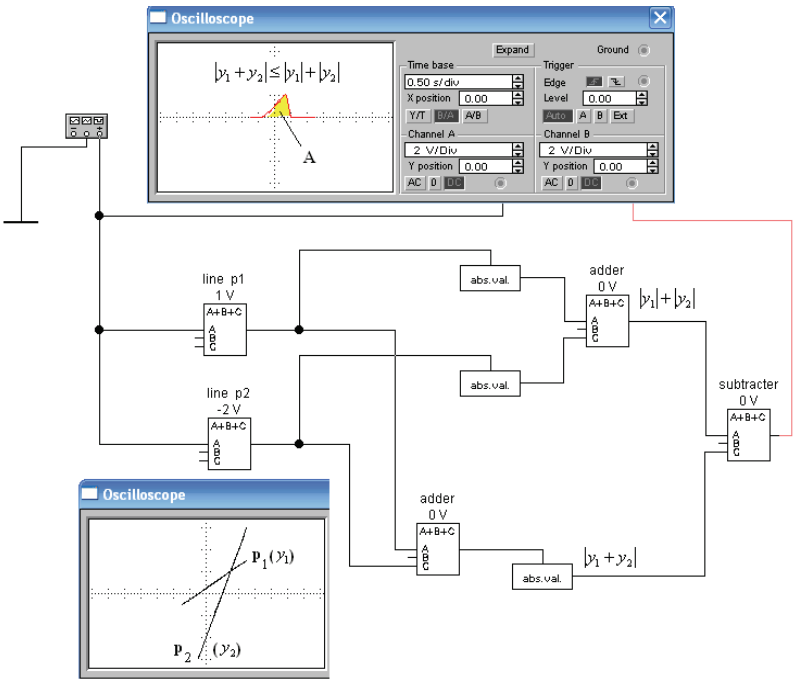


Figure 26. In the A-area, y-coordinates of lines p_1 and p_2 are of opposite signs making subtracted response (to inputs) positive; inequality $|y_1|+|y_2| \geq |y_1+y_2|$ holds

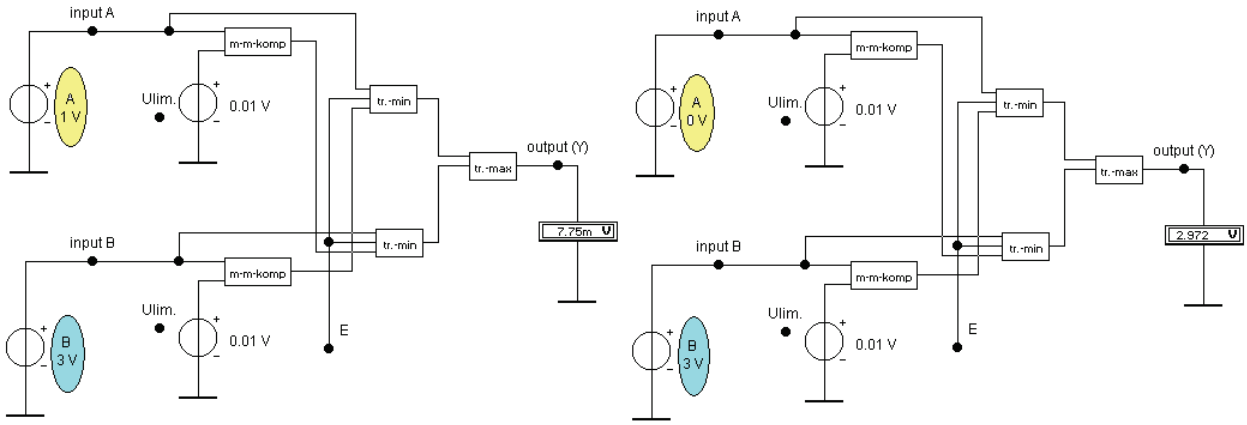


Figure 27. "Exclusive OR gate" can serve as a coordination detector; if two orders $n_A \neq 0$ and $n_B \neq 0$, from two agent are imposed simultaneously, then a kind of coordination mechanism must be established to solve the emergent problem.

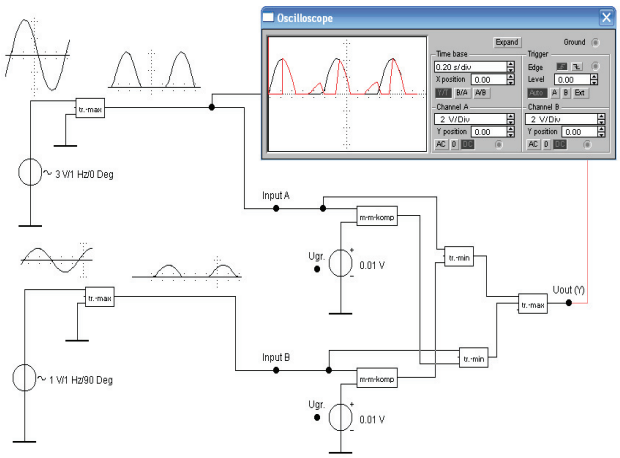


Figure 28. Simulation of the "exclusive OR gate" dynamic response

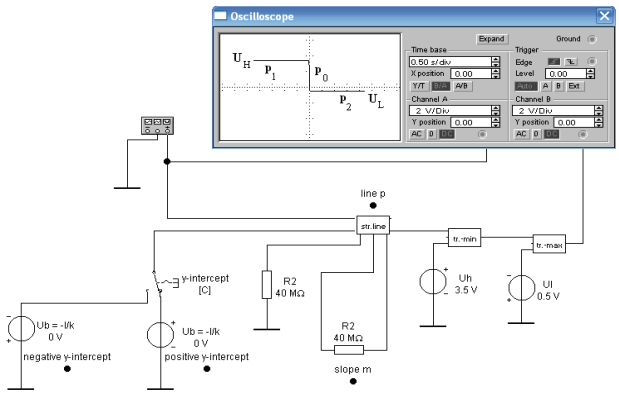


Figure 29. *Min-max* comparator ("m-m-komp");
$u_{out} = \max(U_L, \min(U_H, p_0))$
where $p_0 \equiv u_{out} = m\, u_{in}$ and $m >>$

IX. Conclusion

In this paper (lesson) we have presented only of a part of *min-max* applications that have been developed in the electronic laboratory of Maritime faculty Split as it can be seen from the list below:

- original *min* and *max* circuits
- ***min-max* clipping**
  - **positive clipping circuit**
  - **negative clipping circuit**
  - **positive and negative clipping circuit**
  - **negative peak pass circuit**
  - **positive peak pass circuit**
  - **inverted positive and negative clipping circuit**
    - **D-operators**
    - four-part multiagent diagram
- half wave and full wave rectifiers
- ***clamping circuits***
- nonlinear transfer function synthesis
- **membership function generator(s)**
- ***zero crossing detector***
- ***active peak detector***
- ***simple sort***
- ***complex sort***
- AI algorithm(s)
  - human activity hardware models
  - hardware models of coordination in a multiagent society
- ***min-max circuits in digital environment***
- *min-max* circuits in mathematical environment
  - inequality solving tool
  - theorem solving (proving) tool
  - trigonometry solving tool
  - ***phase plain partition***
- ***small expert systems***
- ***hardware models of syllogisms***

Shown applications are marked (bold). In this presentation "painted applications" have been presented partly ("blue ones"), while "red ones" (in italic) are still waiting to be present in a future.

In our opinion future work has to go in two directions;

- more practical applications that can be used in electronic practice and production
- development of new AI algorithms in the field of coordination

At the very end, it is possible to conclude that *min-max* circuits coming out from our laboratory may have a very perspective future as a problem solving tool in a different technical areas.

References

[1] Theodore F. Bogart, Jr., Electronic Devices and Circuits, Prentice-Hall, Inc., Upper Saddle River, New Jersey, USA, 1997, pp.696-707.

[2] D. Kovačević et al, MIN and MAX Circuit Applications; Function Fitter, MIPRO 2010, 33rd International Convention on Information and Communication Technology, Electronics and Microelectronics, May 24-28, 2010, Opatija, Croatia, pp. 181-186.

[3] D. Kovačević, A. Kovačević, N. Pribačić, Electronics In Mathematics; Idea About a Different Lesson in a Different Classroom, 10th International Educational Technology Conference, Proceedings Book Volume 1, Istanbul, Turkey, May 2010, pp. 496-501.

[4] D. Kovačević, Hardware Modeling of Newspaper Distribution and Sell Process Control Based on Intelligent Agent Network Realization, PhD Thesis, Faculty of Electrical Engineering, Mechanical Engineering, and Naval Architecture, University of Split, Split, Croatia, July, 2012.

[5] D. Kovačević, Sklopovske realizacije elemenata i funkcija teorije neizrazitih skupova, Magistarski rad, FESB – Split, listopad 1992.

[6] D. Kovačević, A. Kovačević, Fuzzy Logic Controller Based on Standard Operational Amplifiers in Fuzzy Logic, Edited by J.F. Baldwin, John Wiley & Sons, Chichester, England, 1996., pp. 231-244.

[7] en.wikipedia.org/wiki/Max-plus algebra, Max Plus Algebra

[8] P. Butkovič, Max-algebra: the linear algebra of combinatorics, Linear algebra and its applications 367, Elsevier, Holland, 2003, pp. 313-335.

[9] en.wikipedia.org/wiki/tropical_geometry, Tropical Geometry, 11.3.2012.

[10] M. Gavalec, Computing matrix period in max-min algebra, Discrete Applied Mathematics, Volume 75, Issue 1, May 16, 1997, pp. 63-70.

[11] S. Kurepa, Uvod u matematiku, Tehnička Knjiga Zagreb, Croatia, May 1975, pp. 109-140.

PAPERS

Tuning Electronic Properties of Molybdenum Disulfide by a Substitution in Metal Sublattice

V.E. Fedorov[*/**a], N.G. Naumov[*], A.N. Lavrov[*], M.S. Tarasenko[*], S.B. Artemkina[*], A.I. Romanenko[**] and M.V. Medvedev[***]

[*]Nikolaev Institute of Inorganic Chemistry, Siberian Branch of Russian Academy of Sciences, 3, Akad. Lavrentiev Ave., Novosibirsk 630090, Russia
[**]Novosibirsk State University, 2, Pirogova st., 630090, Russia
[***]Samsung Advanced Institute of Technology, Suwon, South Korea
[a]e-mail: fed@niic.nsc.ru

Abstract - The electronic properties of molybdenum disulfide have been modified by substitution of molybdenum atoms with non-isovalent Re and Nb atoms. Similar approach results in preparation of semiconducting $Mo_{1-x}Re_xS_2$ and semimetallic $Mo_{1-x}Nb_xS_2$ materials. These solid solutions were exfoliated in organic solvents. Nano size colloidal dispersions were used for preparation of thin films. All materials are characterized by a set of modern methods.

I. Introduction

Layered dichalcogenides of transition metals MQ_2 (M= Ti, Zr, Nb, Ta, Mo, W; Q = S, Se, Te) have been extensively studied for more than three decades [1], but their electronic nature remain a matter of debate to this day [2, 3]. Remarkable electronic properties of dichalcogenides allow use some of them as photocatalysts, battery cathode materials including solar energy conversion units, optoelectronic sensors etc.

Molybdenum disulfide (Figure 1), most typical representative of this family, is a very interesting material with numerous applications. Recently it was shown that single-layer MoS_2 can be used for preparation of field-effect transistor and for other nanoelectronic applications [4-10]. Therefore the electronic properties of MoS_2 are of large interest.

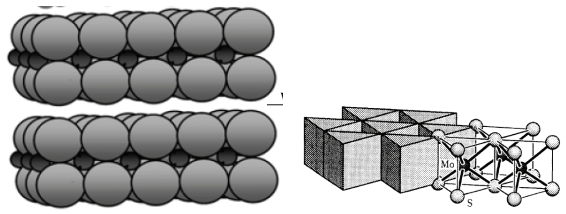


Figure 1. Structure of hexagonal modification of molybdenum disulfide 2H-MoS_2

In present work we made an attempt to modify electronic properties of MoS_2 by non-isovalent substitution of metal atoms.

II. Experimental

A. *Synthesis of MoS_2 and $Mo_{1-x}M_xS_2$ solid solutions*

A series of samples of compositions $Mo_{1-x}M_xS_2$ (M=Nb, Re; $0<x<0.15$) with 2H-MoS_2 structure type was synthesized via ampoule high-temperature method. High purity elements were used for synthesis. Before synthesis powder metals were annealed in hydrogen at 1000°C during 1 hour for removing adsorbed water and traces of oxides. To reach homogeneity the samples were regrinded two times and sealed again. The final thermal treatment was carried out at 1100°C/7 days. According XRD analysis the samples were single phases with some broadening of reflections.

B. *Conductivity measurements*

Conductivity measurements of the samples were carried out on pressed tablets using standard four-probe technique.

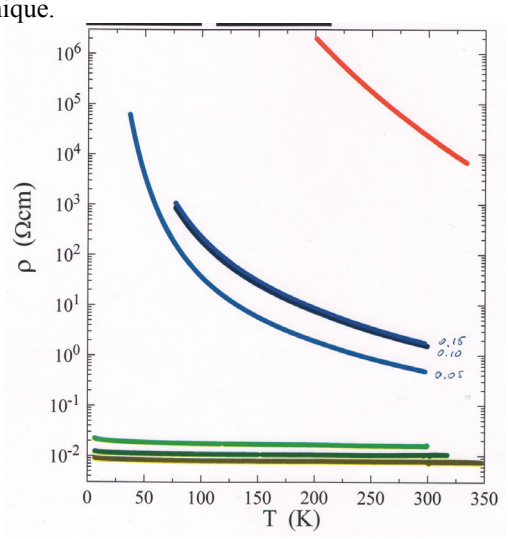
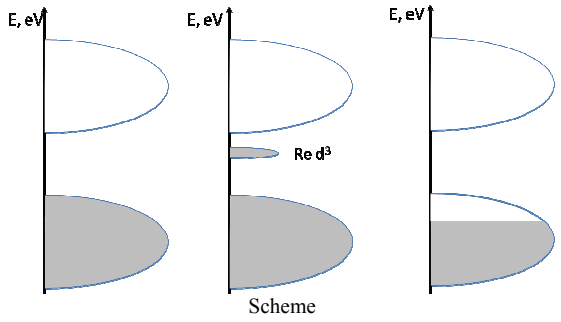


Figure 2. Temperature dependence of resistivity of MoS_2 (red curve) and solid solutions $Mo_{1-x}Re_xS_2$ (blue curves) and $Mo_{1-x}Nb_xS_2$ (green curves)

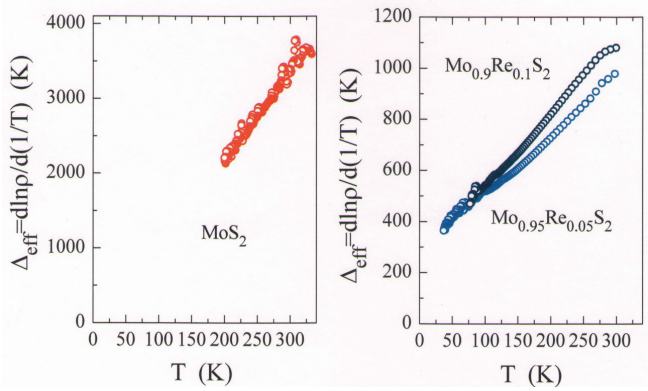
The measurement of temperature dependence of resistivity of undoped MoS_2 shows that in the temperature range 325-200 K resistivity grows exponentially from c.a. 10^4 to 10^7 Ohm·cm (Figure 2, red curve).

This increase can be described as typical semiconductor behavior with substantial gap between valence and conductivity bands (Δ_{eff} ~3500 K at 300 K (0.3 eV)) (Red curve on Figure 2; Scheme, a). Plot $\Delta_{eff} = d\ln\rho/d(1/T)$ demonstrates nearly linear dependence of Δ_{eff} on temperature (Figure 3, a). The value of Δ_{eff} is lower than that for single crystal MoS_2 which is semiconductor indirect-gap with a band gap (c.a. 15 000 K) but is typical for polycrystalline MoS_2 due to numerous stacking defects.



Scheme

Substitution in MoS_2 of Mo atoms having electronic configuration d^2 by electron deficient Nb atoms having electronic configuration d^1 in MS_2 phase (Figure 2, green curves) and electron-rich Re atoms with electronic configuration d^3 in MS_2 phase (Figure 2, blue curves) shows distinct differences in influence of substitution on transport properties.



a). b).

Figure 3. Temperature dependence of Δ_{eff} for MoS_2 (red curve) and solid solutions $Mo_{1-x}Re_xS_2$ (blue curves)

Sample $Mo_{1-x}Re_xS_2$ with x = 0.05 reveals significant drop in the resistance of sample (by c.a. 10^4 times) and decrease of Δ_{eff} in 4 times down to 1000 K (0.086 eV) (Figure 3, b). Such behavior can be preliminary explained as appearance of additional sub-level of Re d^3 states between valence and conductivity bands that is defining Δ_{eff}. (Scheme, b).

Further substitution (samples with x = 0.10 and 0.15) slightly increase resistance of samples being samples with x = 0.10 and 0.15 nearly identical from transport properties point of view. The ratio $\rho(Mo_{0.85}Re_{0.15}S_2)/\rho(Mo_{0.90}Re_{0.10}S_2)$ remains within 1.12-1.20 in whole temperature range. Such dependence can be explained by tendency of Re to form metal clusters, i.e. formation of covalent metal-metal bonds with strong localization of valence electrons and excluding of them from transport. For example, in contrast of MoS_2 with equidistant Mo-Mo, the phase ReS_2 consists of rhombic Re_4 clusters. Thus, all samples $Mo_{1-x}Re_xS_2$ (x = 0.05, 0.10 and 0.15) demonstrate semiconducting behavior with exponential temperature dependence of resistivity.

Substitution of Mo (d^2) by electron-deficient Nb (d^1) in $Mo_{1-x}Nb_xS_2$ effects on transport properties in different manner. Figure 2 (green curves) indicates that the samples with (x = 0.05, 0.10 and 0.15) show metallic-like dependence resistivity with weak increase of ρ down to helium temperatures (from 300 to 4K). Such behavior indicates that insertion of electron-deficient Nb in lattice MoS_2 produces the holes in Mo d^2 conductivity band.

C. Exfoliation of bulk layered solid solutions $Mo_{1-x}Nb_xS_2$ and $Mo_{1-x}Re_xS_2$ (x=0.05, 0.10, 0.15) into colloidal dispersions

It is known several methods for preparation of very thin sheets from bulk layered materials, but they are not numerous. Scotch tape method allows obtain high quality layers but it has some disadvantages: it is labor-consuming, requiring an availability of single crystal starting materials and does not allow preparing the samples of large scale. One of promising techniques for exfoliation of layered bulk materials is chemical method in which bulk material is dispersed into colloidal solutions using proper solvents. From such colloidal dispersions the films with wishful thickness can be obtained by simple filtering through fine filter or by spray technique. Similar approach was applied successfully for preparation of many layered materials including transition metals dichalcogenides [11-12].

Preparation of colloidal dispersions of $Mo_{1-x}Nb_xS_2$ and $Mo_{1-x}Re_xS_2$ was carried out as following.

1 g of the solid solutions $Mo_{1-x}M_xS_2$ was sonicated in an ultrasonic bath with 400 ml of polar organic solvents (dimethylformamide, n-methyl pyrrolidone and others), for 7 h; the dispersion was left to settle down for 12 h and then centrifuged at 2500 rpm for 20 min. According to the dynamic light scattering (DLS) measurements, average lateral size of the particles in all different dispersions is roughly the same (~60 – 260 nm) (Figure 4).

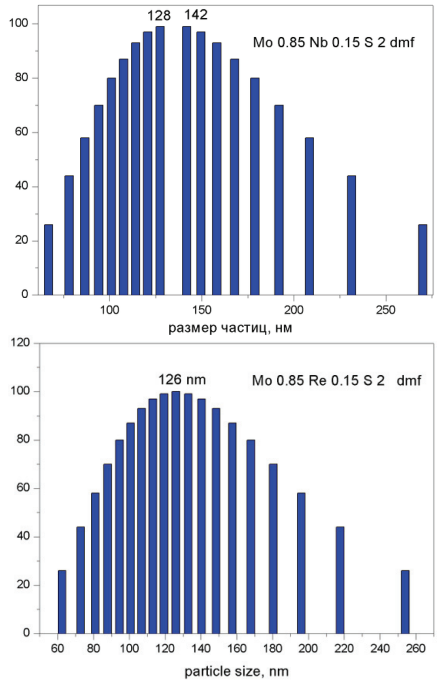


Figure 4. Particle size distribution in DMF dispersions from DLS data

UV-vis spectra registered for DMF dispersions (Figure 5) are identical with two bands at about 620 and 674 nm for $Mo_{1-x}Nb_xS_2$ and 627 and 682 nm for $Mo_{1-x}Re_xS_2$. Both these bands are moved slightly compared with pure MoS_2 sample (610 and 670 nm). The results obtained show that substitutions of molybdenum atoms in cation sublattices of transition metal dichalcogenides give visible changes in their electronic spectra. These data are in accordance with electro-physical properties of the samples.

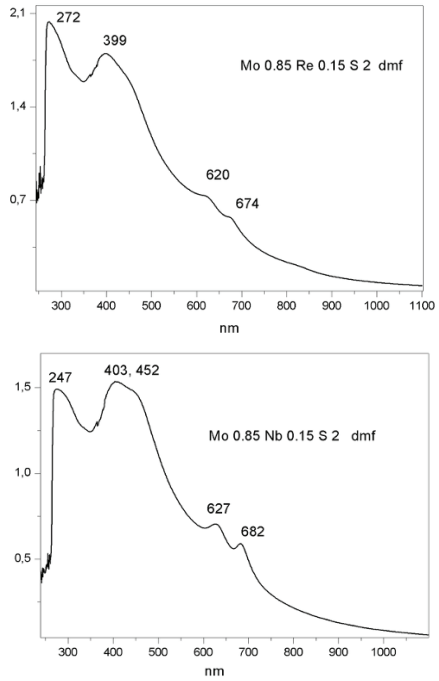


Figure 5. UV-vis spectra of DMF dispersions of $Mo_{0.85}Re_{0.15}S_2$ (above) and $Mo_{0.85}Nb_{0.15}S_2$ (below)

D. Films of layered doped dichalcogenides prepared from colloidal dispersions

Two methods for film preparation were used: *i*). filtration of colloidal dispersions through membrane filter "Whatman anodisc", the pore values is 0.02 μm and *ii*). spray method using colloidal dispersions. Thickness of films can be regulated by a volume of strained liquid dispersion with fixed concentration. In our experiments the film thicknesses were in the interval of 1-5 μm. After soft annealing at 100-200°C the films look as quite smooth surface (Figure 6).

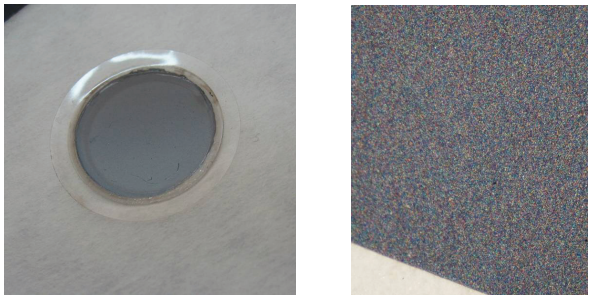


Figure 6. Photograph and optical microscope image of thin films of $Mo_{0.85}Re_{0.15}S_2$

It is important to note that diffractograms for all films are very similar: they demonstrate high degree of texturing that is developed in presence of *00l* reflexes (Figure 7). This is not unexpectedly because flat particles of dispersion should packing up exactly on filter.

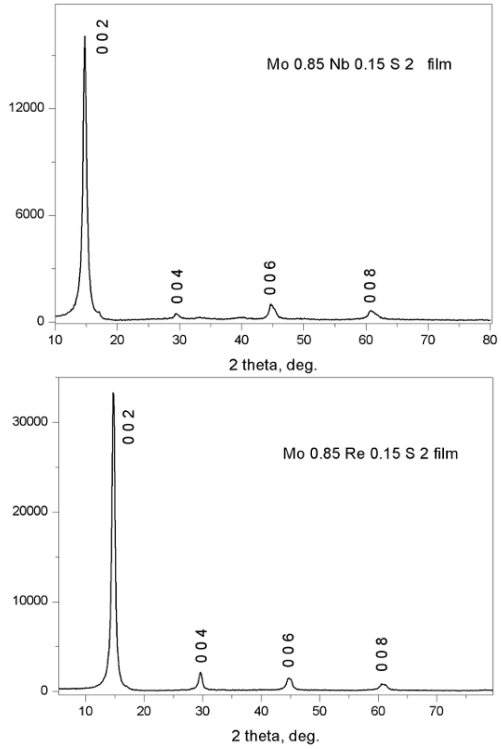


Figure 7. XDR pictures of the films obtained from DMF dispersions of solid solutions $M_{1-x}M'_xQ_2$: $Mo_{0.85}Nb_{0.15}S_2$ (above) and $Mo_{0.85}Re_{0.15}S_2$ (below)

E. Temperature dependence of the resistivity of the films $Mo_{1-x}Nb_xS_2$ and $Mo_{1-x}Re_xS_2$

The temperature dependences of electrical resistivity of the niobium- and rhenium-doped films prepared by spray method are shown in Figure 8. As may be seen from these data, the resistivity of $Mo_{0.85}Nb_{0.15}S_2$ film is invariable practically in the wide temperature interval but it is increased at lower 50K; the resistivity of $Mo_{0.85}Re_{0.15}S_2$ has typical semiconductor behavior, it is greatly increased upon decreased in temperature. The value of band gap *Eg* for $Mo_{0.85}Re_{0.15}S_2$ calculated from temperature dependence of resistivity is close to 1.3 eV.

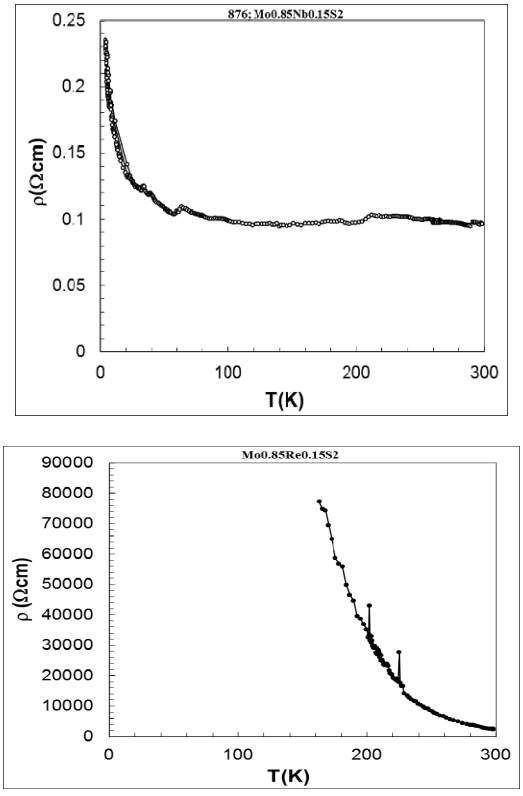


Figure 8. Temperature dependence of the resistivity ρ(T) for $Mo_{0.85}Nb_{0.15}S_2$ (above) and $Mo_{0.85}Re_{0.15}S_2$ (below) films prepared by spray method

III. CONLUSION

For fine tuning transport properties of semiconducting $Mo_{1-x}Re_xS_2$ (Δ_{eff}) and semimetallic $Mo_{1-x}Nb_xS_2$ solid solutions the most perspective materials are compositions with small value of *x*, because this concentration interval gives the materials having wide diapason of transport properties.

ACKNOWLEDGMENT

The study was supported by Samsung Advanced Institute of Technology, Suwon, Republic Korea (GRO Project), and Ministry of Education and Science of Russian Federation (Project-Agreement 8030).

REFERENCES

[1] J.A. Wilson, A.D. Yoffe, "The transition metal dichalcogenides: discussion and interpretation of the observed optical, electrical and structural properties", Adv. Phys., 1969 , vol. 18 , pp. 193-335.

[2] K.S. Novoselov, D. Jiang, F. Schedin, T.J. Booth, V.V. Khotkevich, S.V. Morozov, and A.K. Geim, "Two-dimensional atomic crystals". Proc. Natl. Acad. Sci. U. S. A. 2005, vol. 102, pp. 10451−10453.

[3] H.S.S.R. Matte, A. Gomathi, A.K. Manna, D.J. Late, R. Datta, S.K. Pati, and C.N.R. Rao, "MoS_2 and WS_2 analogues of graphene*",* Angew. Chem., Int. Ed. 2010 , vol. 49, pp. 4059-4062.

[4] B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, and A. Kis, "Single-layer MoS_2 transistors", Nature Nanotechnology, 2011, vol. 6, pp. 147−150.

[5] B. Radisavljevic, M.B. Whitwick, and A. Kis, "Integrated Circuits and Logic Operations Based on Single-Layer MoS_2", ACS Nano 2011, vol. 5, pp. 9934−9938.

[6] B. Radisavljevic, M.B. Whitwick, and A. Kis, "Small-signal amplifier based on single-layer MoS_2", Applied Physics Letters, 2012, vol. 101, pp. 043103 (4).

[7] S. Bertolazzi, J. Brivio, and A. Kis, "Stretching and Breaking of Ultrathin MoS_2", ACS Nano, 2011, vol. 5, pp. 9703–9709.

[8] A. Splendiani, L. Sun, Y. Zhang, T. Li, J. Kim, C.-Y. Chim, G. Galli, and F. Wang, "Emerging Photoluminescence in Monolayer MoS_2", Nano Lett. 2010, vol. 10, pp. 1271−1275.

[9] K.F. Mak, C. Lee, J. Hone, J. Shan, and T.F. Heinz, "Atomically thin MoS_2: a new direct-gap semiconductor", Phys. Rev. Lett., 2010, vol. 105, pp. 136805-136808 .

[10] Q. H. Wang, K. Kalantar-Zadeh, A. Kis, J.N. Coleman and M.S. Strano, "Electronics and optoelectronics of two-dimensional transition metal dichalcogenides", Nature Nanotechnology, 2012, vol. 7, pp. 699-712.

[11] J. N. Coleman*, et al., "*Two-Dimensional Nanosheets Produced by Liquid Exfoliation of Layered Materials", *Science* 2011, vol. 331, pp. 568-571.

[12] Kai-Ge Zhou et al., "A Mixed-Solvent Strategy for Efficient Exfoliation of Inorganic Graphene Analogues", *Angew. Chem*., 2011, vol. 123, pp.1-5.

Alternating Current Electrochemical Anodisation of Silicon on Insulator Layers

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

[1] Ruđer Bošković Institute, Division of Materials Physics, Bijenička cesta 54, Zagreb, Croatia
[2] University of Zagreb, School of Medicine, Department of Physics and Biophysics, Šalata 3b, Zagreb, Croatia
[3] National Institute of Chemistry, Hajdrihova 19, Ljubljana, Slovenia
*ivanda@irb.hr

Abstract - Porous silicon (PSi) samples were prepared by electrochemical anodisation of silicon on insulator layers. Structural and optical properties of prepared samples were investigated by Raman and photoluminescence (PL) spectroscopy and field emission scanning electron microscopy (FE-SEM). The anodisation of silicon on insulator layers was performed by alternating currents (AC) of the mains frequency of 50 Hz. A very intensive PL was observed at the circular edge of the samples that exhibited micrometer sized island-like porous structure, while the central part of the samples showed moderate PL signal. The formation of such porous island-like structures with strong intensity PL was interpreted with stress induced due to difference of the piezoelectric coefficient of silicon and quartz layers (buried SiO_2). Micro-Raman spectra of islands show strong phonon confinement with the cluster size between 1.4 and 3.5 nm.

I. INTRODUCTION

Porous silicon (pSi) was established as a promising optoelectronic material since Canham's discovery of intense pSi PL at room temperature in the visible light spectrum in 1990[1]. However, bulk silicon exhibits up to five orders of magnitude weaker PL than pSi as a material with indirect energy band-gap[2]. Several models have been proposed to describe the PL in pSi. The most promising candidates were quantum confinement models[3], but some evidence suggest that luminescence properties of pSi may not simply be due to quantum confinement, but also due to surface chemistry effects[4].

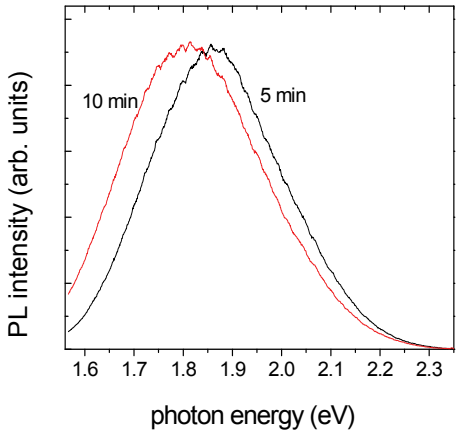
A large amount of work demonstrates that the surface of freshly prepared pSi is hydrogen passivated. However, exposure of the porous silicon to ambient air for prolonged period leads to the slow oxidation of pSi films[5–8]. The presence of silicon oxide on pSi surface stimulates the shift of the PL spectra to shorter (blue) wavelength range[9,10]. The structural data for the pSi are more complicated. Some published works indicate that the pSi layer consists of Si columns and pores or isolated nanocrystallites[11]. On the other hand, pSi may be considered as system of interconnected quantum wells, so-called "quantum sponge"[10].

Silicon can be electrochemically etched in hydrofluoric acid (HF) solutions and at specific current densities, local dissolution of silicon atoms occurs, such that low density porous crystalline silicon is formed. Here we show the results of preparation of porous silicon samples from n-type silicon on insulator layer grown on more conductive n-type silicon substrate. By varying the concentration of 48% HF in ethanol solution micro- and nano-pores of different sizes within the epitaxial layer in dependence on HF concentration were obtained. Etching of silicon-on-insulator (SOI) wafers is of scientific and technological challenge due to the nature of etching process itself, as well as due to the possible application of such silicon in development of new sensors.

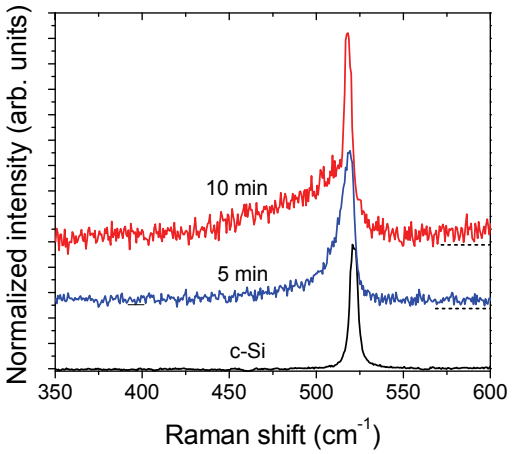
II. EXPERIMENTAL

The commercially available silicon on insulator (SOI) wafers consisting of 40 μm thick device layer with resistivity ρ=20 Ωcm, separated from the 460 μm thick handle with resistivity ρ=10 Ωcm, with 1.5 μm thick SiO_2 layer. The active layer and the substrate has

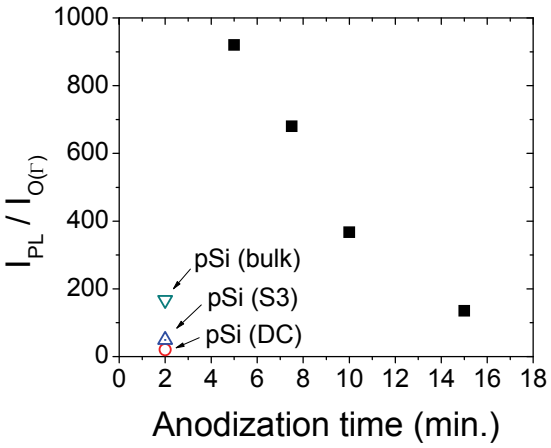
{110} and {100} orientation, respectively. The anodisation process was carried out in an ethanoic solution of hydrofluoric acid with the volume concentration 1:1, by using different AC current densities and anodisation time. The samples S_1(AC) and S_2(AC) were anodised using the current density of 13 mA/cm^2 and 15 mA/cm^2 for duration of 5 and 10 minutes, respectively. The samples S_3(AC) and S_4(AC) were anodised using the current density of 30 mA/cm^2 during 15 and 7.5 minutes, respectively. Following the anodisation process, the electrolyte from the anodisation cell was exchanged with ethanol. After anosidation all samples were once more rinsed in ethanol and twice in pentane. Final pentane rinse was continued for 5 minutes in order to reduce the capillary stress during sample drying, since pentane has a very low surface tension. Pentane was removed and samples were left to dry in air. Porous poly-Si surface morphology and cross-section were explored with FE-SEM using JEOL JSM 7000F field-emission FE-SEM. Photoluminescence and Raman spectra in micro-Raman mode were recorded with Jobin Yvon T64000 triple spectrometer using excitation from 514.5 nm wavelength argon-ion laser. The special care was taken to avoid the laser heating effects on the O(Γ) silicon Raman mode. For this reason the laser beam power at the surface of the sample was kept at 1 mW and in special cases of recording Raman and PL spectra from silicon islands, the laser power on the sample was 0.01 mW. Raman and photoluminescence spectra were recorded from the same spot at the top of the porous islands.



a)



b)



c)

Figure 1. Photoluminescence spectra recorded from the porous island at the edge of the circular adnodisated region of the samples S_1(AC) and S_2(AC) with the peak value at 1.87 and 1.82 eV (a); the corresponding Raman spectra with the spectrum of crystalline silicon for comparison (b); dependence of the ratio of peak intensity of the PL band and O(Γ) phonon mode in dependence on the anodisation time.

III. Results and Discussion

Figure 1a) shows a very strong PL bands recorded from the porous islands found on the edge of circular anodisated region on the sample prepared by with mains frequency AC (50 Hz). The samples S_1(AC) and S_2(AC) were anodised under the current density of 13 mA/cm^2 and 15 mA/cm^2 in a time of 5 and 10 minutes, respectively. While the intensity of

the PL in the centre of the samples was moderate, a very strong PL was observed on the edge of two centimetres in diameter anodisated region of SOI wafer. The Raman spectra from the same islands shows the red-shifted asymmetric Raman band.

The size of nanostructures estimated from the Raman bands bands are 1.4 and 3.5 nm respectively. Figure 1c) shows the comparison of the observed PL intensity with the best results we obtained by different method of anodisation. For this case the PL signal was normalised on the intensity of O(Γ) phonon band, i.e. to the number of excited Si atoms. It is evident more than 5 times stronger intensity for the samples anodised with AC current.

IV. CONCLUSION

Porous silicon (PSi) samples were prepared by electrochemical anodisation of silicon on insulator layers. The extremely strong PL was observed at the circular edge of the anodised region. It was shown that such PL comes from the micrometer sized island-like porous structures. The moderate PL signal, similar to DC anodised silicon, was observed in the centre of the samples. The intensity of the PL peak recorded from the islands on the edge of anodised regions showed dependence on the anodisation time. We assume that such PL bands of porous island-like structures is connected with the the stress produced during anodisation process. The stress is caused by the difference of the piezoelectric coefficient of thin epitaxial silicon layer and buried quartz layer of SOI.

REFERENCES

1. L. T. Canham, Appl. Phys. Lett. 57, 1046 (1990).
2. Fauchet, P. M., C. C. Tsai, L. T. Canham, I. Shimizu, and Y. Aoyahi, 1993, Microcrystalline Semiconductors: Materials and Devices, vol 283.
3. T. Ohno, K. Shiraishi, and T. Ogawa, Phys. Rev. Lett. 69, 2400–2403 (1992).
4. S. Munekuni, T. Yamanaka, Y. Shimogaichi, R. Tohmon, Y. Ohki, K. Nagasawa, and Y. Hama, J. Appl. Phys. 68, 1212 (1990).
5. E.A. Petrova, Mater. Sci. Eng., B 69–70 (2000) 152.
6. K.H. Beckmann, Surf. Sci. 3 (1965) 314.
7. V.M. Dubin, F. Ozaman, J.-N. Chazalviel, Thin Solid Films 255 (1995) 87.
8. Shindi Takeoka, Phys. Rev., B 62 (24) (2000) 820.
9. Yue Zhao, Derren Yang, Dongsheng Li, Minghua Jianga, Mater. Sci. Eng., B 116 (2005) 95.
10. V. Lehmann, U. Go¨ sele, Appl. Phys. Lett. 58 (1991) 856.
11. Levy-Clement, S. Lust, M. Mamor, J. Rappich, Th. Dittrich, Phys. Status Solidi (a) 202 (2005) 1390.

Porous silicon prepared from p^- and p^{++} bulk silicon

Vedran Đerek[1], Romana Baltić[1], Marijan Marcijuš[2], Mira Ristić[2], Zorica Crnjak Orel[3] and Mile Ivanda[1*]
[1] Ruđer Bošković Institute, Division of Materials Physics, Bijenička 54, Zagreb, Croatia
[2] Ruđer Bošković Institute, Division of Materials Chemistry, Bijenička 54, Zagreb, Croatia
[3] National Institute of Chemistry, Hajdrihova 19, Ljubljana, Slovenia
vedran.djerek@irb.hr

Abstract - Macro and meso-porous silicon (PSi) samples were prepared by electrochemical anodisation of p^- and p^{++} doped silicon wafers in hydrofluoric acid (HF) based electrolyte under galvanostatic conditions. Anodisation time and current were varied in order to obtain different PSi morphologies. Samples were characterized by scanning electron microscopy (SEM), photoluminescence (PL) spectroscopy and Raman spectroscopy. Samples prepared on p^- type substrates were macro-porous, showed visible photoluminescence and no observable phonon confinement in Raman spectrum. Porous silicon prepared on p^{++} type substrates was mesoporous, showed no observable photoluminescence and intense phonon confinement was observed in their Raman spectra. Silicon nanocrystal dimensions in PSi samples obtained from photoluminescence and phonon confinement models were correlated with silicon electrochemical dissolution valence n_d. A surface morphology instability is identified in case where electrolyte is more conductive than the silicon substrate.

I. INTRODUCTION

Porous silicon (PSi) consists of interconnected pores embedded in c-Si matrix. It is typically prepared by electrochemical anodisation of crystalline silicon (c-Si) in hydrofluoric acid (HF) based electrolyte. Silicon dissolution valence n_d can be measured easily by relating the total charge transported during the anodisation with the number of dissolved silicon atoms:

$$n_d = \frac{I \cdot t \cdot A_r}{e \cdot N_A \cdot \Delta m} \quad (1)$$

where I is the anodisation current, t anodisation time, A_r relative atomic mass, e electron charge, N_A Avogadro number and Δm difference in mass before and after anodisation[1].

Electrochemical dissolution of silicon is possible by two electrochemical pathways[1] – for currents below the electropolishing current J_{PS} a two-electron oxidation (n_d=2) of silicon is dominant, producing porous silicon:

$$Si + 6F^- + 2H^+ + h^+ \longrightarrow SiF_6^{2-} + H_2 \quad (2)$$

For current densities larger than J_{PS} a four-electron silicon oxidation process is active (n_d=4), resulting in electropolishing:

$$Si + 6F^- + 4h^+ \longrightarrow SiF_6^{2-} \quad (3)$$

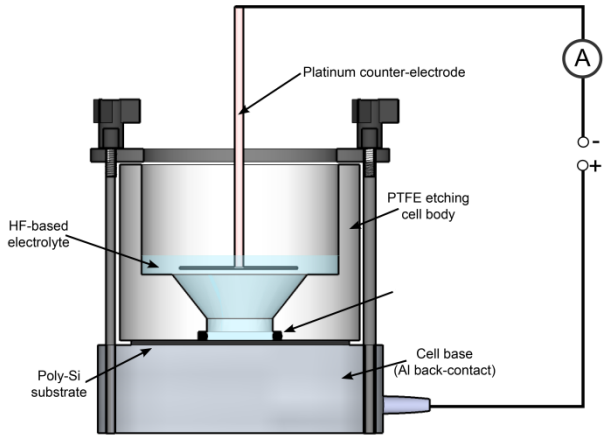


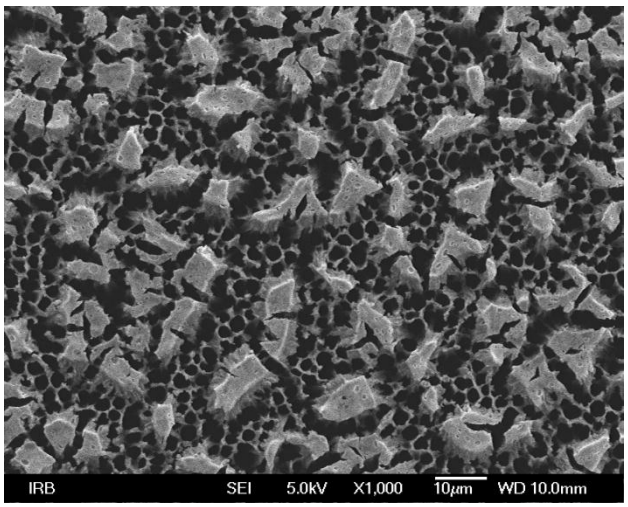
Figure 1. Schematic of experimental setup for preparation of porous silicon.

Pore walls in PSi can be as small as several nanometers, forming sponge-like structures. PSi properties can be engineered by careful control of anodisation parameters such as current density, electrolyte composition and substrate resistivity. Porous silicon can show intensive room temperature photoluminescence in visible spectrum, in contrast to c-Si. Origin of this photoluminescence is a combination of bulk (quantum confinement) and surface effects in silicon nanocrystals present in PSi [2].

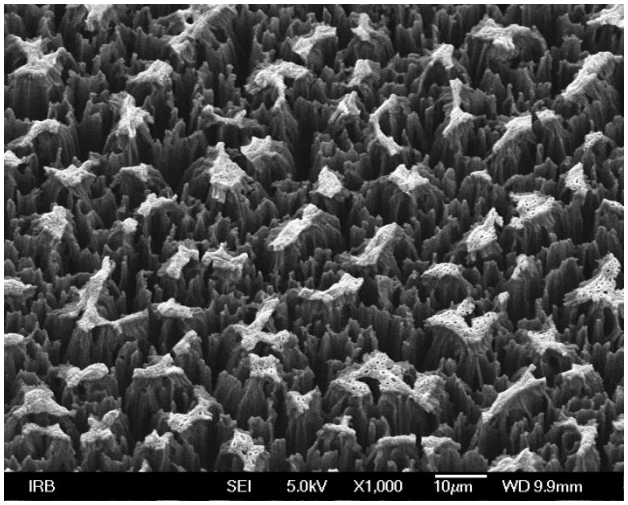
II. EXPERIMENTAL

Porous silicon samples were prepared by electrochemical anodisation of bulk Si wafers. Anodisation was conducted in a hydrofluoric acid (HF) resistant electrolytic cell (Fig. 1) consisting of aluminum base and PTFE cell body. Electrolyte used was aqueous/ethanoic 20% and 5% w/w HF acid prepared by mixing 40% HF with 98% ethanol. Samples were anodized in galvanostatic conditions. Substrates used were boron p^- doped (10 Ω cm < ρ < 23 Ω cm) and p^{++} doped ($\rho \leqslant 0.007$ Ω cm) silicon wafers. High resistivty substrates (p^- doped) were coated with Gallium-Indium eutectic alloy (EGaIn) and painted with SPI conductive silver paint in order to establish ohmic contact to anodisation cell base. Anodisation time and current were varied in order to obtain different PSi morphologies. Bulk Si

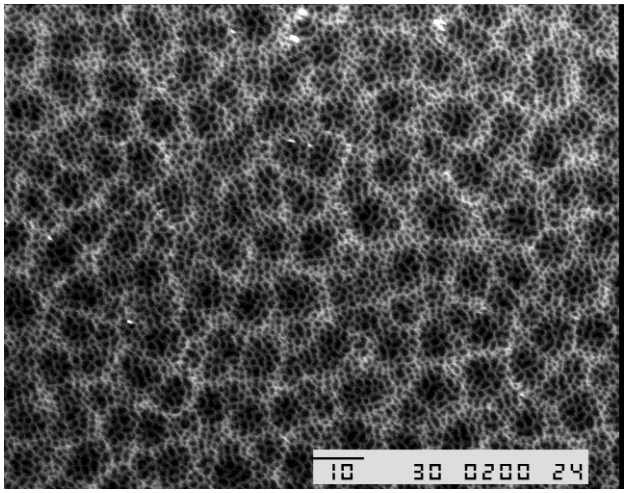
Corresponding author: ivanda@irb.hr



a)



b)



c)

Figure 2. SEM micrographs of PSi prepared from p⁻ doped silicon a) in 20% HF, normal to sample surface b) the same sample tilted 35 degrees c) in 5% HF, normal to sample surface. Magnification x1000 for all micrographs.

substrates were weighed before and after anodisation for measurement of electrochemical dissolution valence n_d.

Samples were characterized by scanning electron microscopy (SEM) using Jeol JSM 7000F field emission SEM and JEOL JSM-T300 SEM, and by photoluminescence (PL) spectroscopy and Raman spectroscopy using Horiba Jobin-Yvon T64000 Raman spectrometer. Photoluminescence and Raman scattering were excited by 514.5 nm argon ion laser line.

A. PSi prepared from p⁻ bulk silicon ("S" samples)

Samples were prepared in 20% and 5% HF acid electrolyte. Samples prepared in 20% HF were anodised at current densities ranging from 10 to 200 mA/cm^2 for 1 to 5 minutes. Surface morphology investigation by SEM showed samples consisted of two layers – a top, cracked one, consisting of islands of material with no pores observable within SEM resolution, and a bottom, macroporous one (Fig. 2a and 2b). Micro-Raman spectroscopy of the samples showed no observable silicon Raman O(Γ) mode shifting or widening. Laser beam power at sample surface was kept at 1 mW during the micro-Raman measurements to avoid sample heating. Photoluminescence spectra were measured in the same setup. Intense photoluminescence was observed from the top, cracked layer, while the bottom, macroporous layer showed no PL. The photoluminescence spectra (Fig. 3) were centered at 1.8-2.1 eV. Silicon nanocrystal sizes can be estimated using the quantum confinement model given in [3,4]:

$$E(d) = E_g + \frac{3.73}{d^{1.39}}(eV) \qquad (3)$$

where E is PL peak energy, d Silicon nanocrsytal diameter and E_g Silicon band-gap energy. For prepared samples quantum confinement model estimates Si nanocrystal diameters ranging from 3,0 to 3,2 nm.

A same type substrate was anodised in 5% HF electrolyte at current density of 5 mA/cm^2 for 120 minutes.

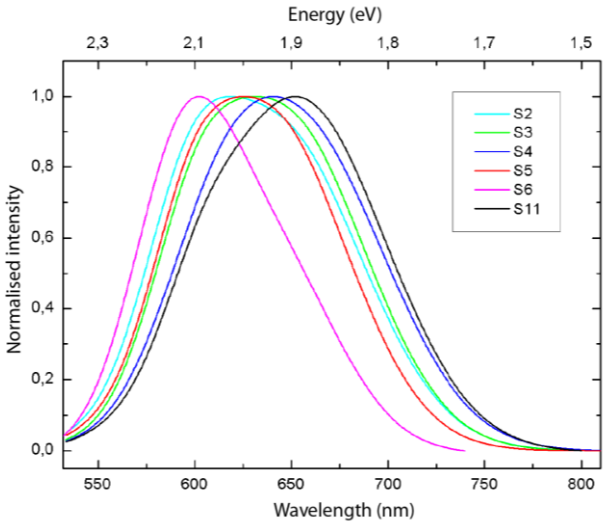


Figure 3. Photoluminescence spectra for PSi prepared from p⁻ doped substrates.

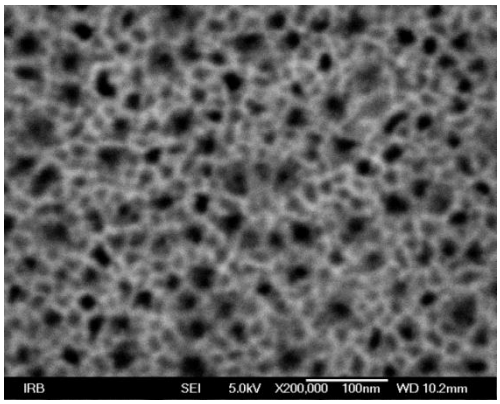


Figure 4. SEM micrograph of PSi prepared from p^{++} doped silicon in 20% HF, normal to sample surface. Magnification x200,000.

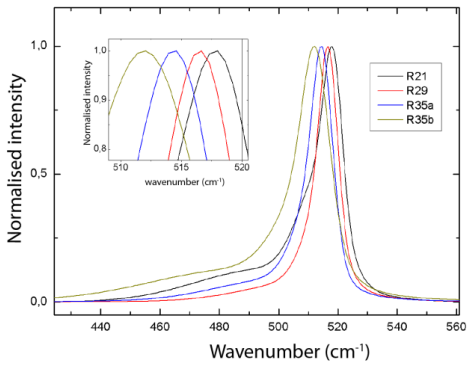
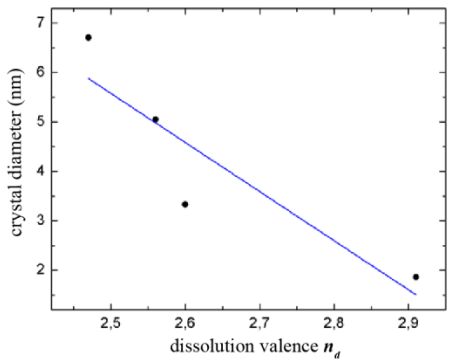
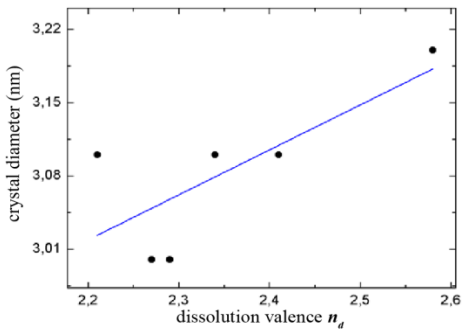


Figure 5. Normalised Raman spectra of PSi prepared from p^{++} doped substrates.



a)



b)

Figure 6. Si nanocrystal size vs. dissolution valence **a)** for *R* and **b)** for *S* samples.

The sample showed homogeneous macroporous surface morphology (Fig. 2c) with no top cracked layer.

Sample showed no observable photoluminescence or Raman silicon O(Γ) mode shift or broadening. The reason for different surface morphologies lies in the ratio of electrolyte and substrate conductivities. As proposed in [5], for homogeneous surface morphology substrate has to be more conductive than the electrolyte, otherwise due to local instabilities large conductive paths form in larger pores, screening the remainder of the bulk material of lower conductivity. Substrate resistivity for *S* samples was 10-23 Ω cm, while 20% and 5% HF electrolytes had resistivities of about 5 Ω cm and 20 Ω cm respectively, so only for 5% HF can the homogeneity condition be fulfilled.

B. PSi prepared from p^{++} bulk silicon ("R" samples)

Samples were prepared in 20% HF acid electrolyte at current densities ranging from 2 to 100 mA/cm^2 for 2 to 30 minutes. Surface morphology was investigated with SEM (Fig. 4), showing closely packed pores of (15±2) nm diameter. All samples showed characteristic asymmetry and red shift in Raman silicon O(Γ) mode (Fig. 5) due to phonon confinement effects [6,7]. Red shifts of 1,3-8,1 cm^{-1} and band widening (ΔFWHM) of up to 1,3-8,1 cm^{-1} were measured. Silicon nanocrystal diameter can be estimated from Raman silicon O(Γ) mode shift using phonon confinement model:

$$\Delta\nu = -A\left(a/L\right)^{\gamma} \qquad (4)$$

where $\Delta\nu$ is Raman mode shift (cm^{-1}), a silicon lattice constant and A and γ empirically obtained values (A=47.41 cm^{-1} and γ=1.44)[6].

For different anodisation parameters nanocrystal diameters ranging from 1,9 to 6,7 nm were estimated. No photoluminescence was observed in *R* samples.

C. Si crystalite diameter vs. dissolution valence

Correlation between Si crystallite diameter obtained from Raman and PL measurements via phonon and quantum confinement models, and dissolution valence n_d was investigated (Fig. 6). For highly doped (*"R"*) samples crystallite diameter apparently decreases with increasing n_d as expected, since for higher n_d electropolishing processes are more intensive, resulting in smaller Si features. High resistivity samples (*"S"*), however apparently showed inverse behavior. Due to small number of measurements and large experimental error, this behavior will be more closely followed through in the future.

III. CONLUSION

Porous silicon samples were prepared from p^- and p^{++} doped crystalline substrates using different anodisation parameters. Raman and photoluminescence spectra are consistent with previously published results. We report a

surface morphology instability tentatively attributed to difference between substrate and electrolyte conductivities. Silicon nanocrystal dimensions in PSi samples obtained from photoluminescence and phonon confinement models were correlated with silicon electrochemical dissolution valence n_d. In high resistivity samples increase of crystallite dimensions with dissolution valence n_d was observed, contrary to expected behavior. This phenomenon will be further investigated in our future works.

REFERENCES

[1] Lehmann, V., Electrochemistry of Silicon: Instrumentation, Science, Materials and Applications, Wiley-VCH Verlag GmbH, 2002.

[2] Canham, L., "Visible photoluminescence from porous silicon", in "Properties of porous silicon," INSPEC, London, 1997, pp 249-255

[3] Delerue, C., Allan, G. and Lannoo, M., "Theoretical aspects of the luminescence of porous silicon," Phys. Rev. B 48.15 11024, 1993

[4] Ledoux, G., J. Gong, F. Huisken, O. Guillois, and C. Reynaud. "Photoluminescence of size-separated silicon nanocrystals: Confirmation of quantum confinement," Appl. Phys. Lett. 80(25), pp 4834-4836, 2002

[5] Solomon I., Wehrspohn R.B., Chazalviel J.-N. and Ozanam F., "Spatial and quantum confinement in crystalline and amorphous porous silicon," J. Non-Cryst. Solids, vol. 227–230, Part 1, pp. 248-253, 1998

[6] Fauchet, P. M., and Ian H. Campbell, "Raman spectroscopy of low-dimensional semiconductors," Critical Reviews in Solid State and Material Sciences 14.S1 79-101, 1988

[7] Zi, J., Buscher, H., Falter, C., Ludwig, W., Zhang, K., & Xie, X., "Raman shifts in Si nanocrystals," Appl. Phys. Lett. 69(2), pp200-202, 1996

Some physical problems in the preparation and analysis of the heavily boron and phosphorous doped polysilicon thin films

M. Očko[1,*], S. Žonja [2] and M. Ivanda[3]
[1]Institute of Physics, Bijenička cesta 46, 10002 Zagreb, Croatia
[2]University of Zagreb, Faculty of Electrical Engineering and Computing Department of Electronics, Microelectronics, Computer and Intelligent Systems, Unska 3, 10000 Zagreb, Croatia
[3]Ruder Boškovic Institute, Bijenička cesta 54, 10002 Zagreb, Croatia
*E-mail: ocko@ifs.hr

Abstract - We discuss problems in the preparation and analysis of the solid solution phase in the case of heavily phosphorus and boron doped polycrystalline silicon. Some currently used methods and the technology of polysilicon thin films preparation are presented. Based on and afore these discussions, we propose a process of preparation of a good quality heavily boron and phosphorus doped silicon thin film samples. The characterisation, with some new methods of characterisations, and the investigation of the stability of such obtained samples are presented. It was also indicated that the problem of the determination of the current carrier concentration is not yet solved even in the heavily doped, metallic, regime in polysilicon..

I. INTRODUCTION

It is difficult to number all the applications of polysilicon today. In the last few decades, polysilicon has affirmed its position as one of the most important materials in integrated circuit technology. Apart from its importance in micro-electro-mechanical-systems [1], it has been studied and used as a material for the gate electrode in metal-oxide-semiconductor devices, for the fabrication of high value resistors, for diffusion sources, conduction lines or for applications in active matrix-liquid crystal displays and photovoltaic conversion of solar energy. Some integrated circuits processes demand highly doped polysilicon layers with surface as less rough as possible [2]. In the field of thermoelectrics, one expects just from polysilicon a breakthrough towards certain cost effectiveness of the TE devices [3].

It is interesting to note that all of this enormous in quantity of various products and, therewith, enormous diverse in physical properties demanded by these very different applications come from a tinny difference in, otherwise, very small dopant concentrations that can be dissolved in silicon. These concentrations are mainly less than 1 at%. Moreover, in many cases doped silicon used, even with such low concentrations are not stable because the equilibrium solubility at room temperature is mainly less than 1 at%.

The other instability, or better to say, metastability of polysilicon is caused by the existence of the grains. But the possibility of producing different grain sizes gives the possibility of production some additional diverse properties within the same dopant concentration..

Without going into technical and technological problems, here we stress some physical properties which lead to difficulties in sample preparations: low dopant equilibrium concentration, great variation of physical properties with concentration and the variation of properties with the sizes of the grains in polysilicon. All of this mentioned instabilities, are the reasons why there exist various preparation techniques and why there is an enormous need for purity, i.e., for extremely clean rooms during sample preparation. Also, one needs a lot of knowledge and experiences to produce a polysilicon material with desired or demanded properties and desired stability for certain application.

In this sort communication, we shall touch only a small part of the overall problem: we shall deal with the preparation of the heavily boron and phosphorus doped polysilicon thin films. In addition, it will be presented the problem of the analysis of polysilicon material, especially, the problem of the concentration determination.

A. Solubility

The main differences in physical properties in a semiconductor come from the different concentration of current carriers. This concentration depends on the possible concentration of the substitutional positions of dopant atoms in semiconductor crystal lattice, i.e., it depends on solubility. Only the substitutional dopant positions insure the current carries in semiconductors.

The solubility of an atom in the matrix of another one is theoretically one of the unresolved problems in solid state physics. Experimental facts are gathered within the Hume-Rothery rules. The solubility of a proper dopant in a semiconductor is, in principle, very low because the dopant valence should be different from the valence of the semiconducting matrix. Equilibrium phase diagrams of the semiconductor-dopant type at especially low dopant concentrations are mainly not known. Precisely speaking, the equilibrium solubility line in these diagrams is not precisely known. Better to say, no one of the diagrams for particular binary system is commonly accepted.

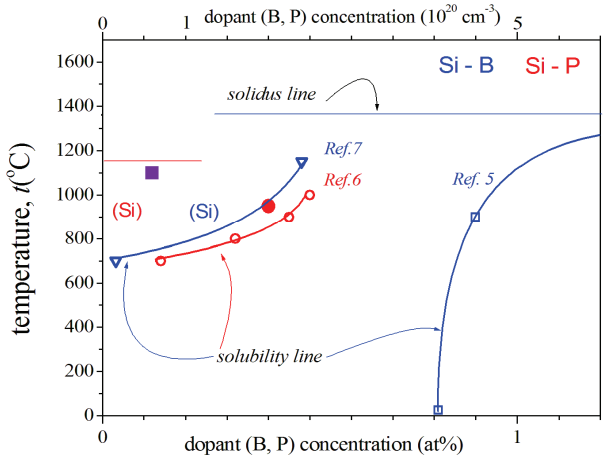


Figure 1. The phase diagrams at low dopant concentrations of Si-P, open circles from Ref. 6, and Si-B, open squares from Ref. 5, open triangles from Ref. 7 are placed within the same graph in order to point very different results found in literature. Closed symbols denote the concentrations and the temperatures of annealing of the samples we are dealing with in this paper. Closed square (far from the corresponding solubility lines) represents the case of the Si:B samples. Closed circle (close to the corresponding solubility line) represents the case of the Si:P samples.

About ten or more various experimental data for the *solubility lines* for of Si-B and Si-P can be found in literature. In Fig. 1, the Si-B and Si-P phase diagrams at the low dopant concentration taking data from Refs 5, 6 and 7 are sketched within the same graph in order to point great differences among the results found in literature. As can be seen, it is not determined and commonly accepted even whether there exists an equilibrium solubility of boron at room temperature [5,7]. But all the known investigations exclude possibility that phosphorus is soluble at room temperature in silicon. Similar problem with solubility is in the case of the other proper dopants, like Sb, As, in silicon. Hence, one should take into consideration the metastability of these materials at room temperature. It is obvious that the preparation of the heavily doped silicon demands special care in sample preparation. Although there is not defined precisely what does it mean *heavily doped*, one can take that the concentrations of these materials are above the metal-isolator transition (MIT) which is at about $4 \div 5 \cdot 10^{18}$ cm^{-3} for boron and phosphorus in silicon [8].

II. SAMPLE PREPARATION

A. *On the sample preparation techniques*

There are in principle two physically different ways to get proper or desired dopant concentrations: a) thermal processes and b) low temperature (room temperature) implantations. Here, we shall give only a short overview of these methods and of the thin layers obtained by these methods in order to discuss the technology of the preparation of the samples we were dealing with.

The thermal processes use the fact that the solubility of dopants is higher at higher temperatures. A consensus on the highest possible solubility of boron and phosphorus is not achieved, but usually it is cited as $4 \cdot 10^{20}$ cm^{-3}, or about 0.8 at% due to the solubility at high temperatures but below solidus line, which is about 1400ºC [9]. References 6 and 7 do not support such high values.

Ion implantation methods, which are performed at room temperature usually, are capable to produce much higher dopant concentrations in substitutional positions and, therewith, by short time annealing, the dopants can be activated. The reported values are even up to $7 \cdot 10^{21}$ cm^{-3}, i.e., more than 10 at% [9]. However, the alloys with the concentrations above solubility limit are rather unstable, especially if they operate at elevated temperatures. The problem of stability will be treated in Sec. 3.3 But there are some other bad characteristics of such obtained thin films. First of all, by high energy implantation, enormous strain is introduced in the samples. Moreover, crystal lattice damage is considerable and this damage should be repaired. It could be done by thermal treatment at elevated temperatures but it is inevitably connected to the loose of the activated dopands. The profile of such obtained samples is in principle not uniform with a maximum whose position depends on the implanted dopant and implantation energy. It is not usual to get a uniform profile from such as-obtained samples. A thermal treatment leads to the lost of the activated dopants. But such profile with different concentration within a sample leads to a diffusion of the dopant ions towards lower concentrations even at very low temperatures. Thus, one could have instabilities of a device, which has working temperature close to room temperature. Also the thickness of the layers cannot be high. The bad characteristics of the ion implantation are discussed because we want to show that the thermal methods we were used had some advantages although much lower concentrations of the activated dopants can be achieved.

In what follows, we present two set of samples we were dealing with. First, we describe the demands on the samples intended for certain application and then it is described their preparation. One of the points of this paper is to show that application and corresponding sample properties determine the way and the method of sample preparation.

B. *Si:B*

The Si:B samples were intended for the investigations of the thermoelectric (TE) properties. It was shown, mainly theoretically, that the maximum of the TE power factor, P, would be at the concentrations between 10^{19} cm^{-3} and 10^{20} cm^{-3}. The theory is inferred for the small gaped semiconductor and for monocrystalline material. We wanted to see if this theoretical prediction is valid for our samples which are polycrystalline ones and are in metallic regime. The demand on the samples was, besides these concentrations, to obtain relatively thick layers in order to get specific heat of the samples as high as possible. Certainly, a high percentage of the activated boron was desired, as well. There was no demand on the sample surface roughness. In the preparation, we used thermal method because it could fulfill all of these demands.

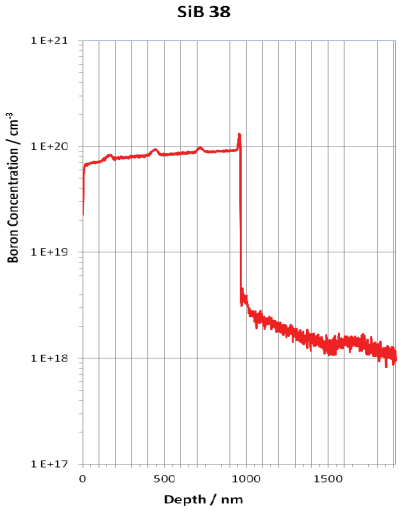


Figure 2. Profile of one of the Si:B samples 3 times δ-doped. The effects of this can be clearly seen even after 1 hour of annealing at 1100 ºC. Compared to the ion implantation, one can take that the concentration is rather homogenous and is about $8 \cdot 10^{19}$ cm^{-3}. The depth is not calibrated and the numbers at x-axis are arbitrary. By the bevel method, the depth of the Si:B samples was determined as 2.1 μm.

The thermal method consists of three steps: deposition, activation and annealing. For the deposition, we used low pressure chemical vapour deposition (LPCVD) technique. Monocrystalline silicon wafers with diameter of 2" and <111> orientation were placed into the LPCVD reactor tube at 750°C for 1 hour at working pressure of 0.3 mbar (0.22 Torr). Fixed BCl_3 volume, V_o (~8 cm^3), was flushed into the reactor tube after SiH_4 flow. By repeating this procedure, different concentration were obtained in, so called, *δ*-doping procedure. The deposition temperature, 750°C, determined that the deposited samples were polycrystalline.

After the deposition, the samples were subjected to the thermal annealing at 1100°C for 1 hr. Thus, the dopant, boron, electrical activation was achieved together with homogenization. The other details of LPCVD method, sample preparation and their preparation for various measurements are described elsewhere [10].

C. *Si:P*

The Si:P heavily doped samples were used in an as diffusion sources in a fabrication of shallow pn junctions [3]. One certainly expected that such materials had *i*) activated dopant concentration to be as high as possible; *ii*) without of any kind of precipitatates or amorphous phase at the grain boundaries if possible. *iii*) A further demand was that the sample surface would be smooth. i*v*) Certainly, a demand was for the *n*-type current carriers. *v*) As the investigations [3] were performed for certain commercial product, one has to take into account also the cost effectiveness of this final product.

In-situ doped silicon thin films were grown by LPCVD at 530°C (amorphous) and 580°C (polycrystalline). All the samples were deposited on (100) silicon wafers. Doping was achieved with a gas mixture of phosphine and siliane in the reactor tube. In this way the homogenous samples were achieved.

Demands *i*) and *ii*) including demand *v*) determined that the optimal phosphorus concentration is $2 \cdot 10^{20}$ cm^{-3} (about 0.4 at%) and that the annealing temperature is 950°C. Closed symbols in Fig.1 show the positions of the Si:B and Si:P samples with respect to the corresponding solubility line. The position of Si:P show that the Si:P samples in cooling down should pass a wide range of temperatures where the solubility is zero. At these temperatures, phosphorus segregates towards grain boundaries where it is deactivated. In that case, the resistivity increases because of smaller number of current carriers, but also due to the amorphous phase at the grain boundaries. Therefore, the annealing procedure cannot be performed like in the case of the Si:B samples. For proper annealing, the rapid thermal annealing (RTA) technique was used [11]. This technique enables short time annealing and rapid cooling down to room temperature. Thus, we were able and we investigated the properties of these samples after 10 s, 20 s, 30 s and 45 s of annealing.

III. ANALYSIS

In this short communication, we cannot not give an overall presentation of the results of the analysis we performed, but we want here only to point the problems in analyzing of the heavily doped polysilicon thin films. Overall review of these investigations will be given elsewhere.

A. *X –ray analysis*

X-ray diffraction (XRD) investigations revealed that the samples deposited as amorphous (at 530°C) became polycrystalline already after 10 s of annealing Topography/morphology of the Si:P films has been determined by X-ray reflectivity (XRR) method. One of the results of these investigations was that the surface of the thin films deposited as amorphous (A), although having larger grains, had smoother surface than the other ones (B).

It is interesting that this characteristic (demand *iii*) was decisive for the application although the resistivity of

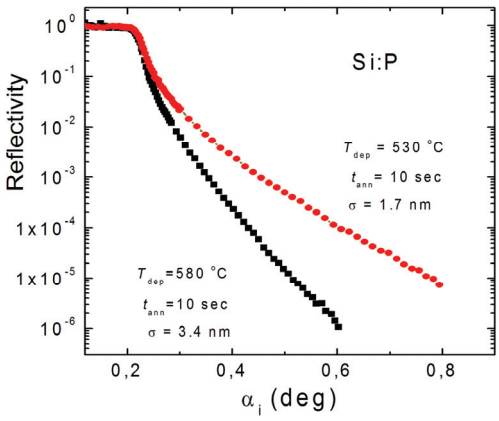


Figure 3. Reflectivity from the surface of two Si:P samples. One of them (circles) was deposited as amorphous (A) and the other one as polycrystalline (B).

These samples had higher resistivity than the samples deposited as polycrystalline.

X-ray scattering is very useful in determination of precipitates within grains or at the grain boundaries (so called small angle scattering methods). Such information would be of great importance in semiconducting physics. However, even in the heavily doped regime the dopant concentrations are too small and x-ray analysis is not able to give such information.

B. Concentration determination

Determination of concentration in the case of polycrystalline materials, as it is known and we shall show, is an unresolved problem in semiconductor physics. Two methods are mostly used for determination of total number of a dopant in a material: Rutherford back-scattering spectrometry (RBS) and Secondary Ion Mass Spectroscopy (SIMS). As RBS cannot detect boron and hydrogen and is useless in our case.

But the main problem is to determine the current carrier concentration. Very often, even in industry, one uses resistivity measurements and compares them to the one of $n(\rho)$ curves published in literature. Such procedure is certainly very uncertain because the resistivity in a polycrystalline material consists of:

a) trapping of current carriers

b) passing current carriers between grains

c) resistivity within grain (phonon contribution, scattering on impurities, ionized and neutral, scattering on disorder, electron-electron scattering and so on).

Trapping mechanism describes well alone the resistivity below critical concentration. At higher concentrations, the role of trapping decreases and resistivity tends to the resistivity of monocrystalline silicon, but the difference is always present. Two effects

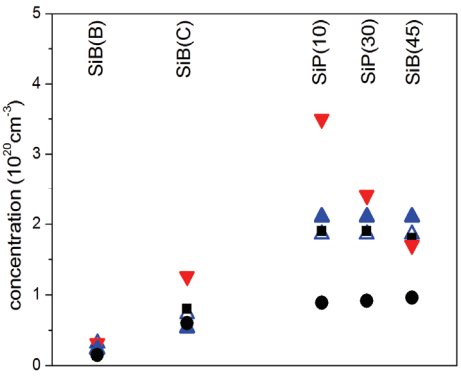


Figure 4. Concentrations of the investigated samples determined from various measurements: resistivity (closed circles), SIMS (closed squares), Hall effect with $r = 1$ in fotmula (1) (down triangles), thermopower with electronic mass $m = m_o$ (up triangles), thermopower with $m = 0.8m_o$ for Si:B and $m = 1.1m_o$ for Si:P.

are produced by trapping. Certain amount of current carriers are trapped and a potential barrier is formed. Both mechanisms increase resistivity. If a dopant segregates to grain boundary the resistivity increases because number of current carriers decreases but, in addition, resistivity might increase also because resistivity between grains might increase.

There are some $n(\rho)$ curves for the case of polycrystal silicon thin films found in literature. They differ a lot among themselves because $n(\rho)$ depends on the grain size and because of possible existance of the segregated phosphorus at the grain boundaries. We took the $n(\rho)$ curve for Si:B from Ref. 12 and for Si:P from Ref. 13. By such obtained concentration we could explain the best our overall results. This was the reason we chose the data from these references.

Because of these problems with resistivity, the Hall effect is much convenient and is most straightforward to determine the current carrier concentration. One takes that the Hall effect gives just the number of current carriers. However, our results show that simple relation inferred in the free electron approximation:

$$n = \frac{r}{qR_{\mathrm{H}}} \qquad (1)$$

where n is concentration of current carriers, q the charge of current carrier, and R_H, the Hall constant. r = 1in the free electron approximation. One usually takes $r = 0.8$ in the case of p-type material and $r = 1$ in the case of n-type material at room temperature [12]. The reason of $r \neq 1$ one takes is the complicate band structure of silicon.

In Fig. 4, we summarise our determination of current carriers for the samples investigated. One can see a great difference between the determined concentrations from the resistivity (closed circles) and from the Hall effect data (closed triangles). But the problem with the Hall effect measurements is that they give larger values than the SIMS gives. Something better agreement, one can get taking $r = 0.8$ for Si:B.

Influence of disorder on the determination of the current carriers is indirectly theoretically investigated in Ref. 14 in the case of complete degeneration. Not going into consideration of this theoretical work here, we may say it might explain these higher values of the concentrations coming from the Hall effect measurements and might explain even their unexpected concentration dependence. The problem is that the values for r coming from this and similar theories are ranged from 0.75 to 1 at room temperature depend on the approximation used.

But our investigations showed that thermopower in the concentration range from $1 \cdot 10^{19}$ cm^{-3} up to concentrations we investigated $2 \cdot 10^{20}$ cm^{-3}, a very important range for various different applications, can be reliable for the determination of concentration (see closed and open up triangles). The considerations and discussion on this finding will be given elsewhere.

The determined values of the concentrations of the current carriers from the resistivities for Si:P are much lower than the one from the other physical quantities but we take them as the reliable values because:

- they are more compatible with some other measurements and considerations
- they are compatible with similar investigations found in

literature [15].

Besides, the current carriers determined from the Hall effect and thermopower are, in general higher or about the concentrations inferred from SIMS, which gives the total number of dopants.

The aimed dopant concentrations of the samples we investigated depended on the application demands (see: Sec. I). For the Si:B samples, it was in the range from $1{\cdot}10^{19}$ cm^{-3} to $1{\cdot}10^{20}$ cm^{-3}. For all the Si:P samples, the aimed concentration was $2{\cdot}10^{20}$ cm^{-3}. The preparation technique, all our considerations, characteristics of the dopants and range of concentrations give us finally that we may conclude: For Si:B, we got even about 100% current carriers with respect to the total number of the introduced dopants. But in the case of Si:P it was only about 45%. Such results are consistent with similar investigations described in Ref. [15]. Browsing literature, we may say that this 45% is one of the better percentages at these high concentrations. The effects of 55% inactivated phosphorous, we observe in rather high resistivity of our samples and on the low temperature behaviour below 50 K in the resistivity and thermopower. These considerations will be given elsewhere.

According to these results, we may offer a conclusion. As it was obtained the solubility in Si:B was roughy about 100% although we did not take much care on the rapid cooling after homogenization at 11000°C, it comes out that our results are compatible with the solubility line of Si:B published in Ref.5.

C. Investigations of stability

In this chapter, we consider the electrical stability of the Si:P samples due to non solubility of phosphorus in silicon at room temperature. As said, these alloys are metastable at low temperature. This means that the obtained microstructure is for a long time is stable, i.e. unchanged considerably at room temperature. But if a device works or be undesirably for certain time at elevated temperature the microstructure changes. Thus the electrical characteristics will not be the same at room temperatures In Fig. 5, we sketched the process of annealing of two samples taking Ref. 9. Annealing at lower temperature, 450°C, has smaller effect than annealing at 550°C. After 1hour, the lost of the current carriers is 0.9% at 450°C, but is 9.5% at 550 °C. Obviously, the process of changing of microstructure is an activation process. Hence, one may expect that the lost of the current carriers would be faster at higher, 600°C.

In the inset, results of annealing for 1 hour at 600°C of our four samples are presented. Not going into details, we may say that the energy barrier for the transforming microstructure is in average equal or higher in the case of our samples. Also, the loss of the current carriers was considerably lower for the samples C and D which are 30 s and 45 s annealed at 950°C before annealing them at 600 °C.

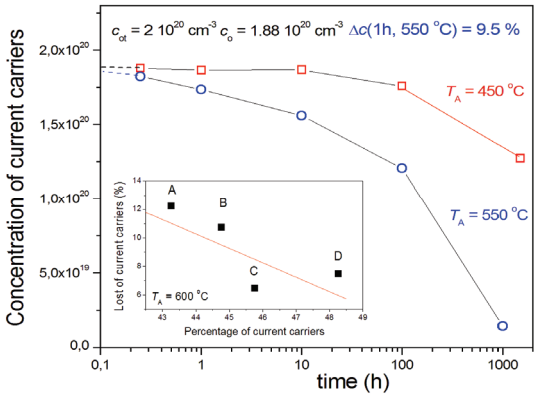


Figure 5. In the figure, we sketched the process of annealing of two samples taking data from Ref. 9. In the inset, our results of annealing for 1 hour at 600°C of four samples are presented. (More in the text)

IV. SUMMARY

In this paper, it was shown how the problem of the insolubility of boron and phosphorus in silicon, as well as various demands on the properties of materials influence the way of preparation of heavily doped poliysilicon thin films. According to the demands for the two sets of the samples we suggested and described the methods and the ways of the sample preparation. It was also point the problem of determination of the current carrier concentration.

ACKNOWLEDGEMENTS

This work has been supported by the Croatian Ministry of Science, Education and Sports (projects 036-0361566-1567, 036-0982904-1642 and 098-0982934-2744). We thanks HYDEH ANALITICAL, Warrington WA5 7UN, UK. for providing us the SIMS measurements.

REFERENCES

[1] J. Lee, Fabrication, Characterization and Application of Multifunctional Microcantilever Heaters, Georgia Inst. of Tech, 2007.

[2] A. D. McConnell, S. Uma, K. E. Goodson, *Thermal cond-uctivity of doped polysilicon layer*, J. Microelectro-mech. Sys, **10,** no. 3, (2001) 360-369.

[3] M. Koričić, *Horizontal Current Bipolar Transistor Struct- u res for Integration with CMOS Technology*, University of Zagreb, 2008.

[4] M. Očko, S. Žonja, M.Ivanda, *Thermoelectric materials: problems and perspetives,* 2010 *Proc. of MIPRO 2010,* p 42-48, Rijeka, Croatia

[5] G. V.Samsonov and V. M. Sleptov, Zh. Neorgan. Khim., **8** (1963) 2009-2011 in R. P. Elliot 1965 *Constitution of binary alloys*, *First supplement*, McGraw-Hill book company, Inc., 1965, p 98

[6] Trumbore F A 1960 *Bell Syst. J*, **39** (1960) ibidem, 716.

[7] G. L. Vick and K. M. Whittle, *Solid Solubility and Diffusion Coefficients of Boron in Silicon*, J. Electrochem. Soc., **116** (1969) 1142-1144.

[8] P. Dai, Y. Zang, and M. P. Sarachik, *Crytical conductivity exponent*, Phys. Rev. Let. **66**, 1914-1917.

[9] M. Finetti, P. Negrini, S. Solmi and D. Nobili, *Electrical Properties and stability of Supersaturated Phosphorus-Doped Silicon Layers*, J. Electrochem. Soc, **128** (1982) 1313.

[10] S. Žonja, M. Očko, M. Ivanda, P. Biljanović, *Low temperature resistivity of the heavily boron doped LPVCD poly-silicon thin*

films, J. Phys. D: Appl. Phys. **41** (2008) 162002.

[11] X. Zhang, T.-Y. Zhang, M. Wong, and Y. Zohar, *Rapid Thermal Annealing of Polysilicon Thin Films*, J. Microelectromechanical systems, **7**(4) (1998) 356-364.

[12] W. R. Thurber, R. L. Mattis. Y. M. Liu, and J. J. Filliben, *Resistivity-Dopnt Density Relationship for Boron-Dpoed Silicon*, J. Electrochem. Soc.: SOLID STATE SCIENCE AND TECHNOLOGY **127** (10) (1980) 2291-2294.

[13] W. R. Thurber, R. L. Mattis. Y. M. Liu, and J. J. Filliben, *Resistivity-Dopnt Density Relationship for Phosphorus-Dpoed Silicon*, J. Electrochem. Soc.: SOLID STATE SCIENCE AND TECHNOLOGY **127** (8) (1980) 1807-1812.; J. C. Plunkett, J. L. Stone and A. Leu, *A Computer algorithm for accurate and repeatable analysis using anodization* and stripping of silicon, Solid-State Electronics, **20** (1977) 447-453

[14] A. G. Mojseev, Physics of semiconductors and diele-ctrics; A study of kinetic effects in isotropic polycrystalline p-type silicon, Russian Physics Journal, **51** (2008) 195-205.

[15] M.Sarret, A. Liba, P. Jourbet, and B. Fortin, *n-type silikon thin films obtained by crystallization in situ phosphorus-doped amorphous silicon films deposited at low pressure*, J. Appl. Phys. **76**(9) (1994) 5492-5498.

Luminescence, Raman and synchrotron XPS study of amorphous Ge_2S_3 based films

V.Mitsa*, M.Ivanda**, O.Gamulin***, R.Holomb*, O. Kondrat*, N. Popovych*, G.Lovas*, S.Petreckiy*, N. Tsud****, V. Matolín*****, K.C. Prince****

*Uzhhorod National University, Institute for Solid State Physics and Chemistry, Uzhhorod, Ukraine
v.mitsa@gmail.com
**Division of Materials Physics, Rudjer Boskovic´ Institute, Bijenic˘ka 54, 10000 Zagreb, Croatia
*** University of Zagreb, School of Medicine, Department of Physics and Biophysics, Šalata 3, 10000 Zagreb, Croatia
****Elettra-Sincrotrone Trieste S.C.p.A., Strada Statale 14, km 163.5, 34149 Basovizza, Trieste, Italy
*****Charles University, Faculty of Mathematics and Physics, Department of Surface and Plasma Science, V Holešovičkách 2, 18000 Prague 8, Czech Republic

Abstract – **Ge_2S_3-based films have been investigated using Raman, photoluminescence and photoemission spectroscopy. For the sake of comparison some of the photoluminescent properties of bulk glasses have been used. The synchrotron photoelectron spectroscopy (SRPES) and XPS spectra of a-Ge_2S_3 films have been measured after the illumination and annealing. The changes in the parameters of the film's core levels induced by near or above band gap light and thermal treatment are discussed in the paper. We consider the possibility that the PL radiation in Ge_2S_3-based films is a surface contaminant effect from native oxidized layer, which might have formed in the air.**

I. INTRODUCTION

The discovery of recombination processes in chalcogenide glassy semiconductors (ChGS) [1] has been the catalyst for the further study of photoluminescence (PL) in binary germanium and arsenide chalcogenides for comparing their band gap energy structure in both crystalline and glassy forms [2-5]. Investigations of PL with respect to the types of structural units and in interconnection with other physical properties have been extended to ternary glassy systems and films based on them [3, 6-8]. Optical quality of the ChGS films at, and near the surface is the most important property for practical optical applications [9]. The tendency to thermal dissociation at higher temperature and formation of Ge-rich species on the surface of films [10] during thermal evaporation of Ge-based glasses have been revealed. The information about high refractive index on the surface of a–GeS_2 films to a depth of 10 nm was used for the compensation of spectral changes of different types of IR interference filters with a-GeS_2 high refractive layers [11]. Recently during the investigation of oxysulfide GeS_2-GeO_2 films [12] it has been found that preparation of a-GeS_2 film by magnetron sputtering is accompanied by the formation of a thin layer at the top of the film, but with lower index of refraction in a layer 20 nm thick. Information about surface contaminants for a-Si films was obtained from the analysis of the high energy PL radiation [13].This was an effect unrelated to the bulk states of the material [13]. In [14] we assumed that the visible PL peak at 2.2-2.3 eV might arise from GeO_x impurities in the structure of GeS_2-based glasses. In such glasses the main peak in PL spectra at 1.96 eV was identified with sulfide-dominated dopant sites, while the shoulder at 2.02 eV was attributed to oxide-dominated sites [15]. The mechanism of luminescence in Ge- based amorphous and nanocrystalline materials is still being disputed [16-21]. It is found that large PL signals correlate with good interface properties [22]. Laser illumination and thermal annealing might be employed in order to get high quality ChGS films [23]. Here, we report the observation of room temperature visible PL emission in amorphous films of As_2S_3-Ge_2S_3 alloys before and after laser illumination with near or above band gap light. Bulk glasses were used for the comparison of film PL properties. Surface characterization of Ge_2S_3–based films on the top was carried out with high resolution synchrotron photoelectron spectroscopy (SRPES). Deeper layers up to 3 nm were investigated by ordinary XPS.

II. MATERIALS AND METHODS

The mixture was synthesized in evacuated (~10^{-3} Pa) quartz ampoules by step-wise gradual heating up to 1223 K in a rocking furnace. At the end of the process each ampoule was quenched to room temperature. Films were obtained by flash evaporation of glassy powder onto c-Si substrate with an Al interlayer. The Raman spectra have been recorded with a DILOR Z-24 triple spectrometer and FT-Raman module of Perkin-Elmer GX spectrometer. The excitation was provided by a green line λ= 514.5 nm (E= 2.41 eV) from an Ar-ion laser, and near infrared line λ= 1064 nm (E=1.17 eV) from a Nd:YAG laser. A projector lamp NL-100 was used for exciting PL spectra of bulk polished glassy samples. IR illumination from the lamp was cut off by a filter based on $CuSO_4$ water solution. The PL of film samples were excited by cw light from an Ar+ laser illumination with near/or above optical gap photon energy 2.41 eV. The irradiation of films was carried out in air at room temperature for ~0.5 h by the same laser with 35 mW/cm^2 intensity.

Photoemission measurements were performed at the Materials Science Beamline of the Elettra Synchrotron light source (Trieste, Italy). The spectrometer was calibrated with the Au $4f_{7/2}$ photoemission line at 84.0 eV from an evaporated gold film. The S 3d, Ge 3d, C 1s and O 1s core level peaks were fitted using a Voigt function with subtraction of a Shirley type background to yield peak position and intensity. Other details of measurements were the same as described in our previous article [24].

III. RESULTS AND DISCUSSION

In PL spectra of $(Ge_2S_3)_x(As_2S_3)_{1-x}$ glasses excited by the lamp, broad PL bands well below the band edge (Figure 1) are seen. The values of the optical pseudogap (E_0) for g-Ge_2S_3 and g-As_2S_3 are equal to 2.16 and to 2.4 eV respectively [3,14]. At x=1, the maximum of the PL spectrum is at 1.12 eV. Near the middle gap position of all maxima of the PL spectra of $(Ge_2S_3)_x(As_2S_3)_{1-x}$ glasses (Figure 1) corresponds to the classical $E_o/2$ rule, and such PL spectra might be explained in terms of the Street-Davis model [25]. Some deviations from this model are discussed in [14,18]. When the mean coordination number of $(Ge_2S_3)_x(As_2S_3)_{1-x}$ glasses is below the value

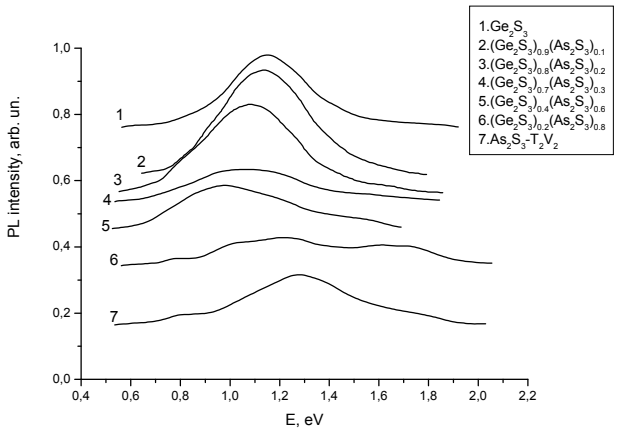


Figure 1. PL spectra of $(Ge_2S_3)_x(As_2S_3)_{1-x}$ glasses measured at nitrogen temperature (composition is shown on insert of Fig.1)

of z=2.67, the intensity of the PL spectrum for x=0.2 decreased and the half width increased (Figure 1). For this composition, the PL spectrum is very similar to the PL spectrum of freshly thermally deposited a-As_2S_3 films [26]. The concentration of As_4S_4 molecules (235 cm^{-1}) for a-As_2S_3 film and for bulk $(Ge_2S_3)_x(As_2S_3)_{1-x}$ glass have maxima near x=0.3 [27]. During excitation of PL spectra of natural realgar (E_g=2.5-2.6 eV) by laser beams with different energies at 2.38 and 2.6 eV, a low energy PL band was found with peak near 0.9 and 1.7 eV peak of the higher energy correspondingly [5]. Exciting PL spectra of ternary glasses (Figure 1) by the lamp with a wide band energy distribution leads to the appearance in PL spectra for samples at x=0 and x=0.2 of two shoulders near 0.9 and 1.7 eV in same position as peaks in the PL spectrum of natural realgar, which were registered by different lasers line excitation [5].

Table 1. The relation between the composition of film samples and mean coordination number

$(Ge_2S_3)_x(As_2S_3)_{1-x}$ $Ge_xAs_{40-x}S_{60}$	Z	d/nm
(Ge2S3)0.3(As2S3)0.7 $Ge_{12}As_{28}S_{60}$	2.52	400
(Ge2S3)0.5(As2S3)0.5 $Ge_{20}As_{20}S_{60}$	2.6	500
(Ge2S3)0.7(As2S3)0.3 $Ge_{28}As_{12}S_{60}$	2.68	750
(Ge2S3)0.9(As2S3)0.1 $Ge_{36}As_4S_{60}$	2.78	450

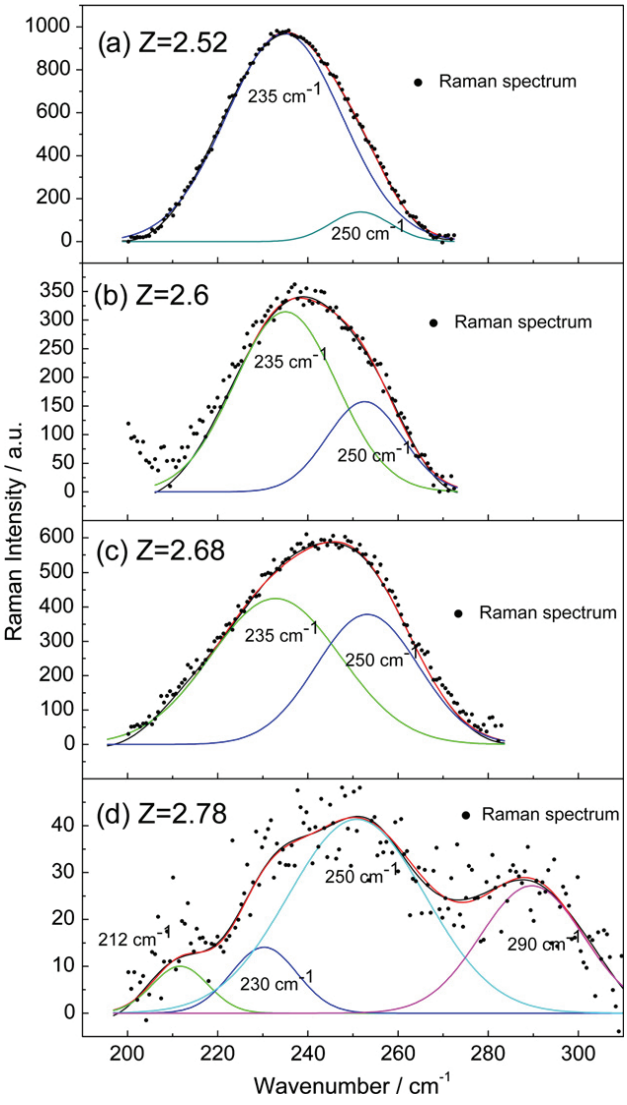


Figure 2. Raman spectra of $(Ge_2S_3)_x(As_2S_3)_{1-x}$ films, E_{ex}=2.41 eV: (a)- x=0.3; (b)-x-0.5; 3-(c)- 0.7; (d)- x=0.9

In the process of room temperature measurement PL spectrum of bulk g-As_2S_3 with realgar-type As_4S_4 molecular inclusions, while the low intensity PL band peaking at 1.7 eV has been also found [14]. We expect

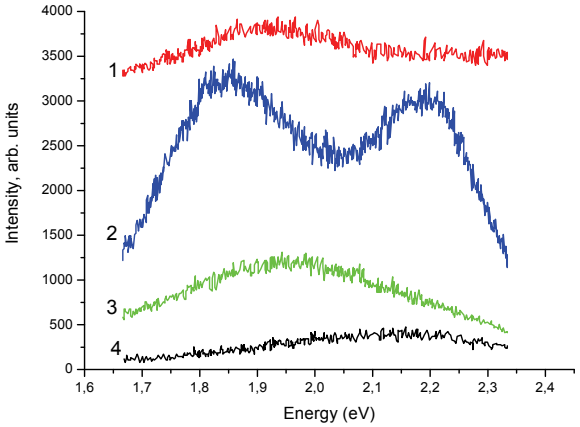


Figure 3. Room temperature PL spectra of $(Ge_2S_3)_x(As_2S_3)_{1-x}$ films, E_{ex}=2.41 eV: 1-x=0.9;2-x-0.7;3-x-0.5; 4-x=0.3

that a similar band could appear at room temperature in the PL spectrum of $(Ge_2S_3)_x(As_2S_3)_{1-x}$ film samples (Table 1) at x=0.3, 0.5 in which Raman spectra we registered intense vibration of p-As_4S_4 molecules (235 cm^{-1}) (Figure 2). However the film sample at x=0.3 gives in the PL spectrum a low intensity wide band PL signal whose maximum is centered at 2.15 eV (Figure 3). For thicknesses of the films from 400 to 500 nm at x=0.5 (Table 1) the intensity of this band slightly increased and its maximum is red shifted to 1.95 eV. Further increasing the Ge content (x=0.7) and thickness of the film to 750 nm (Table 1) leads to the appearance in the PL spectrum of two well separated bands near 1.85 and 2.2 eV (Figure 3, curve 2). When the composition of ternary films is approaching Ge_2S_3 and the thickness of the film decreases to 450 nm (x=0.9) the maximum of PL spectrum is seen near 1.95 eV and there is a shoulder near 2.2 eV (Figure 3, curve 1). The decrease of intensity of PL light and Raman signal (Figure 2) in the last case might be connected with a decrease of the penetration depth of the exciting laser beam at 2.41eV (from the absorption coefficient, $1/\alpha$= 100 nm) due to narrowing of the g-Ge_2S_3 optical band gap (E_o= 2.16 eV). We note that for As_2S_3 based ternary films (E_o= 2.4 eV) the value $1/\alpha$ is a few times more than for g-Ge_2S_3. If we take together the known data about luminescent properties of binary bulk Ge-S glasses and films in the region 1.9-2.5 eV, we might find common PL emission characteristics of the species in other Ge-containing materials. For example, in Ge-doped silica gel glasses was observed room temperature red PL emission with peaks at 1.89 and 2.12 eV for 514.5 nm of an Ar+ laser excitation [28]. Green PL (2.1-2.5 eV) of GeO_2 was reported in nc-Ge or Ge nanocrystals embedded in GeOx ($0 < x < 2$) thin films [29]. In general, depending on the excitation wavelength the PL energy in GeO_x shows variations. Usually a blue PL band (2.8 -3.1 eV) is observed when the PL is excited in the near UV [19], and when the excitation wavelength is longer than the blue wavelengths, the other PL bands (i. e. 1.6-1.9 eV, 2.1-2.3 eV, bands) are observed [28]. The most commonly reported PL band from GeO_x is the broad yellowish-green band located between 2.1-2.3 eV [30]. From the variety of oxygen-deficient defects the theoretical modeling shows that X3Ge–GeX2 (X=OH, OAH3, A=Ge) defect gives only a red/orange PL band at 2.0–2.1 eV [31]. Oxygen related bands in the PL spectrum of an as-synthesized GeO_x nanowire, which was excited at the O K-edge (536.5 eV), showed during fitting procedure four components at 1.90, 2.17, 2.42, and 2.70 eV. Thermal annealing of the GeO_x nanowire leads to the disappearance of the high energy bands at 2.42 and 2.70 eV and decreasing intensity of PL yield [19]. Taking into account the similarity of GeO_x, Ge-based glasses and films PL spectra (Figure 3) we consider the possibility that the PL radiation in Ge_2S_3-based films is a surface contaminant effect [13] from native oxidized layer, which might have formed in the air. In order to determine if this hypothesis is correct, we have analyzed a difference plot of Raman spectra of $(Ge_2S_3)_x(As_2S_3)_{1-x}$ z=2.78, excited by IR and Ar-ion lasers (Figure 4) first of all. Due to the different penetration depth of IR and Ar-ion laser illumination into the film, this curve is characteristic of the surface of the film. Comparing both Figure 2 and Figure 3 in the range 200-300 cm^{-1} shows that the surface is enriched in GeS-like (212,230 cm^{-1}), ethane-like (250 cm^{-1}) and $Ge_{4/4}$ (290 cm^{-1}) structural units (s.u). Detailed assignment of the Raman bands for these samples has been done in [27]. In the range 300-500 cm^{-1} the Raman spectrum g-Ge_2S_3 has two intense bands near 340 and 410 cm^{-1} [33,34], but the intensity of the last band is lower in comparison with the main band at 340 cm^{-1}.The most prominent band near 410 cm^{-1} in glassy GeO_2 is identified as symmetric stretching mode of bridging oxygen in 6-membered GeO_4 rings [17]. In Raman spectra of suboxide this maximum is shifted to 440 cm^{-1} [17]. So in the difference Raman spectrum (Figure 4) the maximum near 420 cm^{-1} is more reasonably ascribed to the presence of GeO_x species on the surface of the film (z=2.78). A new band at 380 cm^{-1} is seen at x=0.9 in the difference plot of the $(Ge_2S_3)_x(As_2S_3)_{1-x}$ film (Figure 2). Similar peculiarities have been noted in the process of Raman investigation of Ge-O-S films [17]. This band was assigned [17] to mixed anion (sulfur and oxygen) germanium tetrahedral sites or phase separation of the glass structure. In order to examine the vibrational modes on the surface of the film at x=0.9 a complementary technique, SRPES and ordinary XPS spectroscopy, was used to look for the presence or absence of GeO_x and GeS_x species. SRPES and XPS analysis has been performed on the as-deposited, illuminated and annealed films. The study was carried out on the Ge 3d, As 3d and S 2p, O 1s and C 1s core peaks in Ge_2S_3 glass and evaporated thin films. The values of the binding energies which were found for different germanium sulfide and oxide compounds [19,28,33,37,41] were used as reference energies during SRPES and XPS analysis. The surface and subsurface composition and their structural assignment obtained from the fitting procedure similar to those shown in (Figure 5) are presented in Tables 2 and 3, respectively.

Table 2.
Composition (atomic percentages) and Ge/S ratio obtained by SRPES (615 eV) and ordinary XPS (1486,6 eV) spectra analysis of amorphous films obtained by flash evaporated $Ge_{40}S_{60}$ glass powder

	Energy eV,	Ge	S	C	O	Ge/S
As deposited	615	20	15	52	13	1.33
Irradiated	615	19	14	54	13	1.35
Annealed	615	32	25	25	18	1.29
As deposited	1486,6	17	17	40	26	1.0
Irradiated	1486,6	14	14	41	31	1.0
Annealed	1486,6	23	24	24	20	0.95

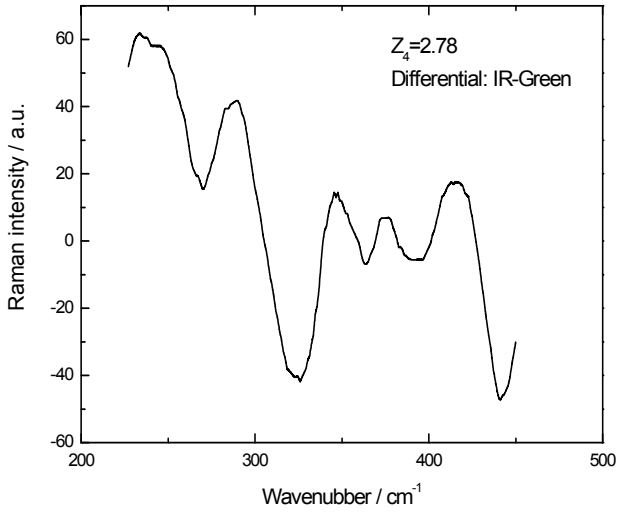


Figure 4. Difference plot between spectra recorded with two different excitation lines: 1.17 and 2.41eV

The C 1s signal was associated mainly with the surface and near-surface region. Carbon is partially removed from the surface and near surface during the annealing process in ultra high vacuum at temperature, less than softening temperature of g-Ge_2S_3 (T_g-30^0), (Table 2). In an earlier XPS investigation of Ge_2S_3-based films [35] the carbon contamination (an atomic percentage of about 15-20 %) has always been determined but was not taken [35] into account during the consideration of surface composition. It can be noted as well, that in Auger spectra of ternary ChGS film during long term aging, the tail of carbon and oxygen signals extended up to 30 nm in depth [36], but 10-20 nm are below the detection limit in the Auger profile [12]. In our case laser illumination of a thin film in air leads to changes of Ge/S ratio on the surface due to the increasing carbonization of the sample surface and the laser stimulated losses of Ge and S (Table 2). Also illumination with energy above the bandgap results in the breakdown of Ge-S bonds and creation of new Ge-O bonds in the irradiated area on the top of the film (Table 3). The analysis of SRPES Ge 3d spectra has shown that after illumination and annealing the concentration of GeO_2 on top reduces and GeO_x increases in deeper layers which is confirmed by the XPS peak signal from 3 nm depth (Table 3). In previous literature for $Ge_{25}Ga_{10}S_{65}$ glasses studied by XPS and XAS, a similar situation can be found [37]. XPS data showed that $Ge_{4/4}$ exists in small quantities in the deeper layers of the illuminated zone and might be connected with the fact that part of the germanium oxides decompose into $Ge_{4/4}$ s.u. after illumination [38]. As already stated, $Ge_{4/4}$ s.u. were found in the illuminated zone too, during measurement of Raman spectra (Figure 2) which confirmed the XPS spectrum assignment (Table 3). It is known that evaporation of Ge-based glasses is often a non-congruent vaporization process which can lead to off-stoichiometry films [10]. In [39] the composition of the thermally deposited film based on $Ge_{40}S_{60}$ glass was determined as $Ge_{46}S_{54}$ and did not matches well with that of the bulk glass. In the last case, the ratio Ge/S =0.85 is slightly lower in comparison with the Ge/S ratio on the top (1.33) and deeper layers (1.0) of flash evaporated samples from the same nominal glass composition (Table 2).

Table 3.
Individual components of Ge 3d spectra determined from curve fitting as evaporated, irradiated and annealed films (Table 1) and their contribution (%) to total peak area. The main ($3d_{5/2}$) peak of the doublet is considered.

Peak number	Core level/ Component	As evaporated sample, peak area, %	Irradiated sample, peak area, %	Annealed sample peak, area, %
	Ge3d, 615 eV			
peak 1	S3Ge-GeS3 and GeS-like	16	13	-
peak 2	$GeS_{4/2}$	33	32	19
peak 3	GeO_x	34	37	57
peak 4	GeO_2	17	18	14
	Ge3d, 1486 eV			
peak 1	$Ge_{4/4}$	-	2	-
peak 2	S_3Ge-GeS_3 and GeS-like	25	22	-
peak 3	$GeS_{4/2}$	43	46	25
peak 4	GeO_x	30	30	29
peak 5	GeO_2	2	12	-

Table 4.
Individual components of Ge 3d spectra determined from curve fitting of bulk glass $Ge_{40}S_{60}$ surface and their contribution (%) to total peak area. The main ($3d_{5/2}$) peak of the doublet is considered.

	Ge3d,1486 eV	Area, %
peak 1	$Ge_{4/4}$	-
peak 2	S_3Ge-GeS_3 and GeS-like	24
peak 3	$GeS_{4/2}$	74
peak 4	GeO_x	-
peak 5	GeO_2	2

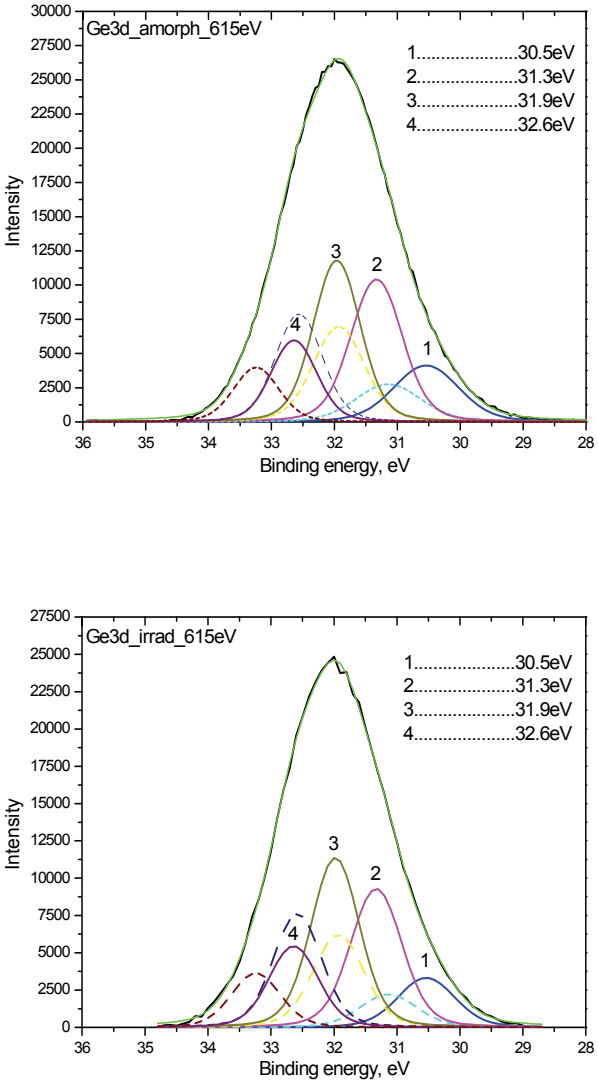


Figure 5. SRPES spectra of the flash evaporated Ge_2S_3 films before and after irradiation, with the fitted peaks

During fitting of SRPES and XPS C 1s core level spectra (not shown here) there is no indication of the Ge-C bonds formation (binding energy 284.3 eV [40]). Only C-S and/or C-O bonds where found at 286 eV. That is why we consider the structural composition of Ge_2S_3 films on the surface and subsurface as mixing of structural units: $GeS_{4/2}$×c-GeS-like×S_3Ge-GeS_3×$GeO_{4/2}$×$GeO_{2/2}$. The ratio $GeS_{4/2}$×c-GeS-like×S_3Ge GeS_3/$GeO_{4/2}$×$GeO_{2/2}$ (see data in Table 3) on the surface of films before illumination is near 1 and drastically differs from those represented by a bulk sample with polished surface (Tabl.4). We suggested that Ge oxidation plays an important role in radiative recombination processes from the surface of Ge_2S_3 - based films. The obtained results showed that annealing removed c-GeS-like×S_3Ge-GeS_3 species from both the surface and near surface region. Annealing did not change the composition and local coordination of the films (Table 3) to those in bulk glass (Table 4.)

IV. Conclusion

We suggested that oxidation of Ge containing species on the top and subsurface layers of Ge_2S_3-based films have a leading role in radiative recombination process from the surface of films. The changes of surface stoichiometry of amorphous Ge_2S_3 films exposed to air are found to be mainly due to germanium oxidization and carbonization of the surface and subsurface layers. The laser illumination of thin films in air leads to changes of Ge/S ratio on the surface and subsurface due increasing carbonization on top and oxidation, carbonization of the sample subsurface layers and the laser stimulated losses of Ge and S. Annealing removed c-GeS-like×S_3Ge-GeS_3 species and partially carbon from both the surface and near surface region, but did not move the surface toward the composition and local coordination of bulk glass.

References

[1] T. Kolomiets, T.K.Mamontova, V.V. Negreskul, "On the recombination radiation of vitreous semiconductors", Phys.Stat. Soldi, vol.27, pp.K15-K17,1968 .

[2] K.Arai, U.Itoh, H.Hamikawa, "Photoluminescence in Ge-S glasses", Jap. J. Appl. Phys,, vol.13, pp.1305-1306, 1974.

[3] V. A.Vasil'Ev, B. T. Kolomiets,.T. N.Mamontova, I. G. Khr., "Radiative recombination in the vitreous semiconductors Ge_2S_3 Ge_2Se_3, and Ge-Pb-S", ZhETF, Pisma v Redaktsiiu, Vol. 21, p.183-185,1975

[4] K.Murayama, M. A.Bösch, "Radiative recombination in crystalline As_2S_3",Physical Review B, vol. 23, pp.6810-6812, 1981.

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

[6] R. P. Wang, S. J. Madden, C. J. Zha, A. V. Rode, and B. Luther-Davies,''Annealing induced phase transformation in amorphous As_2S_3 films,'' J. Appl.Phys., vol.100, 063524-1–4 ,2006.

[7] V.Mitsa, Y.Babinets, Y.Gvardionov and I.Yermolovich, "Photoluminescence in $Ge_xAs_yS_{1-x-y}$ glasses by varying the average coordination number. J. of Non-Crystalline Solids", vol.137&138, pp.959-962, 1991.

[8] N.V. Bondar, N.A. Davydova V.V. Tishchenkoa, M. Vlcek, "Bond-conversion model for photoinduced effects in glassy Ge–S chalcogenides",Journal of Molecular Structure, vol. 555 , pp. 175–178,2000.

[9] A. V. Mitsa, I. V. Fekeshgazy, A. V.Ugrin "Optical coatings based on non-crystalline films with transition substrate-film layers: Sims and Auger profiles",Chalcogenide Letters, vol. 2, pp. 5 – 7, 2005.

[10] Mitsa, V. Mitsa, A. Ugrin, "Mathematical modeling of spectral characteristics of optical coatings with slightly inhomogeneous chalcogenide films", Journal of Optoelectronics and Advanced Materials,vol. 7, pp. 955 – 960, 2005.

[11] A. Macleod, Thin -films optical filters. Third ed., Bristol and Philadelphia: Institute of physics publishing,2001.

[12] C. Maurel, T. Cardinal, P. Vinatier,L. Petit, K. Richardson, N. Carlie, F. Guillen, M.Lahaye, M. Couzi , F. Adamietz, V. Rodriguez, F. Lagugné-Labarthet,V. Nazabal, A.Royon, L. Canioni, "Preparation and characterization of germanium oxysulfide glassy films for optics", Materials Research Bulletin , vol.43,pp. 1179-1187,2008.

[13] B. A. Wilson , "Super-band gap radiation in a-Si", AIP Conf. Proc. 73, pp. 273-277, 1981.

[14] V.M. Mitsa, R.M. Holomb, G. Lovas, M. Ivanda, G.Y. Rudyko, E.G. Gule, I.V. Fekeshgazi, "Room temperature visible luminescence in wide band gap chalcogenide glasses",2012 MIPRO Proceedings of the 35th International Convention,pp. 21-22,2012.

[15] X.Liu, M. Naftaly, A.Iha "Spectroscopic evidence for oxide dopant sites in GeS_2- based glasses using visible photoluminescence from Pr3+ probe ions", Journal of Luminescence, vol.96, pp.227-238,2002.

[16] T.Nakanishi, Y. Tomii, K.Hachiya. "Temperature dependence of the photoinduced fatigue-recovery phenomena of photoluminescence under prolonged irradiation in GeS_2 chalcogenide glass". J. of Non-Crystalline Solids, vol.354, pp.1627-1632, 2008.

[17] D.V.Martin,V.A.Volodin,H.Rinnert, M.Vergnat, "Anomalous temperature dependence of photoluminescence in GeO_x films and GeO_x/SiO_2 nanostructure", Pis'ma v ZhTF, vol.95, pp.472,-476. 2012.

[18] K. Tanaka, Excitation-energy-dependent photoluminescence in glassy As–S and crystalline As_2S_3, Phys. Status Solidi B, vol.1–6 (2013) / DOI 10.1002/pssb.201248519

[19] Mingfa Peng, Yang Li, Jing Gao, Duo Zhang, Zheng Jiang, Xuhui Sun, "Electronic Structure and Photoluminescence Origin of Single-Crystalline Germanium Oxide Nanowires with Green Light Emission", J. Phys. Chem. C, vol.115, pp.11420–11426,2011.

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

[21] K.W.Sun, S.H.Sue, C.W.Liu, "Visible luminescence from Ge quantum dots", Physica E , vol.28,pp.525-530, 2005.

[22] Characterization of Semiconductor Heterostructures and Nanostructures, Ed.Carlo Lamberti, P.B. "Elsevier", 2008

[23] Wang, Rong-Ping, Rode, Andrei; Madden, Steve, Luther-Davies, Barry, "Physical aging of arsenic trisulfide thick films and bulk materials" Journal of the American Ceramic Society, vol.90, pp.1269-1271, 2007.

[24] O. Kondrat , N. Popovich , R. Holomb , V. Mitsa , V. Lyamayev , N. Tsud, V. Cháb ,V. Matolín , K.C. Prince,"Synchrotron radiation photoelectron spectroscopy studies of self-organization in $As_{40}Se_{60}$ nanolayers stored under ambient conditions and after laser irradiation", Journal of Non-Crystalline Solids,vol. 358, pp. 2910–2916,2012.

[25] R.A. Street," Luminescence in amorphous semiconductors", Advances in Physics, vol.25,pp. 397- 415, 1976

[26] A. V. Kolobov, Photo-induced metastability in amorphous semiconductors, Wiley-VCH ,2003.

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

[28] Heqing Yang , Xi Yao , Xingjun Wang , Xiaoxiao Gu , Fujian Wang , "Visible room-temperature photoluminescence of Ge-doped SiO_2 glasses fabricated by a sol–gel process", Optical Materials, vol. 29, pp.631–635,2007.

[29] Y Batra , D Kabiraj , D Kanjilal," Development of Ge nanoparticles embedded in GeO_2 matrix",. J Nanosciens Nanotechnol ,vol.2,pp.4081-4085, 2008 .

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

[31] A. S. Zyubin,A. M.Mebel,S.H.Lin, "Photoluminescence of oxygen-deficient defects in germanium oxides: A quantum chemical study", The Journal of Chemical Physics, vol. 125,pp. 064701-0647011,2006.

[32] Grant S. Henderson , Daniel R. Neuville , Benjamin Cochain , Laurent Cormier ,"The structure of GeO_2–SiO_2 glasses and melts: A Raman spectroscopy study", Journal of Non-Crystalline Solids, vol. 355 ,pp.468–474,2009.

[33] Hiromichi Takebe, Hiroki Maeda, Kenji Morinaga, "Compositional variation in the structure of Ge-S glasses", J. Non-Cryst.Sol , vol.291, pp.14-24,2001.

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

[35] V. Pamukchieva, E. Skordeva, D. Arsova, M.-F. Guimona, D. Gonbeau, **"**Changes in the electronic structure of Ge-As-S thin films after illumination", Journal of Optoelectronics and Advanced Materials, vol. 7, pp. 1265 – 1270,2005.

[36] T.N.Shchurova, N.D. Savchenko, K.O. Popovic, N.Yu. Baran, "Depth profiling of the near-surface layer for $Ge_{33}As_{12}Se_{55}$ thin films", Chemistry, phisics and technology of surface, (Ukraine), vol.1, pp.343–347,2010.

[37] P.N. Lisboa-Filhoa, V.R. Mastelarob,W.H. Schreinera, S.H. Messaddeqc, M. Siu Lib,Y. Messaddeqc, P. Hammerc, S.J.L. Ribeiroc, P. Parentd, C. Laffond ,"Photo-induced effects in $Ge_{25}Ga_{10}S_{65}$ glasses studied by XPS and XAS",Solid State Ionics vol.176,pp.1403–1409,2005.

[38] Peter Broqvist, Jan Felix Binder, Alfredo Pasquarello, "Formation of substoichiometric GeO*x* at the Ge–HfO_2 interface", Applied Physics Letters vol.97, pp.202908- 202911,2010.

[39] M. Mitkova, A. Kovalskiy, H. Jain, Y. Sakaguchi, "Effect of photo-oxidation on the photodiffusion of silverin germanium chalcogenide glasses", Journal of Optoelectronics and Advanced Materials, vol. 11, No. 12, pp. 1899 – 1906,2009.

[40] C.Q. Hua, W.T. Zhenga, B. Zhenga, J.J. Lia, Z.S. Jina, X.M. Baia, H.W. Tiana,Q. Jianga, X.Y. Wanga, J.Q. Zhub, S.H. Mengb, X.D. Heb, J.C. Han, "Chemical bonding of a-$Ge_{1-x}C_x$:H films grown by RF reactive sputtering", Vacuum, vol. 77, pp. 63–68, 2004.

[41] F.Mirabella,R.L.Johnson,J.Ghijsen,"Photoemission investigations of manganese thin films deposited on GeS(001) surfaces", Surface Science, vol.506,pp. 172–182,2002.

Light coupling between the whispering gallery modes of a coated microsphere and a tapered fiber

Davor Ristić[1], Maurizio Mazzola[1], Andrea Chiappini[1], Patrice Féron[2], Alphonse Rasoloniaina[2], Gilles Cibiel[3], Mile Ivanda[4], Giancarlo C. Righini[5,6], Maurizio Ferrari[1]

[1] IFN – CNR CSMFO Lab., Via alla Cascata 56/C Povo, 38123 Trento, Italy.
[2] ENSSAT-FOTON (CNRS-UMR 6082), 22300 Lannion, France.
[3] Centre National d'Etudes Spatiales (CNES), 31401 Toulouse Cedex 9, France
[4]Institut Ruđer Bošković, Bijenička cesta 54, 10000 Zagreb, Croatia
[5] IFAC - CNR, MDF Lab., 50019 Sesto Fiorentino, Italy
[6]Museo Storico della Fisica e Centro di Studi e Ricerche "Enrico Fermi", Piazza del Viminale 2, 00184 Roma, Italy.

Abstract - Spherical microresonators were coated using a $70SiO_2 – 30HfO_2$ glass doped with 0.3 mol% Er^{3+}. The sphere was characterized by coupling a 1480 nm pump laser inside the sphere using a tapered optical fiber and observing the 1530-1580 nm Er^{3+} emission outcoupled with the same tapered fiber. The measurements were done for different coating thicknesses and compared with theoretical calculations to understand the relationship of the detected signal with the whispering gallery mode electric field profiles.

I. INTRODUCTION

Silica microspheres are microsized resonators with a very high quality factor and a low mode volume. The modes of the microsphere are called the whispering gallery modes (WGM) [1, 2]. This makes them very useful for different applications where high light density (non linear optics[3], lasing [4-6], quantum electrodynamics[7]) or sharp resonances (sensing [8], metrology [9]) are required. A common way to couple light inside a microsphere is using evanescent coupling from a tapered fiber. The coupling strength can depend on two parameters: on the overlap of the evanescent fields and on the phase-matching between the whispering gallery modes in the sphere and the propagating mode of the taper. The phase matching condition is primarily dependent on the refractive index of the sphere which cannot be changed once the sphere is produced. We present a method to gradually change the effective refractive index of the sphere by coating the sphere with a material with a known refractive index. Silica-hafnia is a very suitable material for this application since it has very good optical and spectroscopic properties [10,11].

II. EXPERIMENTAL

The microspheres were produced by melting the tip of a standard optical telecommunications fiber. The diameter of the microsphere was found to be 140±10 μm using a standard optical microscope. The microspheres were coated with a film with the molar composition $70SiO_2$-$30HfO_2$ activated with 0.3% mol Er^{3+}. The films were coated using a dip coating technique. The sol was prepared by mixing a solution of $Si(OC_2H_5)_4$ (TEOS), HCl, H_2O, CH_3CH_2OH (ethanol) and $HfOCl_2$. The molar ratio between TEOS, water and the catalyst (HCl) was $n(H_2O) : n(Si(OC_2H_5)_4) : n(HCl) = 200:100:1$. The molar ratio between the TEOS and $HfOCl_2$ was chosen to be $n(Si(OC_2H_5)_4) : n(HfOCl_2) = 70:30$ on the basis of previous experiments [12]. The overall amount of Si and Hf in the sol was controlled so that $c(Si(OC_2H_5)_4) + c(HfOCl_2) = 4.48\ 10^{-4}$ mol/mL in order to have a sol viscosity which makes the dipping rate 30 nm/dip. Er $(NO_3)_3$ x $5H_2O$ was added to activate the film so that the molar concentration of Er^{3+} in the sol is 0.3 %. The mixture of TEOS, HCl and H_2O was prehydrolised at 65 ^oC for 1 hour before adding the $HfOCl_2$ and the Er $(NO_3)_3$ x $5H_2O$. The entire mixture was then stirred for 16 h until the sol became transparent. The microsphere was dipped 30 times into the sol. The dipping speed was 40 mm/min and the dipping time 20s. After each dip the microsphere was densified at 900 ^oC for 50 s and after every 10 dips an additional densification was made for 2 min. At the end of the deposition a final densification for 5 min at 900 ^oC was made.

The measurement of the whispering gallery modes was done using a microsphere-taper coupling setup. A 1480 nm (S-band) fibre diode laser was used for the pumping. The pump laser was made pass through a tapered fiber in order to achieve coupling with the microsphere. The coupling is achieved by moving the taper close to the sphere so that the evanescent field of the taper and of the

whispering gallery modes of the microsphere overlap. The pump laser is coupled inside the microsphere where it is absorbed by Er^{3+} ions embedded in the coating which in turn emit in the C-band wavelength region around 1550 nm. The emission will be greatly enhanced at wavelengths which correspond to the WGMs of the microsphere. This emitted light is then coupled out of the sphere and into the taper and detected using a Anritsu MS9710B optical spectrum analyzer (OSA).

III. THEORY

To calculate the eigenfrequencies of the whispering gallery modes of a coated sphere the Aden-Kerker generalization of the Mie scattering theory can be used [13,14]. The whispering gallery modes will be characterized by three integer numbers *n*, *l* and *m* as is the case for any system with spherical symmetry. The radial part of the electric and magnetic fields of the solution of the radial Helmholtz equation is proportional to the Bessel function in the sphere core, a linear combination of the Bessel and Neumann functions in the coating and the Neumann function in the surrounding medium (usually air). The continuity of the tangential components of the electric and magnetic fields at the core-coating and coating-air interfaces leads to a system of four equations which can be solved numerically to obtain the eigenfrequencies of the WGMs. The eigenfrequencies will be degenerate in respect to *m* and will generally be dependent only on *l* and *n*. Furthermore, the modes closest to the surface of the sphere will be usually those where *n*=0. Once the eigenfrequencies are known the effective refractive index can be defined as $n_{eff}=l/(kR)$ where $k=2\pi/\lambda$ is the eigenfrequency of the mode and *R* is the radius of the sphere [15]. In Figure 1 is shown the effective index of the TE fundamental mode in dependence on the sphere diameter for a mode at 1.56 μm and for the coating refractive index of 1.6 corresponding to $70SiO_2$-$30HfO_2$ and sphere refractive index of 1.44 corresponding to silica. We can see that the effective refractive index changes gradually between the values for a blank SiO_2 sphere and a blank $70SiO_2$-$30HfO_2$ sphere upon increasing the coating thickness. For coating thicknesses above 2 μm the sphere is very close to the bulk $70SiO_2$-$30HfO_2$ sphere case. Since the refractive index of the taper is 1.44, the effective index of the propagation mode of the taper will be 1.44 or less. Therefore from Figure 2 we can see that with increasing thickness from 0 to 2 μm the phase mismatch between the sphere and the taper will increase, saturating to a certain value above 2 μm thickness. If the coupling condition is dependent on the phase mismatch [16] we would expect an decrease of the coupling efficiency with increasing thickness. Figure 2 shows the mode profiles for the fundamental TE mode at 1.56 μm for a sphere with 0.2 μm, 0.6 μm and 1.2 μm thick coating, respectively. The overall increase of the percentage of the mode confined inside the coating with the increase of the coating thickness is evident. Since the coating is the optically active part of the sphere, the increase of the percentage of the electric filed inside the coating should lead to the increase of the luminescence intensity of the WGMs. Therefore the overall luminescence intensity should depend on two factors related to the coating thickness. With the increase of the coating thickness the phase mismatch should lead to the decrease of the luminescence

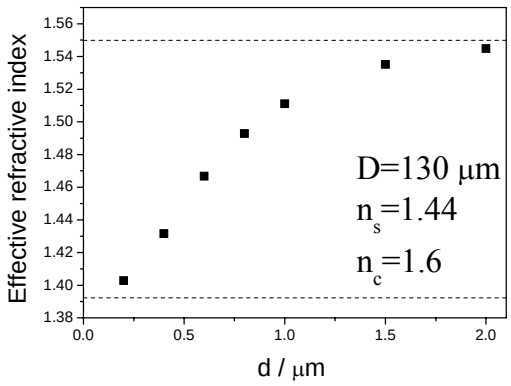


Figure 1. The effective refractive index of the fundamental (*n*=0) TE WGM of the coated microspheres at 1560 nm in respect of the coating thickness. The dashed lines correspond to the border cases of no coating (blank sphere) or a very thick coating (bulk sphere with n=1.6).

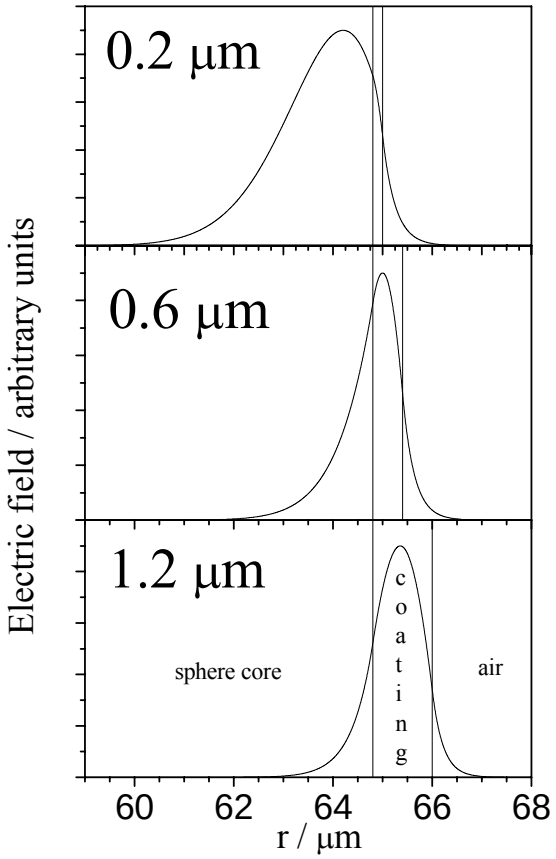


Figure 2. The electric filed profiles of the fundamental (*n*=0) TE WGM of the coated microspheres at 1.56 μm. The refractive index of the sphere and coating is 1.44 and 1.6, respectively. In the image the coating thicknesses are reported.

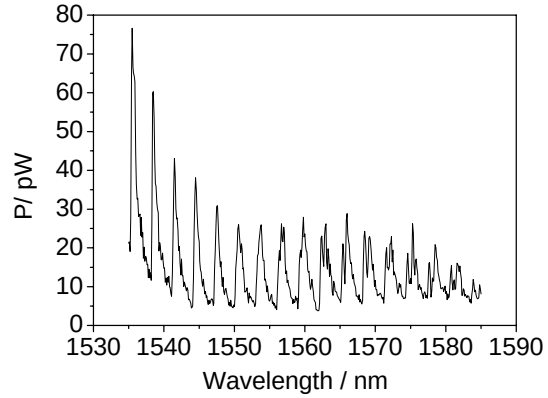


Figure 3. The WGM luminesnce for a coated sphere after 50 dips

intensity while the increased confinement of light inside the optically active layer should lead to the increase of the luminescence intensity.

IV. RESULTS

The sphere was coated using a $70SiO_2$-$30HfO_2$ activated with 0.3% mol Er^{3+}. This material has a refractive index of 1.6. The dipping rate, when dipping on a flat 3'' silica slide using the same protocol, is around 30 nm/dip as measured my m-line [10]. Although it is possible that the dipping rate on a microsphere can differ from the dipping rate on a silica slide, we assume that this difference is slight. A typical luminescence spectrum of the WGMs is shown in Figure 3. The WGMs with a free spectral range of about 3.4 nm corresponding to different values of *l* are clearly visible. The luminescence was measured every 5 dips starting from 20 dips to 50 dips. Special care was taken to always choose the same coupling point between the sphere and the taper for all the measurements. In Figure 4 is shown the overall

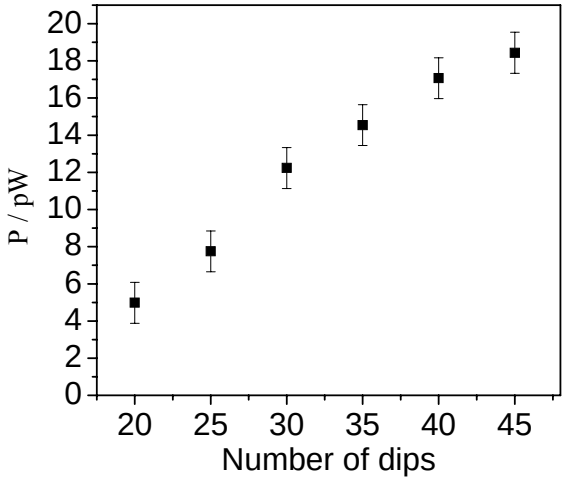


Figure 4. The average overall WGMs luminescence intensity in respect to the number of dips

luminescence intensity in respect to the number of dips. A continuous increase of the luminescence intensity with the number of dips is clearly visible. This leads to the conclusion that the phase matching condition is not the most important one for this kind of coupling scheme. The fact that increased thickness means more active material on the sphere appears to be more important.

V. CONCLUSION

We have presented the procedure for the coating of silica microspheres using a sol-gel dip-coating technique. We coated the microspheres with optically activated silica-hafnia thin films and studied the effect of the increase of the coating thickness on the WGM luminescence intensity. We concluded that the phase matching between the WGMs of the coated sphere and the tapered fiber is not so important as the density of the excited ions.

ACKNOWLEDGMENT

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

REFERENCES

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

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

[3] I. Razdolskiy, S. Berneschi, G. Nunzi Conti, S. Pelli, T. V. Murzina, G. C. Righini and S. Soria, "Hybrid microspheres for nonlinear Kerr switching devices", Optics Express vol. 19, pp. 9523-9528, 2011.

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

[5] Z. Cai, A. Chardon, H. Xu, P. Féron and G. Stephan, "Laser characteristics at 1535 nm and thermal effects of an Er:Yb phosphate glass microchip pumped by Ti:sapphire laser", Optics Commun., vol. 203, pp. 301, 2002.

[6] S. M. Spillane, T. J. Kippenberg and K. J. Vahala, "Ultralow-threshold Raman laser using a spherical dielectric microcavity", Nature, vol. 415, pp. 621-623, 2002.

[7] D. W. Vernooy, A. Furusawa, N. Ph. Georgiades, V. S. Ilchenko, and H. J. Kimble, "Cavity QED with high-Q whispering gallery modes", Phys. Rev. A, vol. 57, pp. R2293-R2296, 1998.

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

[9] J. Li, H. Lee, T. Chen, and K. J. Vahala, "Low-Pump-Power, Low-Phase-Noise, and Microwave to Millimeter-Wave Repetition

Rate Operation in Microcombs", Phys Rev Lett., vol. 109, pp. 233901, 2012.

[10] R. R. Goncalves, G. Carturan, L. Zampedri, M. Ferrari, M. Montagna, G. C. Righini, S. Pelli, S. J. L. Ribeiro Y. and Messaddeq, "Erbium-activated HfO2-based waveguides for photonics", Opt. Mater., vol. 25, pp. 131-139, 2004.

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

[12] A. Chiappini, A. Chiasera, C. Armellini, S. Varas, A. Carpentiero, M. Mazzola, E. Moser, S. Berneschi, G. C. Righini, M. Ferrari, "Sol-gel-derived photonic structures: fabrication, assessment, and application", Journal of Sol-Gel Science and Technology, vol. 60, pp. 408-425, 2011.

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

[14] M. Han and A. Wang, "Temperature compensation of optical microresonators using a surface layer with negative thermo-optic coefficient", Optics Letters, vol. 32, pp. 1800, 2007.

[15] A. Chiasera, Y. Dumeige, P. Féron, M. Ferrari, Y. Jestin, G. Nunzi Conti, S. Pelli, S. Soria and G. C. Righini, "Spherical whispering-gallery-mode microresonators", Laser & Photon. Rev., vol. 4, pp. 458, 2010.

[16] B. E. Little, "Analytic theory of coupling from tapered fibers and half-blocks into microsphere resonators", Journal of Lightwave technology, vol. 17, pp. 704, 1999.

Eu^{3+}/ Yb^{3+}- activated nanocrystals of lanthanide phosphates

A. Lukowiak[*,**], A. Bednarkiewicz[*], R.J. Wiglusz[*], W. Strek[*], L. Kepinski[*], L. Krajczyk[*], D. Ristic[**], A. Chiasera[**], and M. Ferrari[**]

[*] Institute of Low Temperature and Structure Research, PAS, Okolna 2, 50-422 Wroclaw, Poland
[**] CNR-IFN, CSMFO Lab., Via alla Cascata 56/c, Povo, 38123 Trento, Italy
A.Lukowiak@int.pan.wroc.pl

Abstract - The paper presents structural and luminescence studies of two lanthanide phosphate composites codoped with different concentrations of Eu^{3+} and Yb^{3+} ions. The powders of polycrystalline phosphates were synthesized by the modified Pechini method. The composition and morphology of the samples were characterized using inductively coupled plasma atomic emission spectrometry and transmission electron microscope. The crystal structures of the powders were obtained from the powder diffraction data and were refined using the Rietveld method showing the existence of two or three phases in the materials. The absorption, excitation, and emission spectra of the phosphates have been studied. The upconversion luminescence has been observed in the Eu^{3+}/Yb^{3+}-doped samples and the energy transfer processes were discussed.

I. INTRODUCTION

Rare earth phosphate materials have been extensively investigated due to their structural diversity and optical properties that make them suitable for several photonic applications [1]. Although the lattice parameters and spectroscopic properties of numerous lanthanide tetraphosphates (such as $LiLn(PO_3)_4$) have been reported previously [2-4], the research concentrated mostly on single crystals. First reports concerning nanocrystalline materials have been published only recently [5-8]. The material is highly interesting due to large distance between cations (e.g. 5.66 Å in $LiNd(PO_3)_4$ [2]) that prevents non-radiative concentration quenching even for stoichiometric crystals and make them suitable for low threshold lasers or efficient phosphors.

Europium and terbium lanthanides are very attractive luminescence centers due to large energy gap between the emitting level ($^5D_0$ for Eu^{3+} and $^5D_4$ for Tb^{3+}) and low states (^7F_J, J=0..6). Both ions, when codoped with sensitizing Yb^{3+} ions may be potentially used for many practical applications like quantum cutting for solar energy converters or luminescent bio-markers. In the first case, high energy photons are cut into two or more low energy photons in NIR region (≈1μm) that match silicon energy gap [9]. In the second case, NIR photoexcitated emission of Eu^{3+} or Tb^{3+} could be interesting alternative to Yb/Er, Yb/Tm, Yb/Ho upconverting luminescent nanomarkers to broad the selection of available tags or serve as donors in FRET (*Fluorescence Resonance Energy Transfer*) based bio-assays [10, 11].

Europium doped polyphosphate (nano)powders have been studied recently by Matraszek et al. [6], Wiglusz et al. [7], Ferhi et al. [8], and Han et al. [12].

Here we present the synthesis of phosphate nanopowders codoped with Eu^{3+} and Yb^{3+} ions and their structural and spectroscopic assessment.

II. EXPERIMENTAL

A. Synthesis

Two polyphosphates were synthesized by the modified Pechini method with the nominal composition $LiEu_{0.1}La_{0.5}Yb_{0.4}(PO_3)_4$ and $LiEu_{0.2}Yb_{0.8}(PO_3)_4$ that were further abbreviated as **10Eu40Yb** and **20Eu80Yb,** respectively. Appropriate amounts of the lanthanide oxides (Eu_2O_3, Yb_2O_3, and La_2O_3) with the purity of at least 99.99%, as well as lithium carbonate, were digested in diluted, hot nitric acid in order to obtain the nitrate salts soluble in water. The substrates were mixed together and $(NH_4)_2HPO_4$ dissolved in water was gradually added. The nitric acid was used to obtain low pH of the solution in order to avoid the precipitation. Then, the citric acid and ethylene glycol were added into the solution and mixed. The sol was dried at 80 °C for 72 h and afterwards, heat treated at 850 °C for 12 h.

B. Analytical methods

The composition of the powders was determined using inductively coupled plasma atomic emission spectrometry analysis (ICP-AES). The structure and morphology were characterized by X-ray diffraction (Stoe STADI P transmission diffractometer; Cu $K\alpha_1$ radiation 1.54060 Å) and high resolution transmission electron microscope (Philips CM20 SuperTwin TEM). The absorption spectra

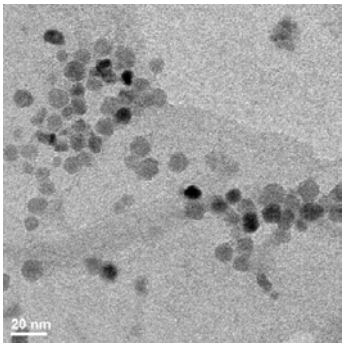


Figure 1. The TEM images of Eu10Yb40 sample

of the powders were obtained in reflectance mode using Varian Cary 2300 spectrometer. Emission spectra were measured with a Jobin-Yvon THR 1000 spectrophotometer equipped with Hamamatsu R928 photomultiplier and 1200 l/mm grating. As the excitation source, pulsed (10Hz) 4th harmonics (266 nm) of Nd:YAG laser was used. A LeCroy Wave Surfer 425 digital oscilloscope was used for data collection. For the upconversion studies, samples were excited with 976 nm radiation by a diode laser. The excitation spectra (λ_{mon}= 592 and 976 nm) were measured with Spectra Pro 750 system. The 450W xenon arc lamp was used as an excitation source. A 1800 l/mm gratings have been used in both excitation and emission monochromators. Excitation spectra have been corrected for the excitation light intensity. All the measurements were done at room temperature.

III. Results and Discussion

The sol-gel (modified Pechini) method was used to prepare polycrystalline lanthanide phosphates. The ICP-AES has confirmed the composition of the powders. However, it has also indicated that traces of erbium ions were presented in the materials despite using ultra-pure lanthanide oxides during synthesis. In the final composition of the investigated samples, 0.005 mol% of Er^{3+} was detected.

The TEM image of 10Eu40Yb is shown in Fig. 1. The nanoparticles exhibit diameters between 5 and 15 nm. Similar morphology was observed for 20Eu80Yb.

The XRD diffraction patterns show a complex structure of the samples (Fig. 2). The 10Eu40Yb and 20Eu80Yb powders are composed of three and two phosphate phases, respectively. Table I gives unit cell parameters, crystal cell volume, phases percentage, and refined factors calculated for the investigated samples according to the Rietveld method.

The absorption spectra (Fig. 3) evidence the presence of both europium and ytterbium ions in the 10Eu40Yb and 20Eu80Yb powders. Typical absorption bands at 360 nm, 380 nm, and 394 nm correspond to absorption from the ground state $^7F_0$ level of Eu^{3+} ions to $^5D_4$, ^5G_J, $^5L_7$, and $^5L_6$ levels (Fig. 3(a)). In the visible range, two bands assigned to the transition from $^7F_0$ to $^5D_2$ and $^5D_1$ levels, are observed (Fig. 3(a)). Additionally, Yb^{3+} absorption at 940–1020 nm ($^2F_{7/2}\rightarrow{}^2F_{5/2}$) may be found (Fig. 3(b)). The traces of erbium ions were not detected in the absorption spectra.

Under UV (266 nm) excitation, characteristic emission spectra of Eu^{3+} ions (Fig. 4(a)) were observed. The spectra are similar to the results obtained earlier for the nanoparticles of $LiEu(PO_3)_4$ [7]. The observed bands with the maxima at 578 nm, 590 nm, 611 nm, 649 nm, and 698 nm are assigned to the transitions from the excited $^5D_0$ state to ^7F_J ground state levels (J = 0, 1, 2, 3, and 4). Two components of $^5D_0\rightarrow{}^7F_0$ transition, found for 10Eu40Yb, suggest that Eu^{3+} ions occupied more than one crystallographic site in the structure. The intensities of the $^5D_0\rightarrow{}^7F_1$ and $^5D_0\rightarrow{}^7F_2$ transitions are comparable

TABLE I. Unit cell parameters, crystal cell volume, phase percentage, and refined factors calculated according to the Rietveld method.

Sample		Cell parameters (sample / reference*)	Phase percentage [%]	Rw [%]
10Eu 40Yb	$LiLnP_4O_{12}$	***a* = 9.801(0)Å** / 9.844(2)Å ***b* = 7.087(7)Å** / 7.008(3)Å ***c* = 13.298(3)Å** / 13.25(2)Å β = **90.02(0)°** / 90.1(2)° ***V* = 923.79Å^3** / 914.07Å^3	65.05	1.24
	LnP_3O_9	***a* = 13.941(9)Å** / 14.0630(5)Å ***b* = 20.325(0)Å** / 19.9811(7)Å ***c* = 9.930(9)Å** / 10.0357(3)Å β = **127.51(9)°** / 127.600(2)° ***V* = 2232.02 Å^3** / 2234.23 Å^3	21.90	1.24
	$LnPO_4$	***a* = 6.826(7)Å** / 6.860(1)Å ***c* = 5.978(2)Å** / 6.003 (1)Å ***V* = 278.61Å^3** / 282.5Å^3	13.05	1.24
20Eu 80Yb	$LiYbP_4O_{12}$	***a* = 9.782(6)Å** / 9.844(2)Å ***b* = 7.067(7)Å** / 7.008(3)Å ***c* = 13.259(4)Å** / 13.25(2)Å β = **90.09(3)°** / 90.1(2)° ***V* = 916.65Å^3** / 914.07Å^3	51.96	2.22
	LnP_3O_9	***a* = 14.043(9) Å** / 14.0630(5) Å ***b* = 20.040(7) Å** / 19.9811(7) Å ***c* = 10.00(4)Å** / 10.0357(3) Å β = **127.47(2) °** / 127.600(2)° ***V* = 2234.62 Å^3** / 2234.23 Å^3	48.04	2.22

*reference crystallographic data – $LiLa_xYb_{1-x}P_4O_{12}$ [13], EuP_3O_9 [14], $ErPO_4$ [15]

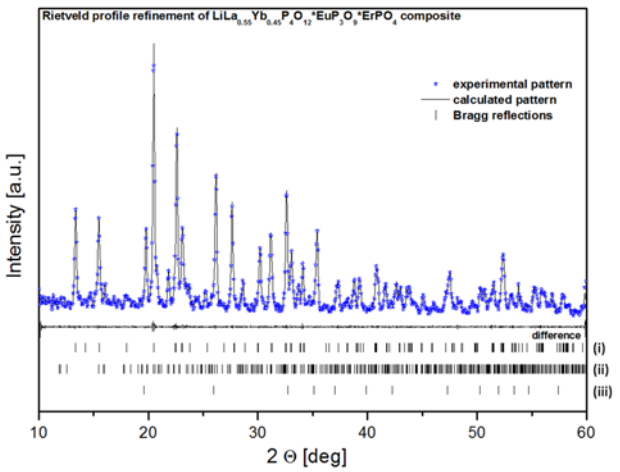


Figure 2. XRD diffraction pattern and Rietveld profile refinement of Eu10Yb40 sample ((i) Card No. 14-ICDS [13], (ii) Card No. 420116-ICDS [14], and (iii) Card No. 36052-ICDS [15])

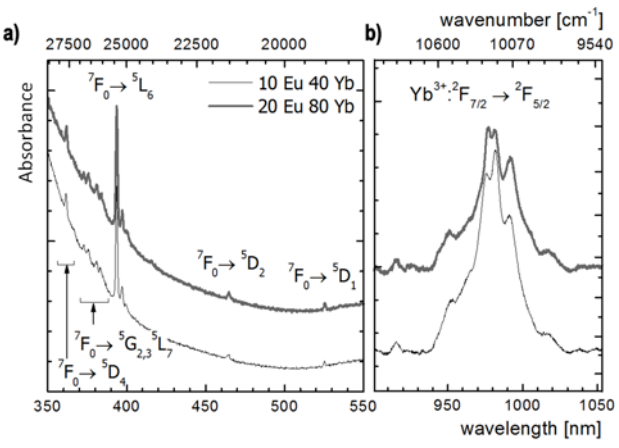


Figure 3. Absorption spectra of Eu^{3+}/Yb^{3+} codoped phosphates

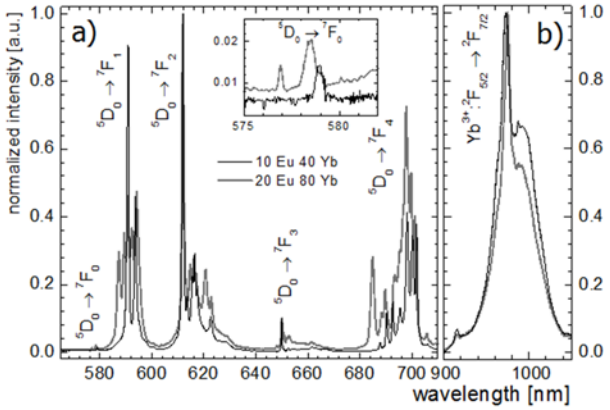


Figure 4. Stokes luminescence spectra of Eu^{3+}/Yb^{3+} codoped phosphates under UV (266 nm) excitation (a). The Yb^{3+} emission through Eu^{3+} pumping can be observed (b)

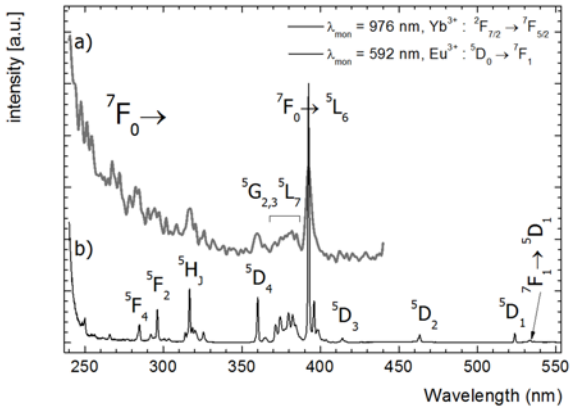


Figure 5. The excitation spectra of 10Eu40Yb under monitoring wavelengths corresponding to $^2F_{5/2}\rightarrow{}^2F_{7/2}$ emission of Yb^{3+} (a) and $^5D_0\rightarrow{}^7F_1$ emission of Eu^{3+} (b)

and also relatively high intensity of $^5D_0\rightarrow{}^7F_4$ is observed indicating a low local distortion at the Eu^{3+} site. All of the transitions are well resolved into Stark components due to splitting of the lower state energy levels of Eu^{3+}. The degeneracy is completed removed for the 10Eu40Yb sample and the well resolved peaks suggest the low symmetry of the Eu^{3+} surrounding. In the case of 20Eu80Yb, the emission band assigned to $^5D_0\rightarrow{}^7F_1$ exhibits only two Stark components indicating a higher local site symmetry.

The wide emission band with a maximum at 976 nm was also observed under UV excitation due to the emission from $^2F_{5/2}$ level of Yb^{3+} (Fig. 4(b)). This emission originates from energy down-conversion from Eu^{3+} to Yb^{3+} ions, and has rarely been studied [16-18]. The possible energy transfer mechanisms between these ions are (i) a cooperative energy transfer from an Eu^{3+} ion in excited $^5D_2$ state to a pair of Yb^{3+} ions that emit two infrared photons and/or (ii) a cross-relaxation process between the $^5D_0\rightarrow{}^7F_6$ (Eu^{3+}) and $^2F_{7/2}\rightarrow{}^2F_{5/2}$ (Yb^{3+}) levels [17, 19]. In the second case, the difference between the respective Eu^{3+} and Yb^{3+} energy levels ($\Delta E\approx2200$ cm^{-1}) may be compensated by phonon emission. The cut-off vibrational energy of these phosphate crystals is around 1250 cm^{-1}.

The photoluminescence excitation spectrum of 10Eu40Yb with monitoring the $^5D_0\rightarrow{}^7F_1$ emission of Eu^{3+} at 592 nm is shown in Fig. 5. The spectrum shows several bands in the UV-VIS range which correspond to the absorption bands of Eu^{3+} (Fig. 3). These absorption bands appear also in the excitation spectrum when monitoring $^2F_{5/2}\rightarrow{}^2F_{7/2}$ emission of Yb^{3+} at 976 nm which is an evidence of the energy transfer (down-conversion) from Eu^{3+} to Yb^{3+}.

The emission from the excited $^5D_0$ state of Eu^{3+} was also observed at similar wavelengths (bands at 578, 590, 611, 647, and 698 nm, Fig. 6) upon 976 nm excitation. Moreover, one can also observe bands in the range of 520–565 nm that can be assigned to the $^5D_1\rightarrow{}^7F_{1,2,3}$ transitions. On the other hand, the origin of some bands in this range can be related to the presence of Er^{3+}. In the blue range, the samples showed intense luminescence with a maximum at 475 nm originating from a cooperative Yb^{3+} pairs emission. The shoulder at around 500 nm may indicate a luminescence arising from $^5D_1\rightarrow{}^7F_0$ and $^5D_2\rightarrow{}^7F_J$ transitions of Eu^{3+} (or Er^{3+}: $^4F_{7/2}\rightarrow{}^4I_{15/2}$). Moreover, a few bands with low intensity were registered in the range of 410–450 nm probably due to the emission from $^5D_3$ level. The emission from higher energy levels of Eu^{3+} (i.e. ^5D_J, J=1, 2, 3, and 4) was also reported in some Eu^{3+}/Yb^{3+} codoped samples under NIR excitation [16, 20]. In other reports [21, 22], similar additional luminescence bands in the blue and green spectral range were associated to the impurities of Er^{3+} and Tm^{3+} ions. We suggest that these emission bands originate from $^5D_{1,2}$ levels of Eu^{3+}, nevertheless, the overlap with Er^{3+} emission ($^2H_{11/2}\rightarrow{}^4I_{15/2}$, $^4S_{13/2}\rightarrow{}^4I_{15/2}$ or $^4F_{9/2}\rightarrow{}^4I_{15/2}$) cannot be excluded in this range. In up-conversion emission mode, the Eu^{3+} emission is relatively weaker than the cooperative Yb-Yb pair emission. Additionally, under 980 nm excitation, the 650 nm band is comparable in intensity with $^5D_0\rightarrow{}^7F_{1,2}$ emission bands, which is not the case for Stokes emission (Fig. 4), where $^5D_0\rightarrow{}^7F_3$ is located at this wavelength. This observation suggests that the band may originate from $Yb^{3+}\rightarrow Er^{3+}$ ($^4F_{9/2}\rightarrow{}^4I_{15/2}$) ETU (*energy transfer upconversion*) emission, which is known to be at least 1-2 orders of magnitude more efficient than the cooperative upconversion [23].

The upconversion emission observed with the NIR excitation could be explained on the basis of the cooperative energy transfer from Yb^{3+} ions to Eu^{3+} ions

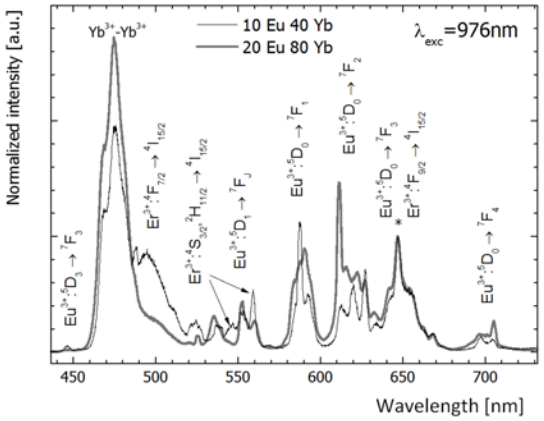


Figure 6. Upconversion emission spectra of Eu^{3+}/Yb^{3+} codoped phosphates under NIR (976 nm) excitation

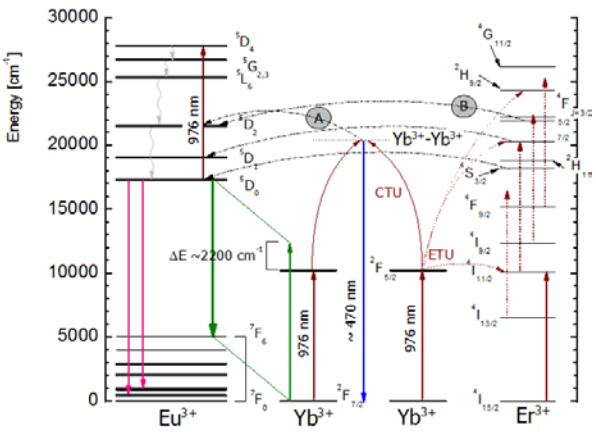


Figure 7. Energy level diagram for $Eu^{3+}/Er^{3+}/Yb^{3+}$ codoped phosphates and possible mechanisms of cooperative energy transfer (A) and energy transfer upconversion (B)

[18, 20, 24, 25]. For phosphate samples doped only with Eu^{3+} no fluorescence is observed upon excitation at 976 nm. However, Yb^{3+} absorbs strongly this radiation and two excited ions cooperatively give a photon with higher energy at 475 nm. Nonradiative energy transfer can occur from Yb^{3+}–Yb^{3+} pair state to $^5D_2$ or $^5D_1$ level of Eu^{3+} that relaxes non-radiatively to the metastable $^5D_0$ state. The Eu^{3+} ions in this state can reabsorb 976 nm photons and get excited to the $^5D_4$ level [16]. A nonradiative relaxation from $^5D_4$ to other ^5D_J levels (e.g. $^5D_1$) give rise to different radiative transitions observed in a wide range of visible spectrum. The schematic mechanism of these processes is shown in the energy level diagram in Fig. 7. The possible energy transfer, mentioned before, between Yb^{3+} and Er^{3+} and between Er^{3+} and Eu^{3+} is also presented in this graph [26, 27].

The luminescence dynamics of the $^5D_0$ state of Eu^{3+} ions in phosphates have been investigated upon 266 nm excitation. The decay curves exhibit a single exponential profile with lifetimes of 2.52 and 3.69 ms for 10Eu40Yb and 20Eu80Yb, respectively. The latter sample exhibits over 30% longer luminescence lifetime in comparison to 10Eu40Yb. This is a further evidence that the Eu^{3+} ions in 20Eu80Yb undergo a higher local site symmetry leading to a reduced electric dipole oscillator strength and resulting in a lengthening of the lifetime. However, due to the high rare earth content in this sample, radiation trapping should be taken into account [28].

IV. Conclusion

Two polyphosphates nanocrystals were synthesized by the modified Pechini method with the nominal composition of $LiEu_{0.1}La_{0.5}Yb_{0.4}(PO_3)_4$ and $LiEu_{0.2}Yb_{0.8}(PO_3)_4$. XRD diffraction patterns showed that nanocrystals are composed of different phosphate phases depending on the Eu/Yb content. Three phases were present for the nanocrystals with 10 mol% Eu^{3+} and 40 mol% Yb^{3+} and two phases for those with 20 mol% Eu^{3+} and 80 mol% Yb^{3+}. The luminescence spectra confirmed a different local structure for the Eu^{3+} ions in these two matrices with more ordered local symmetry in case of the phosphates with higher concentration of Eu/Yb ions. This behavior was reflected by the decay measurements of the $^5D_0$ state of Eu^{3+} ions that gave a longer lifetime in the case of the more concentrated sample. Excitation spectra of Yb^{3+} ions clearly demonstrate an efficient back energy transfer from Eu^{3+} ions. The direct Yb^{3+} to Eu^{3+} energy transfer, achieved by cooperative process, is evidenced by upconversion luminescence obtained upon excitation at 976 nm. An erbium contamination (0.005 mol%) derived from the precursors was measured by ICP-AES and was evidenced by upconverted luminescence.

Although these systems deserve further investigations concerning the luminescence efficiency as a function of the composition, certainly they present useful spectroscopic properties that are worthy of exploitation both for developing glass ceramic waveguides structures and transparent ceramics [29].

Acknowledgment

A.L. thanks the Polish Ministry of Science and Higher Education for the financial support from the “Mobility Plus” Program. A.B. acknowledges support from Polish National Science Centre Grant No. N N507 584938. Part of the research was performed in the framework of NSBMO project (2010-2013) of the Provincia Autonoma di Trento.

References

[1] C.C. Santos, I. Guedes, C.-K. Loong, L.A. Boatner, A.L. Moura, “Spectroscopic properties of Er^{3+}-doped lead phosphate glasses for photonic application”, J. Phys. D: Appl. Phys. 2010, 43, pp. 025102.

[2] M T de Araujo4, C Jacinto4 and M V D Vermelho T. Yamada, K. Otsuka, J. Nakano, „Fluorescence in lithium neodymium ultraphosphate single crystals”, J. Appl. Phys. 1974, 45, pp. 5096-5097.

[3] M.F. Moktar, N. Kbir-Ariguib, M. Trabelsi, "Etude du systéme $La(P_3O_9)$-$LiPO_3$ donnes cristallographiques sur $LiLa(PO_3)_4$", J. Solid State Chem. 1981, 38, pp. 130-132.

[4] J. Zhu, W.D. Cheng, D.S. Wu, H. Zhang, Y.J. Gong, H.N Tong, D. Zhao, “A series of lithium rare earth polyphosphates $[LiLn(PO_3)_4]$ (Ln = La, Eu, Gd) and their structural, optical, and electronic properties”, Eur. J. Inorg. Chem. 2007, 2, pp. 285-290.

[5] S. Briche, D. Zambon, D. Boyer, G. Chadeyron, R. Mahiou, “Sol–gel derived $Y(PO_3)_3$ polyphosphate: Synthesis and characterization”, Opt. Mater. 2006, 28, pp. 615-620.

[6] A. Matraszek, L. Macalik, I. Szczygieł, P. Godlewska, P. Solarz, J. Hanuza, „ Luminescence and optical absorption studies of submicro-dimensional cerium ortho- and metaphosphates doped with Eu^{3+} ions”, J. Alloys Compd. 2008, 451, pp. 254-257.

[7] R.J. Wiglusz, R. Pazik, A. Lukowiak, W. Strek, “Synthesis, structure and optical properties of $LiEu(PO_3)_4$ nanoparticles”, Inorg. Chem. 2011, 50, pp. 1321-1330.

[8] M. Ferhi , K. Horchani-Naifer, M. Ferid, “Spectroscopy properties of Eu^{3+}-doped $KLa(PO_3)_4$ and $LiLa(PO_3)_4$ powders”, Opt. Mater. 2011, 34, pp. 12-18.

[9] G. Alombert-Goget, C. Armellini, S. Berneschi, A. Chiappini, A. Chiasera, M. Ferrari, S. Guddala, E. Moser, S. Pelli, D.N. Rao, G.C. Righini, “Tb^{3+}/Yb^{3+} co-activated Silica-Hafnia glass ceramic waveguides”, Opt. Mater. 2010, 33, pp 227-230.

[10] S. Jiang, Y. Zhang, “Upconversion nanoparticle-based FRET system for study of siRNA in live cells”, Langmuir 2010, 26, pp. 6689-6694.

[11] T. Soukka, T. Rantanen, K. Kuningas, “Photon upconversion in homogeneous fluorescence-based bioanalytical assays”. In: O.S. Wolfbeis (Ed.), “Fluorescence methods and applications: Spectroscopy, imaging, and probes”, 2008, pp. 188-200.

[12] B. Han, H. Liang, H. Ni, Q. Su, G. Yang, J. Shi, G. Zhang, “Intense red light emission of Eu^{3+}-doped $LiGd(PO_3)_4$ for

mercury-free lamps and plasma display panels application", Opt. Express, 2009, 17, pp. 7138-7144.

[13] Card No. 14-ICDS, Fachinformationszentrum Karlsruhe, 2012

[14] Card No. 420116-ICDS, Fachinformationszentrum Karlsruhe, 2012

[15] Card No. 36052-ICDS, Fachinformationszentrum Karlsruhe, 2012

[16] G. Kaur, S.K. Singh, S.B. Rai, "Eu^{3+} and Yb^{3+} codoped Gd_2O_3 single phase nanophosphor: An enhanced monochromatic red emission through cooperative upconversion and downconversion", J. Appl. Phys. 2010, 107, pp. 073514.

[17] X. Wei, J. Zhao, W. Zhang, Y. Li, M. Yin, "Cooperative energy transfer in Eu^{3+}, Yb^{3+} codoped Y_2O_3 phosphor", J. Rare Earth 2010, 28, pp. 166-170.

[18] W. Strek, P. Deren, A. Bednarkiewicz, "Cooperative processes in $KYb(WO_4)_2$ crystal doped with Eu^{3+} and Tb^{3+} ions, J. Lumin. 2000, 87-89, pp. 999-1001.

[19] M.K. Lau, J.H. Hao, "Near-infrared quantum cutting in Eu^{3+}-Yb^{3+} co-doped YAG through downconversion for silicon solar cell", Energy Procedia 2012, 15, pp.129-134.

[20] Y. Dwivedi, S.N. Thakur, S.B. Rai, "Study of frequency upconversion in Yb^{3+}/Eu^{3+} by cooperative energy transfer in oxyfluoroborate glass matrix", Appl. Phys. B 2007, 89, pp. 45-51.

[21] N. Rakov, D.F. Amaral, R.B. Guimarães, G.S. Maciel, "Spectroscopic analysis of Eu^{3+} and Eu^{3+}:Yb^{3+}-doped yttrium silicate crystalline powders prepared by combustion synthesis", J. Appl. Phys. 2010, 108, 073501.

[22] R. Martin-Rodriguez, R. Valiente, S. Polizzi, M. Bettinelli, A. Speghnini, F. Piccinelli, "Upconversion luminescence in nanocrystals of $Gd_3Ga_5O_{12}$ and $Y_3Al_5O_{12}$ doped with Tb^{3+}−Yb^{3+} and Eu^{3+}−Yb^{3+}", J. Phys. Chem. C 2009, 113, pp. 12195-12200.

[23] L. Wang, X. Xue, H. Chen, D. Zhao, W. Qin, "Unusual radiative transitions of Eu^{3+} ions in Yb/Er/Eu tri-doped $NaYF_4$ nanocrystals under infrared excitation", Chem. Phys. Lett. 2010, 485, pp. 183-186.

[24] Y. Dwivedi, D.K. Rai, S.B. Rai, "Stokes and anti-Stokes luminescence from Eu/Yb:BaB_4O_7 nanocrystals", Opt. Mater. 2010, 32, pp. 913–919.

[25] G.S. Maciel, A. Biswas, P.N. Prasad, "Infrared-to-visible Eu^{3+} energy upconversion due to cooperative energy transfer from an Yb^{3+} ion pair in a sol-gel processed multi-component silica glass", Opt. Commun. 2000, 178, pp. 65-69.

[26] F. Auzel, "Upconversion and anti-Stokes processes with f and d ions in solids", Chem. Rev. 2004, 104, pp. 139-174.

[27] Y.Q. Shenga, J. Liu, L.L. Xu, D. Zhai, Z.G. Zhang, W.W. Cao, "Effect of Eu^{3+} codoping on upconversion luminescence in Y_2O_3:Er^{3+}, Yb^{3+} nanocrystals", Solid State Commun. 2010, 150, pp. 1048-1051.

[28] M. Mattarelli, M. Montagna, A. Chiasera, M. Ferrari, L. Zampedri, G.C. Righini, L.M. Fortes, M.C. Gonçalves, L. F. Santos, R.M. Almeida, "Self-absorption and radiation trapping in Er3+-doped TeO2-based glasses", Europhysics Lett. 2005, 71, pp. 394-399.

[29] A. Chiappini, A. Chiasera, C. Armellini, S. Varas, A. Carpentiero, M. Mazzola, E. Moser, S. Berneschi, G.C. Righini, M. Ferrari, "Sol-gel-derived photonic structures: fabrication, assessment, and application", J. Sol-Gel Sci. Technol. 2011, 60, pp. 408-425.

Numerical Study and Performance Analysis of Carbone Nanotube Field Effect Transistors

Rechem Djamil*, Benkara Salima** and Lamamra Kheireddine**
* Laboratory of active components and materials University Larbi Ben M'hidi Oum El Bouaghi 04000, Algeria.
** Department of science and technology, Faculty of science and technology, University Larbi Ben M'hidi, Oum El Bouaghi 04000, Algeria.
e-mail: rechem_dj@yahoo.fr

Abstract - As transistors are scaled down to nanometers, the theory and structure of nanometers devices such as carbon nanotubes field effect transistors (CNTFET) are being extensively studied. Self consistent solution of the Poisson and Schrödinger equations is performed using the non-equilibrium Green's function (NEGF) formalism to investigate the impact of nanotube diameter, gate oxide thickness and high-k gate dielectric permittivity parameters on the coaxially gate, n-type CNTFET. Our results show that the nanotube diameter and gate oxide thickness influences the ION/IOFF current ratio, the drain induced barrier lowering (DIBL), the subthreshold slop as well as transconductance and drain conductance. Furthermore, in this work we focus on the impact of high-k gate dielectric permittivity on the performance of CNTFETs. Using high-k dielectric is caused by the enhancement in device characteristics. A good agreement with numerical simulation results is obtained.

I. INTRODUCTION

The carbon nanotube field effect transistor (CNTFET) is a promising device to supersede the MOSFET at the end of the technology roadmap of CMOS [1-4]. Since the first reports of single-walled carbon nanotubes (CNTs) in 1993 [5], they have been the subject of intense interest for basic and applied research. In CNTFETs, ballistic or near-ballistic transport phenomena have been observed [7-9], and the existing design infrastructure and fabrication process of CMOS-based MOSFETs can be also used for CNTFETs [10–12].

Rapid progress in the field has recently made it possible to fabricate digital and analogue CNTFET bases circuits, such as logic gate, static memory cells [13-14]. A CNTFET with a short gate width is expected to have marked advantages in terms of operational frequency and will be a subject of future studies.

We present in this paper, a scaling study of carbon nanotube field-effect transistors (CNTFETs). We investigate the scaling issues in device performance focusing on I_{ON}/I_{OFF} current ratio, the drain induced barrier lowering (DIBL), subthreshold slope as well as transconductance and drain conductance with different carbon nanotube (CNT) diameters and gate oxide thicknesses.

Silicon dioxide (SiO_2) has facing the scaling limitation due to direct tunneling current that limits low power application and reliability problem. Recently, many high-k materials such as HfO_2 (k=16), ZrO_2 (k=25), Ta_2O_5 (k=50) and TiO_2 (k≈35-100) [15-19] have been introduced as alternative gate dielectrics to overcome leakage current problem. To consider the benefits of both high-k dielectrics and CNTFETs, in this paper we present also the effects of high-k gate dielectric on the performances of CNTFETs. We have simulated the structure using a two-dimensional (2-D) quantum simulation.

II. DEVICE STRUCTURE

The modeled device, a coaxially gate, n-type CNTFET is schematically shown in Fig. 1(a) and 1(b). The nanotube length is 50 nm, consisting of ~1.2×10^4 carbon atoms. The intrinsic channel length is 20 nm, and the doped source/drain length is 15 nm.

To simulate the comportment of a CNTFET, the following model is used at different nanotube diameters, gate oxide thicknesses and high-k gate dielectric constants.

The chiralities of the CNTs used are (13,0), (16,0), (19,0), (23,0), (25,0). The gate oxide thicknesses (tox) used are 1.5 nm, 3 nm, 4.5 nm, 6 nm and 7 nm. In addition the gate dielectric constants are 3.9, 16, 25, 50, and 80. These dielectric constants correspond to the dielectric constants reported for SiO_2, HfO_2, ZrO_2, Ta_2O_5, and TiO_2, respectively.

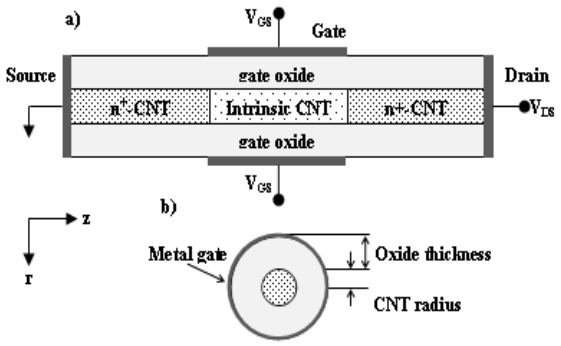


Figure 1. Schematic diagrams of the modeled, coaxially CNTFET

III. MODELING AND SIMULATION

In order to simulate the proposed devices [20], self consistent solution of the Poisson and Schrödinger equations has been performed within the NEGF formalism with a mode space approach, developed in [21-22], Poisson equation is written as:

$$\nabla^2 V(r,z) = \frac{-q}{\varepsilon}[p - n + N_D^+ + N_A^- + n_T] \quad (1)$$

The electron and hole concentrations are computed by solving the Schrödinger equation.The electron density is computed from the density of states (DOS), derived by the NEGF formalism. It can be calculated as follows:

$$n = \int_{Ei}^{+\infty} [D_S f(E - E_{FS}) + D_D f(E - E_{FD})] dE \quad (2)$$

Once, self consistency is achieved, the source-drain current can be expressed as:

$$I = \frac{4q}{h} \int T(E)[f(E - E_{FS}) - f(E - E_{FD})] dE \quad (3)$$

IV. SIMULATION RESULTS

A. *Effect of nanotube diametre*

The nanotube diameter is varied from 1 nm, corresponding to the chirality vector of (13,0), to about 2 nm, corresponding to chirality vector of (25,0). While the gate oxide thickness is fixed at 1.5 nm and high-k gate dielectric is fixed at k=16. The I_{ON}/I_{OFF} current ratio of the CNTFET for different chirality vectors, i.e. different diameters of the zigzag CNT, is depicted in Fig. 2. I_{ON} is obtained at V_{GS} = 1 V and V_{DS} = 1 V, I_{OFF} is determined as the current obtained for V_{GS} = 0 V and V_{DS} = 1 V. As shown in Fig. 2, I_{ON}/I_{OFF} current ratio is improved with increase in the nanotube diameter. This comes from the correlation of the band-structure with the CNT diameter. Using a larger diameter reduces the bandgap, therefore both the ON-current and the leakage current I_{OFF} increase, and I_{ON} increases rapidly.

Thus a significant increase of the I_{ON}/I_{OFF} current ratio is observed. Moreover, the I_{ON}/I_{OFF} current ratio is about 10^6 at a larger diameter, indicating that the CNTFETs have a good current switch capability. So this point must be carefully taken into consideration to achieve the suitable electrical characteristics.

For short channel devices, application of a high drain-to-source bias can shorten the threshold voltage and increase the off-currents. This is known as drain induced barrier lowering (DIBL). In CNTFETs, the DIBL effect is still a primitive problem and open for further study [23]. Fig. 3 shows the effect of varying the nanotube diameter on the DIBL. The DIBL is accessed using the classical formula:

$$DIBL = \frac{V_{TH}(\text{low } V_{DS}) - V_{TH}(\text{high } V_{DS})}{\text{high } V_{DS} - \text{low } V_{DS}} \quad (4)$$

From the simulation results, it can be drawn that the DIBL is considerably improved with decreasing the nanotube diameter; therefore, the control of gate on the channel becomes stronger. One notes a reduction around 53% of DIBL when CNT chiralities varying from (13, 0) CNT to (25, 0) CNT.

Another important parameter is the subthreshold slope S that measures the how efficiently the gate controls conduction though the channel. It is defined as $S = \Delta V_{GS}/\Delta \log 10\ (I_{DS})$. In a transistor with ohmic source and drain contacts (as in a conventional Si MOSFET), S is limited by thermionic emission over the channel and is ~(kB T/ q)ln10. Thus, at room temperature, its limiting value is 60 mV/dec. A small subthreshold slope is desired for low threshold voltage and low-power operation for FETs scaled down to small size. Fig. 3 shows the simulated subthreshold slope as a function of the nanotube diameter. It can be observed that when the nanotube diameter decreases, S decreases slightly (practically remains constant around 67 mV/decade).

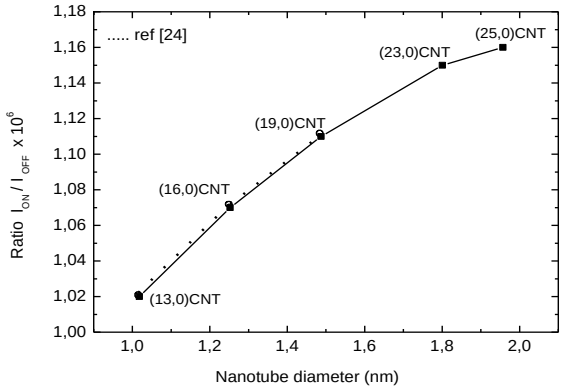


Figure 2. I_{ON}/I_{OFF} current ratio as a function of the nanotube diameter

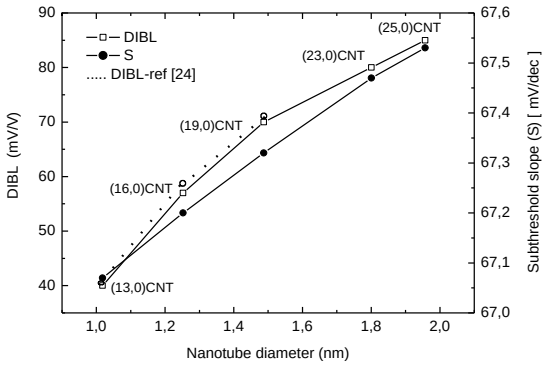


Figure 3. Drain induced barrier lowering (DIBL) and subthreshold slop versus nanotube diameter

The transconductance, an important device parameter, is defined as measure of device gain and is directly related to the circuit speed. The transconductance gm curve is obtained by differentiating the drain current I_{DS} with respect to the gate voltage V_{GS} at a given drain bias. $g_m = \frac{\partial I_{DS}}{\partial V_{GS}}$

As can be seen from Fig. 4, when the nanotube diameter increases, the transconductance gm increases. The drain conductance defined by $g_d = \frac{\partial I_{DS}}{\partial V_{DS}}$ is presented also in Fig. 4 for different values of nanotube diameter. It is observed that gd is higher for larger nanotube diameter. One notes a voltage gain gm/gd around ~25 whatever the value of nanotube diameter is. To check the accuracy of our numerical calculations, our results were compared with simulated results presented in ref [24] (dotted lines) .good agreement has been found in figures 2 and 3.

B. Effect of gate oxide thikcness

In this section, we investigate the effect of different gate oxide thickness (tox) on the CNTFET performance. Fig. 5 represents the evolution of I_{ON}/I_{OFF} current ratio with gate oxide thickness varying from 1.5 nm to 7 nm.

As shown in the figure, a larger gate oxide thickness lowers I_{ON}/I_{OFF}. This leads to reduce the control of the gate over the channel region.

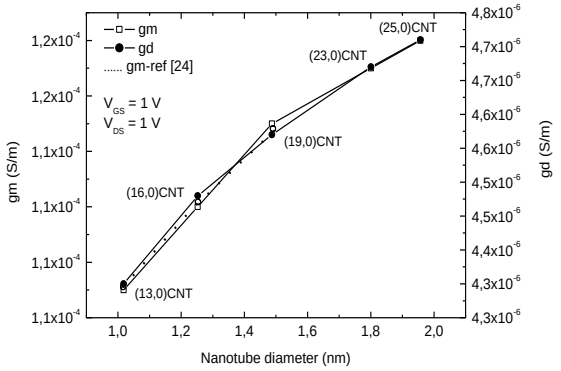


Figure 4. Variation of gm and gd as a function of nanotube diameter

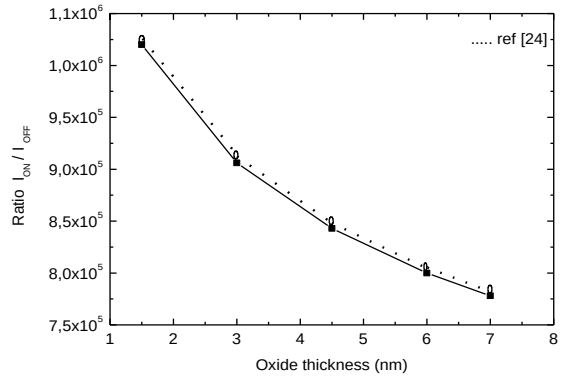


Figure 5. I_{ON} /I_{OFF} current ratio as a function of the oxide thickness.

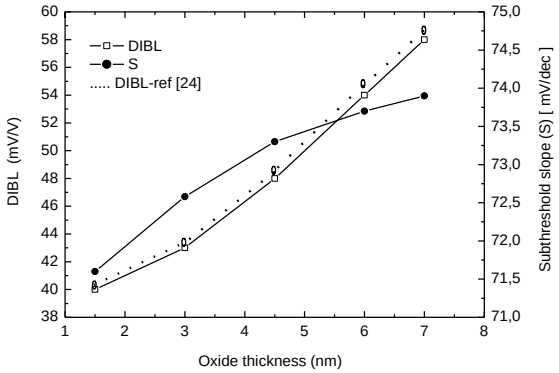


Figure 6. Drain induced barrier lowering (DIBL) and subthreshold slop versus oxide thickness

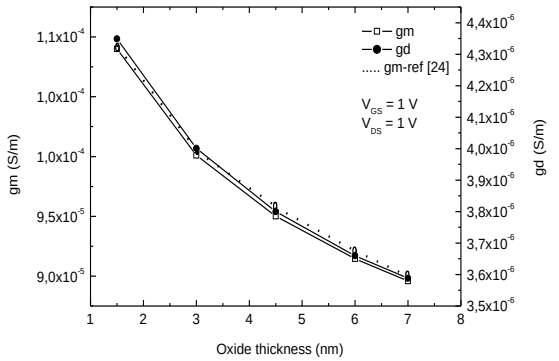


Figure 7. Variation of gm and gd as a function of oxide thickness.

The DIBL is calculated as a function of gate oxide thickness. This is illustrated in Fig. 6, an important reduction in DIBL around 30 % can be observed when tox changes from 7 nm down to 1.5 nm.

The effect of gate oxide thickness on transconductance gm and drain conductance, gd, is shown in Fig. 7. A larger gate oxide thickness leads to a smaller gd and gm. We can conclude that the variation of the gate oxide thickness has a larger effect because it determines the controllability of the gate over the channel region. Our results were compared with simulated results presented in ref [24] (dotted lines) .good agreement has been found in figures 4-7.

C. Effect of high –k gate dielectric permettivity

In this section, to highlight our analysis, we varying the gate dielectric permittivity from 3,9 to 80. Fig. 8 shows the evolution of I_{ON}/I_{OFF} current ratio as function of the gate dielectric permittivity (k). It can be seen from the figure that with increasing k the I_{ON}/I_{OFF} current ratio increases ($\geq 10^4$ is typically desirable in logic applications) and leads to efficient gate switching.

The DIBL and subthreshold slope (S) are calculated as a function of gate dielectric permittivity (k). These are illustrated in Fig. 9. As seen in this figure, the DIBL and subthreshold slope improve with increasing k and the control of gate on the channel becomes stronger. One notes a considerable reduction of DIBL from 180 mV/V to

about 20 mV/V when gate dielectric permittivity increases.

As shown in Fig. 10, it is seen that as gate dielectric permittivity increases the transconductance gm and the drain conductance gd continues to increase. Consequently, voltage gain is higher for high-k gate dielectric. Noting that when the gate dielectric permittivity increases the CNTFET characteristics is not affected by the fringing-induced barrier lowering (FIBL). This is due to the coaxially CNTFET structure. Our results were compared with simulated results presented in ref [25] (dotted lines) .good agreement has been found in figures 8.

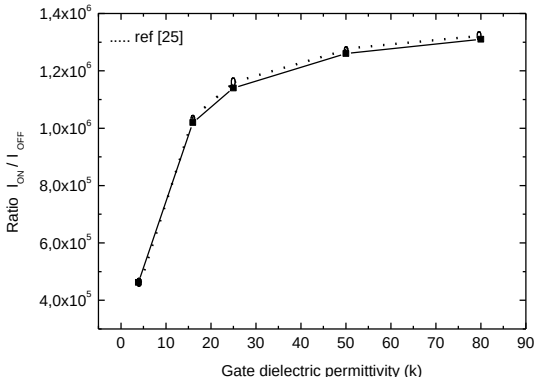


Figure 8. I_{ON} /I_{OFF} current ratio as a function of the gate dielectric permittivity.

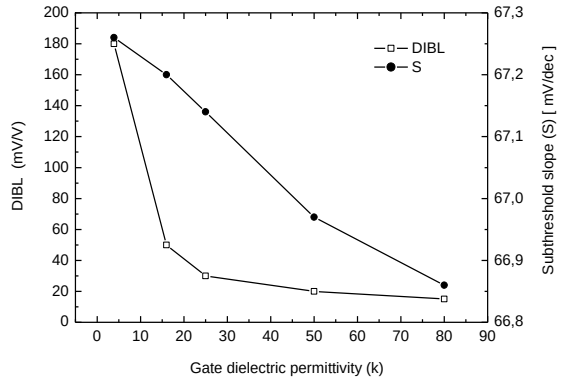


Figure 9. Drain induced barrier lowering (DIBL) and subthreshold slop versus gate dielectric permittivity.

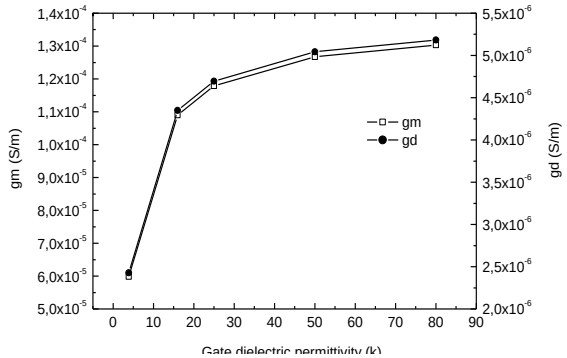


Figure 10. Variation of gm and gd as a function of the gate dielectric permittivity.

V. CONCLUSION

Based on the I_{ON}/I_{OFF} current ratio, drain induced barrier lowering (DIBL), subthreshold slope, transconductance and drain conductance variation the CNTFET behavior is evaluated. We concluded that using large CNT diameter and thinner gate oxide are caused by the enhancement in on-state current, transconductance and drain conductance. In addition, off-state current, DIBL and subthreshold slope improve in CNTFETs with thinner gate oxide, but they become worse in CNTFETs with large nanotube diameter.

Furthermore it is found that using high-k dielectric is caused by I_{ON}/I_{OFF} current ratio increases and leads to high on-state current. Also DIBL is significantly improved in high-k gate dielectric, on the other hand we found that the influence of high-k gate electric on subthreshold slope is slight. It is also found that using high-k gate dielectric in CNTFET is caused by the enhancement in transconductance and drain conductance, consequently, voltage gain is higher.

REFERENCES

[1] International Technology Roadmap for Semiconductors, http://public.itrs.net/.

[2] S. Tans, A. Verschueren, and C. Dekker, Nature., vol. 393, pp 49, 1998.

[3] H. J. Dai, A. Javey, E. Pop, D. Mann, and Y. Lu, NANO, vol. 1, pp. 1, 2006.

[4] Ph. Avouris, Phys. World, vol. 20, pp. 40, 2007.

[5] S. Iijima, and T. Ichihashi, "Single-shell carbon nanotubes of 1-nm diameter," Nature, vol. 363,pp. 603-605, 1993.

[6] S. J. Wind, J. Appenzeller, and P. Avouris, "Lateral scaling in carbon nanotube field-effect transistors," Phys. Rev. Lett, vol. 91, pp. 058–301, 2003.

[7] Y. Park, S. Rosenblatt, Y. Yaish, V. Sazonova, H. Ustunel, S. Brag, T. A. Arias, P.W. Brouwer, and P. L. McEuen, "Electron–phonon scattering in metallic single-walled carbon nanotubes," Nano Lett, vol. 4, pp. 517–520, 2004.

[8] W. Liang, M. Bockrath, D. Bozovic, J. H. Hafner, M. Tinkham, H. Park, and F. Perot, "Interference in a nanotube electron waveguide," Nature, vol. 411, pp. 665–669, 2001.

[9] J. Guo, A. Javey, H. Dai, and M. Lundstrom, "Performance analysis and design optimization of near ballistic carbon nanotube field-effect transistors," IEEE Trans Electron Dev, pp. 703–706, 2004.

[10] H. Ohnaka, Y. Kojima, S. Kishimoto, Y. Ohno, and T. Mizutani, "Fabrication of carbon nanotube field effect transistors using plasma-enhanced chemical vapor deposition grown nanotubes," Jpn. J. Appl Phys, vol. 45, pp. 5485–5489, 2006.

[11] J. Li, Q. Zhang, Y. Yan, S. Li, and L. Chen, "Fabrication of carbon nanotube field-effect transistors by fluidic alignment technique," IEEE Trans Nanotechnol, vol. 6, pp. 481–484, 2007.

[12] V. Derycke, R. Martel, J. Appenzeller, and Ph Avouris, "Carbon nanotube inter-and intramolecular logic gates," Nano. Lett, vol. 9, pp. 453–456, 2001.

[13] A. Bachtold, P. Hadley, T. Nakanishi, and C. Dekker, "Logic circuits with carbon nanotube transistors," Science, vol. 294, pp. 1317–1320, 2001.

[14] J. L. Autran, R. Device, C. Chaneliere, and B. Balland, "Fabrication and characterization of Si-MOSFETs with PECVD amorphous Ta_2O_5 gate insulator," IEEE Electron Device Lett., vol. 18, pp. 447-449, Sept 1997.

[15] J. H. Lee, K. Koh, N. I. Lee, M. H. Cho, Y. K. Ki, J. S. Jeon, K. H. Cho, H. S. Shin, M. H. Kim, K. Fujihara, H. K. Kang, and J. T. Moon, "Effect of polysilicon gate on the flatband voltage shift and mobility degradation for ALD-AL2O3 gate dielectric," in IEDM Tech. Dig., pp. 645-648, 2000.

[16] R. Choi, K. Onishi, S. K. Chang, H. J. Cho, Y. H. Kim, S. Krishnan, M. S. Akbar, and J. C. Lee, "Effects of deuterium anneal on MOSFETs with HfO2 gate dielectrics," IEEE Electron Device Lett, vol. 24, pp. 144-146, Feb 2003.

[17] W. J. Qi, R. Nieh, B. H. Lee, K. Onishi, and L. Kang, "Performance of MOSFETs with ultrathin ZrO_2 and Zr silicate gate dielectrics," in IEEE Symp. VLSI Tech. Dig, pp. 15-16, 2000.

[18] G. D. Wilk, R. M. Wallace, and J. M. Anthony, "High-k gate dielectrics: Current status and materials properties considerations," J. Appl. Phys, vol. 89, No. 10, pp. 5243–5275, May 2001.

[19] J. Guo, S. Datta, MS. Lundstrom, and MP. Anantram, "Towards multiscale modeling of carbon nanotube transistors," Int J Mult Comp Eng,vol. 2, pp. 257–276, 2004.

[20] Liu Xinghui et al., " Study on transport characteristics of CNTFET based on NEGF theory," 2011 IEEE International Conference of Electron Devices and SolidState Circuits (EDSSC). 17-18 Nov. 2011

[21] J. Guo, S. Datta, MS. Lundstrom, and MP. Anantram, "Towards multiscale modeling of carbon nanotube transistors," Int J Mult Comp Eng,vol. 2, pp. 257–276, 2004.

[22] S. O. Koswatta, M. S. Lundstrom, M. P. Anantram, and D. E. Nikonov, "Simulation of phonon-assisted band-to-band tunneling in carbon nanotube field-effect transistors," Appl. Phys. Lett., 87, 253107, 2005.

[23] Z. Hailiang, S. Bingcai, C. Yaqing, Z. Minxuang, and H. Yue , "Numerical study of DIBL effect in Carbon Nanotube-FETs, Nanotechnology," IEEE. NANO 2009, 9th IEEE Conference on Publication, pp. 752–755, 2009.

[24] Z. Arefinia, A.A. Orouji, "Novel attributes in scaling issues of carbon nanotube field-effect transistors," microelectronics journal 40 (2009) 5-9.

[25] Z. Arefinia, A.A. Orouji, "Investigation of the novel attributes of a carbon nanotube FET with high-k gate dielectrics," Physica E: Low-dimensional Syst. Nanostructures 40 (2008) 3068–3071.

New Silicon Reach-Through Avalanche Photodiodes with Enhanced Sensitivity in the DUV/UV Wavelength Range

D. Grubišić, A. Shah
Laser Components DG, Inc., Tempe, Arizona, USA
dgrubisic@laser-components.com

Abstract - Many applications, particularly in the medical and bio-medical fields, require highly sensitive detectors in the short wavelength range (blue and below); so far there have been few commercially available avalanche photodetectors with adequate performance to satisfy demands of such applications. We have developed a new silicon reach-through avalanche photodiode (RAPD) operating in the blue wavelength range with improved sensitivity and noise performance superior to any similar detectors available on the market today. This has been accomplished thanks to an innovative, ultra-thin and highly doped surface top layer grown using a new epitaxial technique developed at Delft University of Technology in The Netherlands. An important additional advantage of this new RAPD is its unmatched noise performance over the widest commercially available wavelength range from 250 nm to 1100 nm. Short wavelength response is limited only by the transmission of the package window optimized for wavelengths starting at 250 nm and up, but devices can operate at wavelength even below 100 nm if assembled in a vacuum.

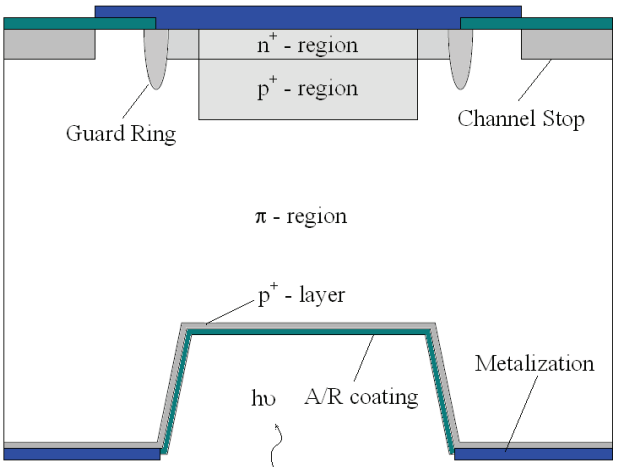
I. INTRODUCTION

Silicon reach-through avalanche photodiodes (RAPD) are types of avalanche photodiodes (APD) made on high quality, float zone, 5000 ohm-cm silicon material as opposed to regular APDs made on epitaxial silicon wafers and as such have inferior performance to the RAPD. Silicon RAPD is highly sensitive photo detector used in low background applications requiring low noise and high bandwidth sensors. Operating in the avalanche mode, the RAPD devices generate a cascade of moving carrier pairs for each absorbed photon, creating measurable signal current even when extremely weak photon fluxes are present [1]. The avalanche multiplication process is based on carrier multiplication in a strong electric field region. Carriers acquire sufficient energy between collisions to eject valence electrons from the lattice, and in that way create multiple electron-hole pairs. The efficiency of new carriers' generation is governed by ionization coefficients of both carriers and is a strong function of the electric field and temperature. Theoretical considerations of the process are complicated, but extensive studies done over the years [2] have provided comprehensive insight into how the devices operate and allow for selection of the optimum configuration needed for each different application. In this paper we will describe how an already wide sensitivity range of silicon RAPD devices is being extended deep into the shorter wavelengths i.e. blue to deep ultra violet (DUV). This was accomplished by applying a novel thin film growth technique. The thin and highly doped layer produced with this deposition technique minimizes short wavelengths' surface absorption loses thus significantly improving device sensitivity range for wavelengths less than 400 nm. This improvement has been sought for a long time, but latest advances in the bio-medical field have made the need for such high performance sensors even more urgent.

II. GENERAL DESCRIPTION

A. Silicon RAPD Structure

The silicon RAPD structure described in this paper is a $p^+ - \pi - p - n^+$ type [3]. In addition to highly doped contact layers on both sides of the structure, it consists of a narrow high-field region where the multiplication takes place and a much wider low-field region in which the incoming photons are absorbed. The design, impurity concentration, and electric field profiles are shown in Fig. 1.



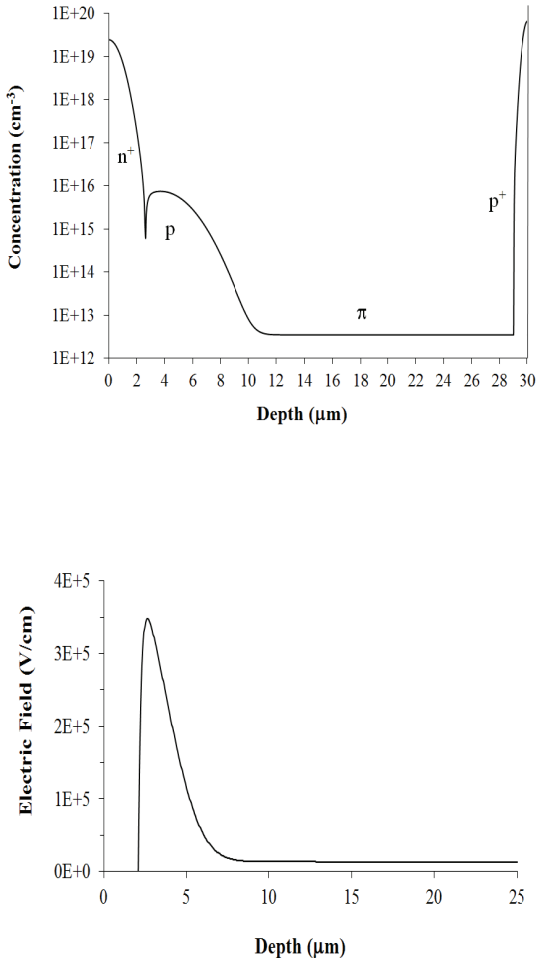


Figure 1. Silicon RAPD structure, impurity concentration and electric field profiles

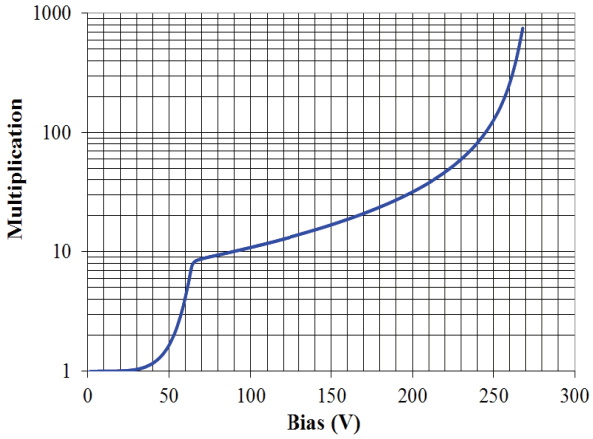


Figure 2. Typical multiplication gain characteristics measured on silicon RAPD

B. Multiplication Gain

Impurity profiles, p-type first and n-type later, are adjusted so that when the depletion edge, driven by an external bias, reaches the high resistivity π region, the peak electric field at the internal p-n junction is set to be just slightly below that required to initiate the avalanche breakdown. The depletion layer spreads rapidly out to the p+ contact when the external bias is incrementally increased beyond the point described in the previous sentence. However, the electric field throughout the device increases only slowly due to significant and fast increase in the depletion layer width. This results in a gradual shape of the multiplication characteristic shown in Fig. 2. To have the device operating in the linear mode the operating voltage range is limited by the requirement for avoidance of the avalanche breakdown. At the point of the avalanche breakdown the multiplication gain is infinite, resulting in an improper operation of the device. The proper linear mode operation of the device requires large, but finite multiplication gain.

The operating voltages of the typical devices are in the 200 V to 400 V range, because carrier multiplication and photon absorption regions are geometrically separated as shown in Fig. 1. An impractical voltage well in excess of 1000 V would have to be used if two regions were combined into a single region. The separation of the multiplication and absorption regions results in a complex electric filed profile with the device noise performance depending significantly on the structure and impurity profiles.

Both signal and noise considerations demand that the multiplication process is initiated by carriers having a higher ionization coefficient. Electrons are higher ionization coefficient carriers in silicon and for that reason RAPDs are designed so that the light enters at the p^+ region surface. Under normal operating conditions both π and p regions are fully depleted, allowing for efficient collection of photo-generated electron-hole pairs. Photons have been absorbed in the π region and then separated by that region's electric field, which is approximately $3x10^4$ V/cm. The electrons are subsequently swept into the p region where the electric field is sufficiently high to cause multiplication by impact ionization. Figure 3. shows a positional dependence of the electrons multiplication gain in an RAPD structure. The characteristic is a result of calculations using a one dimensional model implemented in a spreadsheet macro program [4]. The generated holes in the π region are on the other hand swept toward the p^+ layer. They do not experience any gain and constitute a very small component of the overall hole current. The hole current consists mostly of the holes created by impact ionization in the multiplication region. The response time of silicon RAPD is determined primarily by the hole transit time and ranges from 400 psec to 5 nsec, depending on the width of the absorption layer.

The situation when light is entering on the n^+ layer side is possible as well, such as in the case of epitaxial

APDs, but those devices are noisier and are used in less demanding applications.

C. Noise

Theoretical noise studies demonstrated that the noise component associated with the avalanche mechanism in the silicon RAPD devices is lower if only electrons are being multiplied. The Excess Noise Factor, F, quantifies the above effect as described below.

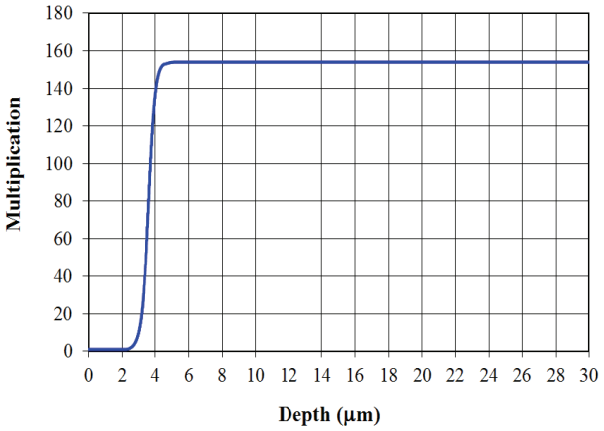


Figure 3. Positional dependence of electrons multiplication gain in the RAPD

The noise density, i_n^2, in an RAPD is given by the following equation [5]:

$$i_n^2 = 2q(I_s + (I_{dm}+I_{bk})M^2F) \quad (1)$$

- I_s - Surface component of the dark current, which does not get multiplied,
- I_{dm} - Bulk component of the dark current undergoing multiplication,
- I_{bk} - Background generated current,
- M – Multiplication gain,
- F – Excess Noise Factor.

The Excess Noise Factor, F, is given by [3]:

$$F = k_{eff}M + (1 - k_{eff})(2 - 1/M) \quad (2)$$

- k_{eff} – the effective ratio of the hole to electron ionization coefficients.

The noise density is reduced by lowering the value of the Excess Noise Factor as shown by (1). k_{eff} is the primary quantity affecting the value of the Excess Noise Factor. Its value is always less than one, because the hole ionization coefficient is always smaller than the electron ionization coefficient in silicon. For multiplication gains higher than 10 and k_{eff} less than 0.1, (2) can be simplified to $F = 2 + k_{eff}M$. Consequently, low k_{eff} is highly desirable in order to achieve a low noise RAPD operation. In practical terms it means that RAPDs with lower value of k_{eff} will be able to operate at higher gains with a lower noise level.

D. Quantum Efficiency

Light absorption in the silicon RAPD is governed by the silicon absorption coefficient shown in Fig. 4 [6]. Quantum efficiency is a fraction of the incident photon flux, which generates electron-hole pairs in the absorption region and contributes to the signal current

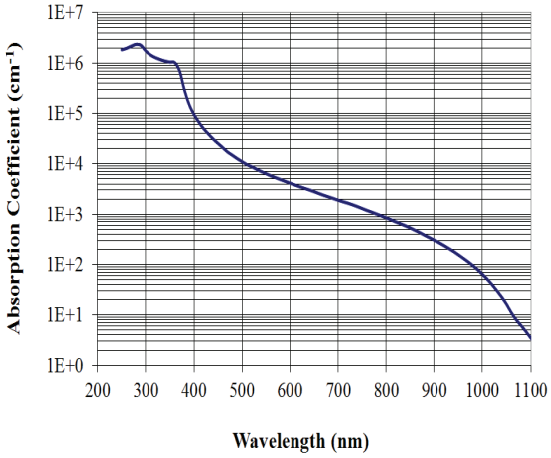


Figure 4. Absorption coefficient versus wavelength for silicon

Quantum efficiency, η is given by [4]:

$$\eta = \frac{(1-r_1)(1-\exp(-\alpha w))(1+r_2\exp(-\alpha w))\exp(-\alpha d)}{1-r_1r_2\exp(-2\alpha(w+d))} \quad (3)$$

where r_1 is the reflectivity of the front surface; r_2 is the reflectivity of the back; d is the effective thickness of the light entry "dead" layer; w is the effective thickness of the absorption layer; and α is the absorption coefficient of silicon in the wavelength range of interest. The light entry "dead" layer is the p^+ layer. The carriers generated in the layer do not participate in the multiplication process and thus do not contribute to the operation of the device. In order to maximize η, r_1 should be made as small as possible by using an anti-reflecting coating, αw should be greater than unity, αd should be much less than unity, and r_2 should be as close to one as possible.

For wavelengths below 800 nm (3) takes form:

$$\eta = (1-r_1)\exp(-\alpha d) \quad (4)$$

Equation (4) shows that the contact layer on the light entrance side is critical and has to be as thin as practically possible if an increase in the sensitivity at short wavelengths is desired.

III. SPECTRAL COVERAGE IMPROVEMENTS

Based on the RAPD quantum efficiency discussion it is clear that thickness of the "dead" layer on the side where light enters the structure should be minimized to improve the short wavelength response of the device, but it has been difficult to accomplish such a thin and highly doped layer using standard processes common in today's semiconductor wafer fabrication. The silicon absorption depth in the wavelength range of interest is shown in Fig. 5 where the shortest '1/e' penetration depth is around 5 nm, implying that if the "dead" layer in our RAPD were 5 nm thick we would expect a 40% quantum efficiency at 250 nm. The absorption depth at the longer wavelengths is much longer making the RAPD operation far less sensitive to the thickness of the "dead" layer at those long wavelengths.

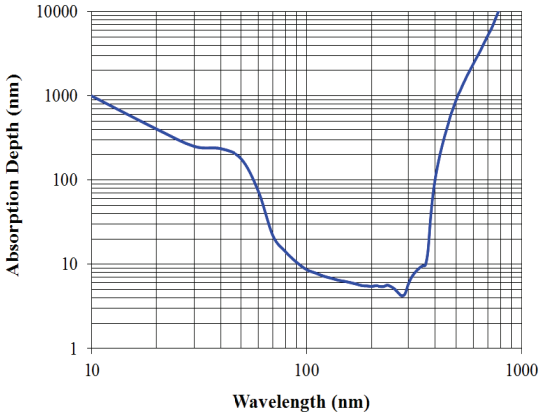


Figure 5. Penetration depth versus incident radiation wavelength in Si [7]

Since the p^+ layer is the contact as well as the surface passivation layer, it has to be as highly doped as possible to prevent the internal electric field from reaching the surface of device and minimizing the series resistance of the layer. The problem is that the p^+ layer is also the top layer on the surface where light enters the device and therefore it has to be as thin as possible. Those two opposite requirements have been difficult to reconcile up until now, because the high dopant solubility in silicon needed to achieve high surface concentrations is possible only at high temperatures, making sufficiently shallow layers created by diffusion, implantation or standard high temperature epitaxy impossible to produce. Fortunately, a novel epitaxial technique developed by the Silicon Device Integration Group at Delft University of Technology [8] provides the means of growing such thin and highly doped layers at lower temperatures, between 500 and 700 ^oC, while still achieving surface doping levels exceeding 10^{20} cm^{-3}. With the help from the Silicon Device Integration Group, their innovative fabrication technique has been successfully incorporated in our RAPD processing sequence, yielding the devices described in this paper.

The deposition of the highly doped and thin p^+ layer was performed in an ASM Epsilon One reactor using diborane B_2H_6 and hydrogen H_2 as the gas source and carrier gas, respectively. The target layer thickness has been set to 3 nm. The layer deposition rate has been calibrated using variable angle ellipsometry measurements and confirmed by a Transmission Electron Microscope (TEM). An example of such a layer is shown in Fig. 6.

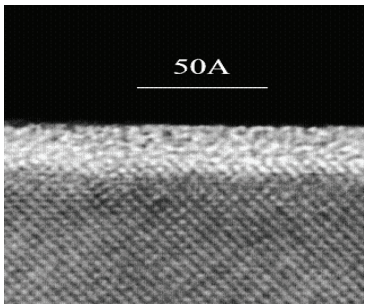


Figure 6. Thin highly doped epitaxial layer grown on silicon

Once wafer processing has been completed, the devices were mounted onto TO-18 header, example shown in Fig. 7, and closed in a hermetically sealed package with the cap window made from the fused quartz with a special antireflective coating optimized for the wavelength range from 255 to 700 nm.

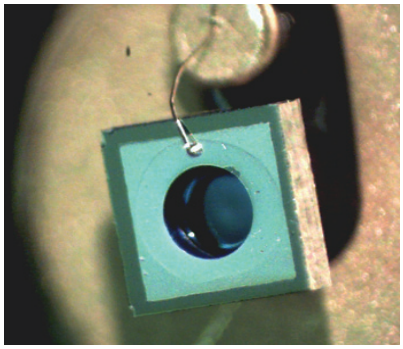


Figure 7. RAPD mounted on TO-18 header (SUR500S8)

SUR500S8, which is Laser Components DG, Inc. (LCDG) product designation for this new silicon RAPD, has current-voltage (I-V), noise and spectral characteristics measured using setup depicted in Fig. 8. SUR500S8 will be used from now on as the designation for the new silicon RAPD device.

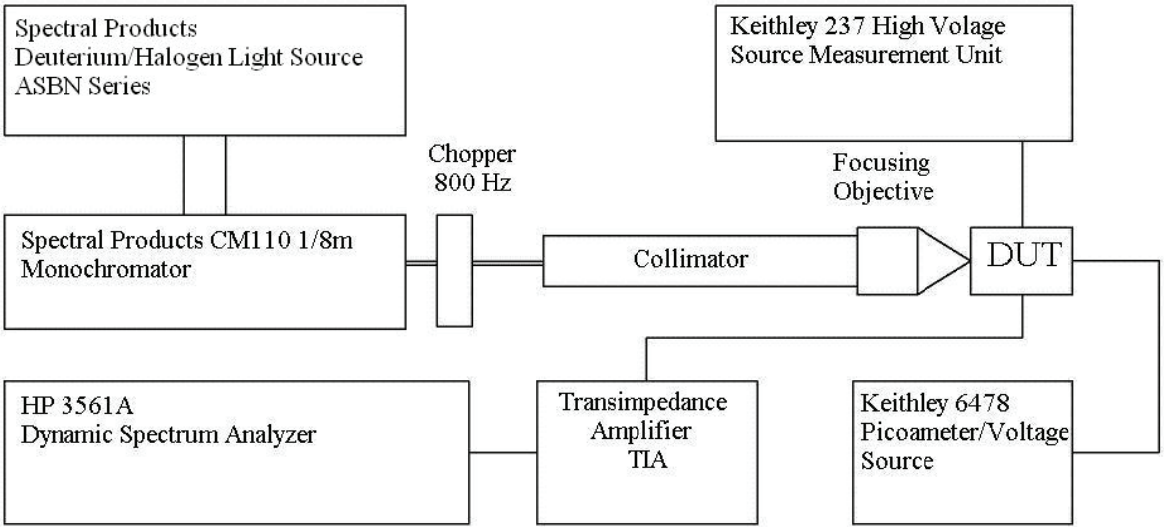


Figure 8. Measurement setup

The light delivery part of the measurement setup consists of the Spectral Products ASBN Series deuterium/halogen light source combined with the Spectral Products CM110 1/8m monochromator and a grating optimized for the wavelengths range of interest. It delivers the light to a device under test (DUT) using a 600 μm core quartz fiber and special optics. The upper part of the setup including the Keithley 237 High Voltage SMU is used in DC measurements of the current flowing inside the device under dark and illuminated conditions. During those measurements the chopper is disabled and it is not obstructing the light path. Typical I-V characteristics obtained in such measurements are shown in Fig. 9, where the I-V characteristic measured under illumination shows the typical reach-through step seen also in the multiplication characteristic. The gradual increase of the current with the bias for the I-V curve measured under illuminations closely corresponds to the gradual increase of the multiplication gain shown in Fig. 2.

Spectral and noise characteristics of the device are measured using the instruments depicted in the lower section of the setup. The light output from the monochromator is chopped at 800 Hz and SUR500S8 is biased by the Keithley 6487 Picoameter/Voltage Source to operate at a multiplication gain close to 100. The signal from the SUR500S8 was conditioned by the transimpedance amplifier and monitored by the HP 3561A Dynamic Spectrum Analyzer. The short wavelength region i.e. up to 400 nm of the spectral characteristic is measured using the deuterium light source and the monochromator with the grating optimized for that wavelength range. The remainder of the characteristic, i.e. from 400 nm up, utilized the halogen light source and an appropriate grating

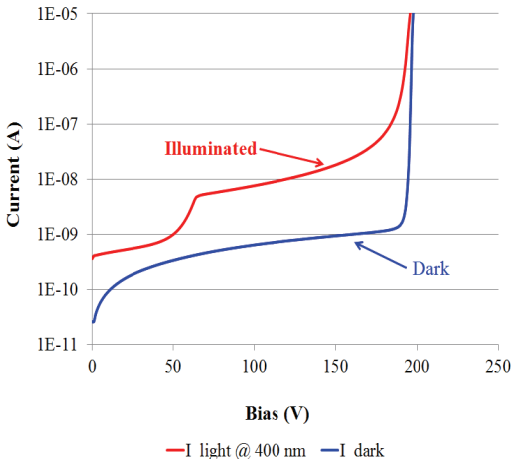


Figure 9. SUR500S8 I-V characteristics

optimized for that wavelength range. A calibrated OL 730-5EC UV-enhanced silicon reference detector from Gooch & Housego was used to measure optical power from the monochromator.

Measurement results are shown in Fig. 10 where the new SUR500S8 device is compared with the LCDG standard RAPD device, SAR500S3. It has to be

emphasized that the SAR500S3 device was assembled using the same fused quartz cap as the SUR500S8 in order to keep the comparison meaningful.

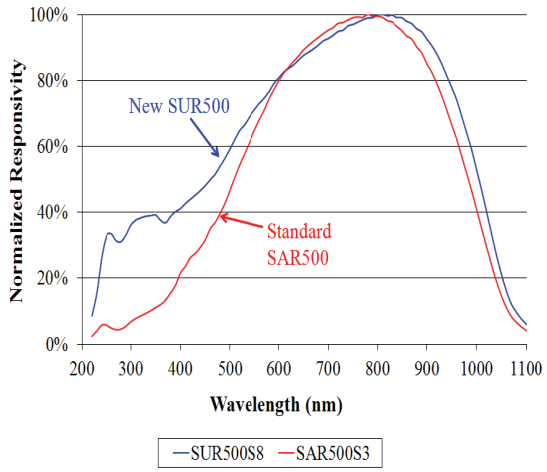


Figure 10. Comparison of spectral characteristics of standard and new RAPD biased at M = 120

Significant improvements in the responsivity [9] at the shorter wavelengths i.e. below 400 nm are clearly visible in Fig. 10, but are even more evident when changes in quantum efficiencies of both devices are compared in Fig. 11. A SUR500S8 responsivity characteristic is also shown in Fig. 11 to demonstrate the responsivity level attainable by such devices for very short wavelengths. This is an important consideration because in bio-medical applications the signal levels are so low that it is practically impossible to detect a signal unless the device responsivity at the wavelengths is high. Another important consideration is minimizing the noise density at the bias levels of interest. Measurements of the responsivity at 400 nm and the noise density as a function of the bias performed on the new silicon RAPD are shown in Fig. 12. The noise density is bias independent in a wide range and up to the responsivity of about 50 A/W, indicating that the low noise device operation is possible while at the same time achieving high responsivity levels. The optimum bias voltage is the voltage at which noise starts to increase more rapidly than the signal itself.

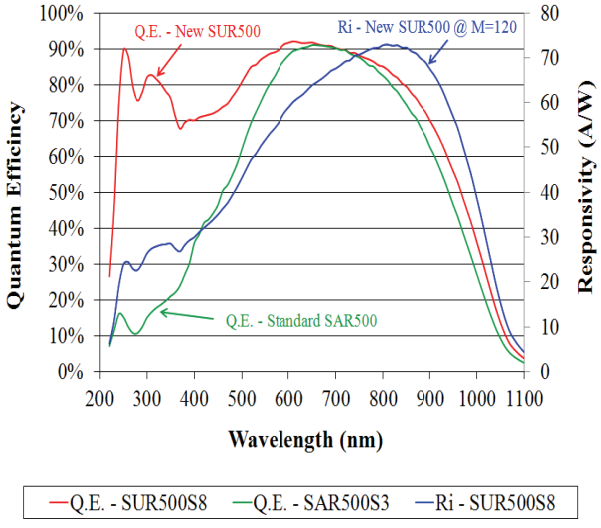


Figure 11. SUR500S8 quantum efficiency improvements

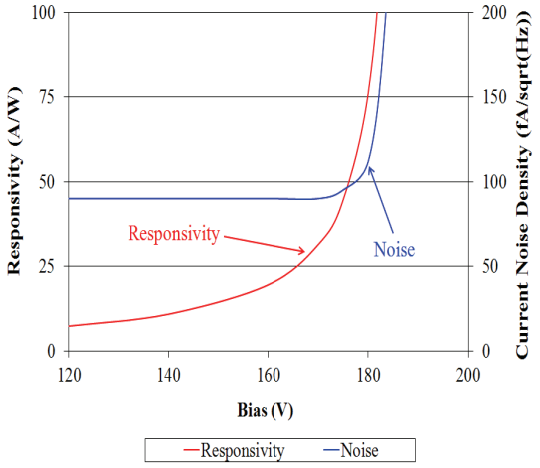


Figure 12. SUR500S8 responsivity at 400 nm and dark noise characteristics

However, it is difficult to see from Fig. 12 what that optimum operating bias would be because responsivity and noise density are changing very rapidly. The Noise Equivalent Power (NEP) [10] graph of the device as a function of bias shown in Fig. 13 is a better representation of the device performance and it clearly shows the optimum bias of the device.

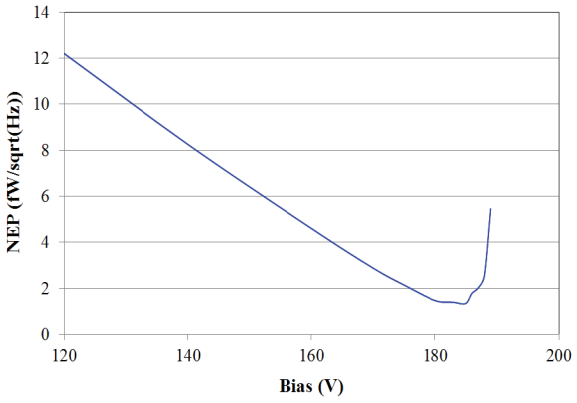


Figure 13. SUR500S8 Noise Equivalent Power

Based on the above noise measurements, it appears that the SUR500S8 can be operated at the responsivity higher than 100 A/W at 400 nm. However, measurements described in this paper were done having no background light, while in actual applications there always will be some background, which will increase the level of noise in the device and shift the NEP characteristic up and to the left. The NEP characteristic shift lowers the optimum responsivity of the device over the complete wavelengths range. Since different applications have different background light levels it is hard to predict what would be the optimum operating point for each individual case. For that reason, each receiver is optimized in such a way that the SUR500S8 bias is adjusted under operating conditions for each application so that its noise is equal to the noise of the preamplifier.

IV. CONCLUSION

Using an innovative semiconductor processing technique, a new class of silicon RAPD devices has been produced that extends its usable operating range deep into the UV wavelengths in order to meet increasing demands for such devices in bio-medical applications. This has been accomplished thanks to an innovative, ultra thin and highly doped surface top layer, grown using a new epitaxial technique developed at Delft University of Technology in The Netherlands. In principle, the new silicon RAPD could operate very well even at wavelengths below 100 nm in a vacuum. Assembled into a typical TO package, the new silicon RAPD can operate at low noise levels at the responsivity in excess of 50 A/W with quantum efficiency peaking as high as 90% in the wavelength range between 255 nm and 400 nm.

ACKNOWLEDGMENT

D. Grubišić wants to express gratitude and appreciation to the Silicon Device Integration Group at Delft University of Technology, and Prof. Nanver in particular, for their help and support on this project. The authors would like to thank Diana Convay and Karl Weiss of Arizona State University LeRoy Eyring Center for Solid State Science for their help with ellipsometry and TEM characterizations. D. Grubišić would also like to thank Dr. Davorin Babić for helpful and insightful conversations during the paper preparation.

REFERENCES

[1] W.T. Tsang (Ed.), “Lightwave Communications Technology, Part D Photodetectors”, Semiconductors and Semimetals, Volume 22, Academic Press Inc., 1985.

[2] P.P. Webb, R.J. McIntyre, and J. Conrady, “Properties of avalanche photodiodes,” RCA Review, Vol 35, June 1974.

[3] Laser Components GmbH, “Silicon avalanche photodiodes”, Application Manual, http://www.lasercomponents.com/fileadmin/user_upload/home/Datasheets/lc/applikationsreport/si-apds.pdf , (February 1, 2013)

[4] P.P. Webb, Private Communications, May 4, 2004.

[5] R.J. McIntyre, “Multiplication noise in uniform avalanche photodiodes,” IEEE Trans. On Electron Devices, Vol. ED 13, January 1966.

[6] M.A. Green and M. Keevers, "Optical properties of intrinsic silicon at 300 K," Progress in Photovoltaics, p. 189-192, 1995.

[7] E.D. Palik (Ed.), “Handbook of Optical Constants of Solids”, Academic Press Inc., Orlando, Florida, 1985.

[8] F. Sarubbi, T.L.M. Scholtes, and L.K. Nanver, “Chemical Vapor deposition of a-boron layers on silicon for controlled nanometer-deep p^+n junction formation”, J. Electron. Mater., vol. 39, p. 162-173, 2010.

[9] R.D. Hudson, Jr., “Infrared System Engineering”, John Wiley & Sons Inc., 1969.

[10] S.M. Sze, “Physics of Semiconductor Devices”, Second Edition, John Wiley & Sons Inc., 1981.

175,000 Device-Hours Operation of AlGaN/GaN HEMTs on Diamond at 200°C Channel Temperature

Dubravko I. Babić, *Senior Member*, Quentin Diduck, *Member*, Chandra S. Khandavalli, Daniel Francis, Firooz N. Faili, *Member*, and Felix Ejeckam

***Abstract*—Hundred and seventy-five thousand device-hours of operating life at channel temperatures above 200°C is demonstrated on AlGaN/GaN HEMTs fabricated using GaN-on-diamond technology for the first time. No catastrophic failures and no drain-current drift larger than 10% from turning the devices on were recorded throughout this two-year DC test**.

***Index Terms*—diamond, gallium nitride, semiconductor device reliability, transistors**

I. Introduction

THE introduction of new semiconductor-device technologies carries a burden of proof that not only their performance and price can offer advantages over exiting technologies, but that the devices exhibit reliability sufficient for commercial deployment. The interest in AlGaN/GaN high-electron mobility transistors (HEMT) as potential work-horse of high-power RF and power-management applications has spurred great interest in their reliability [1], [2], [3]. AlGaN/GaN HEMTs are traditionally fabricated on sapphire, silicon carbide, or silicon. Growing GaN on these heterogeneous substrates necessitates the incorporation of nucleation layers between the substrate and the high-quality GaN epilayers. This approach carries multiple disadvantages from thermal management point of view: some substrates have poor thermal conductivity, while the nucleation layers exhibit significant thermal boundary resistance in addition to their low bulk thermal conductivity (< 25 W/mK) resulting from the high dislocation density and use of AlGaN ternary alloys. The recent attempts to minimize or alleviate some of these problems comprise attempts to integrate GaN with diamond [4], [5], [6], [7], [8], [9]. The approach of Group4 Labs, in which GaN epilayers are atomically attached to chemical-vapor deposited (CVD) diamond wafers has up to now exhibited great progress by demonstrating 100-mm GaN-on-diamond wafers [10], devices with f_T ~ 85 GHz [11], and X-band amplifiers [12]. GaN-on-diamond technology not only replaces the original, growth substrate with diamond but it also allows the removal of the poorly thermally conductive nucleation layers prior to attaching the epilayers to diamond.

The reliability of devices built by processes in which epilayers are transferred from the growth substrate to another substrate, such as, epitaxial liftoff, wafer bonding, and atomic attachment, understandingly undergo special scrutiny of the components community because mechanical handling of thin epilayers is more likely to introduce damage and stresses to the semiconductor epilayers and/or the device than traditional semiconductor fabrication in which the epilayers stay attached to the as-grown substrate. In addition, atomic attachment relies on several high-temperature process steps. Therefore, the central question posed by many interested users of this technology is whether GaN-on-diamond devices will last. This paper contains the first report of high-temperature operating life (HTOL) experiments on GaN-on-diamond high-electron mobility transistors.

The primary objective of this experiment was to check whether these devices would "last through the weekend". The result was surprising, as up to date over 175,000 device-hours have been collected on devices that operate with channel temperatures above 200°C, I_{DS} ~ 60 mA/mm, and V_{DS} up to 48 V. We describe the device design, the experimental setup, and the results of the first two-year operating life test on GaN-on-diamond HEMTs.

II. Device Structures and Life-Test Details

AlGaN/GaN high-electron mobility transistors (HEMT) were fabricated on GaN-on-diamond wafers with layer structure shown in Figure 1. Gallium-nitride on silicon wafers were purchased with standard epilayer design from Nitronex, Inc. and the GaN epilayers were transferred onto diamond using Group4 Labs' atomic attachment process [10]. The wafers used in this work were of the first generation of GaN-on-diamond devices in which the nucleation layers (shown in Figure 1**)** are still present. The wafers were processed at MicroGaN in Ulm, Germany. Free-standing GaN-on-diamond wafer diameter was 24 mm and the wafers were attached to 100-mm silicon wafers for processing.

The layout contained a variety of two gate HEMTs with no air-bridges. The gate metallization was Ni/Au, while silicon nitride was used for passivation. Device physical dimensions

Manuscript received March 11, 2012. This work was supported in part by the US NAVY SBIR Phase II Contract number N0002409C4164 and AFRL Contract number FA865009C5404.

Dubravko I. Babić, Quentin Diduck, Chandra S. Khandavalli, Daniel Francis, Firooz Faili, Felix Ejeckam are with Group4 Labs, Inc., Fremont, California, CA 94539. (e-mail: dubravko.babic@group4labs.com).

were $W = 2 \times 400$ µm, $L_G = 1.2$ and 2 µm, $L_{GD} = 3$, 4, and 5 µm. We could not resolve the effect of these dimensional variations on the performance nor the reliability, hence, in the test we treated the devices as being identical. The wafers were diced using Q-switched Nd:YAG laser [13]. The final chip size was 1.5 x 1.5 mm and it contained two HEMTs per chip. The chips were packaged into Stratedge 580286 packages using silver epoxy. The reason for silver epoxy was that solder could not wet the rough diamond back-surface; the back of the diamond wafer on these experimental devices was not coated with metal. The package lid was omitted.

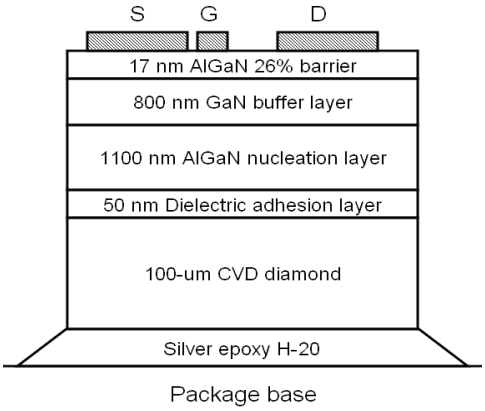


Figure 1 – GaN-on-diamond HEMT epilayer structure

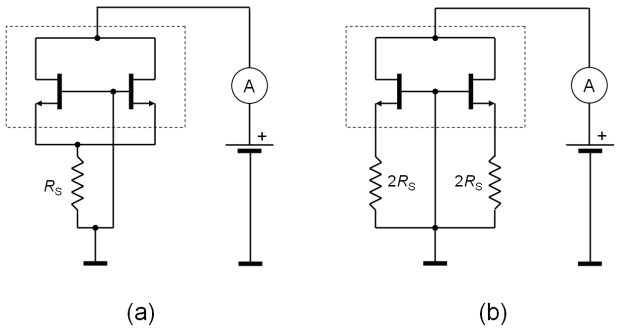


Figure 2 – Electrical wiring: (a) common source resistor, and (b) separate source resistors. The two transistors in the dashed-line box represent the two halves of a two-source HEMT on a chip.

The devices were self-biased according to the electrical scheme shown in Figure 2 which allows biasing and real time measurement of I_{DS} at constant V_{DS} with a single wire per device (also: single power supply and one meter per device). The transistor was stabilized using short bondwires and negative feedback (R_S) chip resistors inside the package. The two electrical circuit diagrams shown in Figure 2 are equivalent provided that the two sides of the two-gate transistors have identical DC characteristics. We first used diagram from Figure 2(a) with $R_S = 25$ Ω (Table 1, Groups 1-3), but subsequently realized that the circuit layout with two separate sources, common to microwave transistors, enables one to use separate source DC operating point stabilizing resistors R_S as shown in Figure 2(b) and photo in Figure 3. Using separate source resistors in each of the two sources reduces the difference between the drain currents in the two transistor halves when the transconductance differs on the two transistor halves. We estimated the reduction in the difference between the drain currents of the two halves for the transistors under test. For typical values of $I_{DSS} \sim 70$ mA and $V_{TH} \sim$ −1.3 V at operating temperature, using separate 10 Ω ($2R_S$) resistor in each source, as in case (b), rather than connecting both sources to ground via a single 5 Ω (R_S) resistor, as in case (a), results in about 20% reduction in the sensitivity of the difference between the drain currents to device parameters. This separate biasing was implemented on Group 4 devices in Table 1.

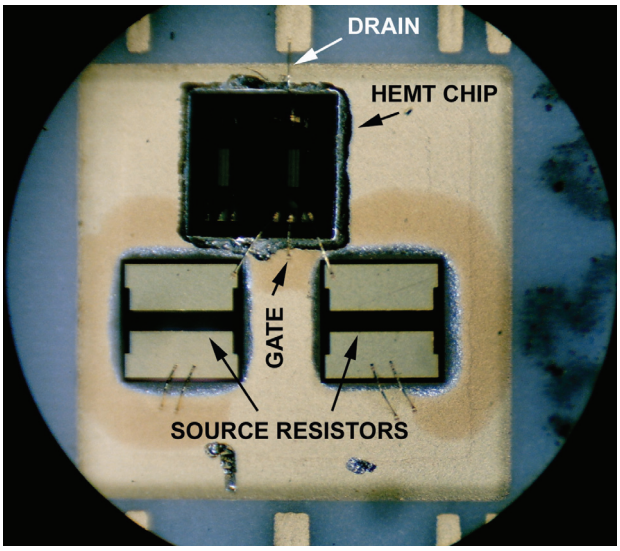


Figure 3 – Photograph of the wirebonding of a single two-source HEMT with two source resistors according to schematic shown in Figure 2.

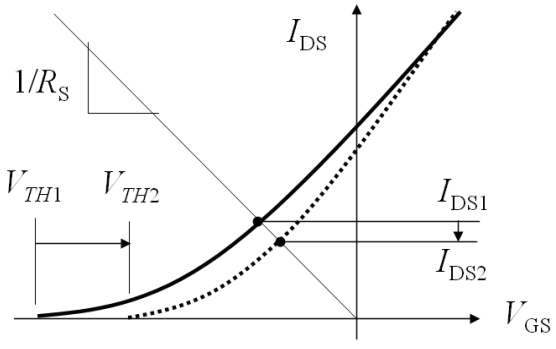


Figure 4 – The effect of transconductance or threshold voltage drift on measured drain current

The test system contained a number of 24V and 48V DC power supplies; one power supply and a dedicated analog panel ammeter per device. Analog meters offer practically infinite failure-free operation (no batteries) and no common electrical power supply lines. (Note that with the return path (ground) being common to all the devices in Figure 2, both

ammeter terminals have to be floating to make independent measurements of the drain current. The drain current data acquisition was done with no interruption to the device operation.

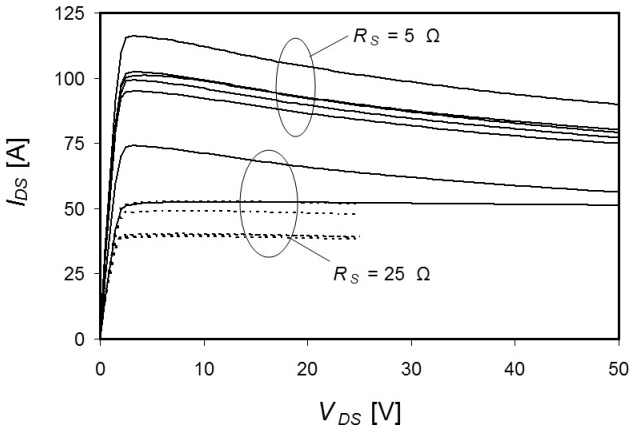


Figure 5 – Typical room-temperature DC characteristics (I_{DS} vs. V_{DS}) for devices under test connected using circuit diagram shown in Figure 12. The reduction in I_{DS} with voltage is due to self-heating and it is more pronounced for devices biased at 48V.

III. RESULTS

The operating characteristics of the HEMTs can drift in many ways during a life test. Using the circuit layout shown in Figure 2, we are primarily sensitive to the changes in I_{DSS} and V_{TH}. We do not monitor gate leakage. Figure 4 illustrates how drift in V_{TH} and/or I_{DSS} affects the I_{DS} readout. Clearly, if I_{DSS} reduces with time, the readout I_{DS} will follow proportionally. If the threshold voltage V_{TH} changes, the transfer characteristics will shift and I_{DSS} will change again producing a change in the readout I_{DS}. In Figure 4, this is illustrated with I_{DS1} changing to I_{DS2} as V_{TH1} changes to V_{TH2}. The goal is to note any gross change in device characteristics and observe any catastrophic failures. Inasmuch as there was no prior knowledge of what the outcome of this test would be, the devices were planned for a range of different stress levels: The two different bias voltages V_{DS}, two different starting drain current levels I_{DS}, and two different package base temperatures T_B were grouped into four groups of conditions listed in Table 1.

The typical measured DC drain-to-ground characteristics are shown in Figure 5, where the gradual reduction with drain current is due to self-heating. The power dissipated on the source resistors (< 60 mW) is small in comparison with the power dissipated on the transistors as seen from Table 1. The thermal resistance on a sample device was measured using Liquid Crystal Thermography (LCT) [14] to be ≈17°C/W. The nematic-isotropic transition temperature of the liquid crystal used was T_{LC} = 136°C. Figure 6 shows the LCT measurement results: variation of power required to increase the temperature from the base temperature T_B to the T_{LC}. The slope of the linear regression line gives the thermal resistance $dP/dT_B = -\Theta_{TH}$. The data correlation coefficient was equal to 99.7%

Table 1 – Summary of conditions under which the devices were operated.

Group	Ch	T_B [°C]	I [mA]	R_S [Ω]	V_{DS} [V]	P [W]	T_{CH} [°C]
1	B7	152	32.5	25	24	0.73	164
	B8	152	31	25	24	0.70	164
	B9	152	34	25	24	0.76	165
	B10	152	30	25	24	0.68	163
2	A2	176	24.5	25	24	0.56	185
	A3	176	25	25	24	0.57	186
	A4	176	27.5	25	24	0.62	187
	A5	176	27.5	25	24	0.62	187
	A6	176	28	25	24	0.63	187
	A7	176	27.5	25	24	0.62	187
3	A15	176	34.5	25	48	1.60	203
	A16	176	30.5	25	48	1.42	200
	A18	176	33.5	25	48	1.55	202
4	A11	176	49	5	48	2.33	216
	A12	176	51	5	48	2.42	217
	A13	176	41.5	5	48	1.97	210
	A14	176	54	5	48	2.56	220
	A17	176	45	5	48	2.14	212
	A19	176	58	5	48	2.75	223
	A20	176	47.5	5	48	2.26	214

Due to different bias voltage and currents flowing though the devices, the power dissipated on the devices varied. Table 1 shows a list of devices currently on test with the bias conditions (V_{DS}, R_S, measured I_{DS} and power P), package base temperature (T_B) and estimated channel temperature ($T_{CH} = T_B + \Theta_{TH} \cdot P$). The power value in Table 1 has been corrected for the power dissipated on the source resistor R_S. Ten of the devices were operated at channel temperatures above 200°C with cumulative number of device hours above 175,000. The total number of device·hours for all devices is over 350,000. The data readout was performed manually which caused a certain amount fluctuation in the values read. The results for all devices tested are shown in Figure 7 where we start the plot after the burin-in time of 10 hours (time zero).

The results show that (a) no early failures nor catastrophic failures occurred in the first 17,500 hours of operation, and (b) the drift in the I_{DS} value was less than 10% of the starting value at time zero. If one interprets the no failure data to represent random failure portion of a failure-rate curve, i.e., cumulative failure distribution being exponential, one can estimate the upper bound on the failure rate. The upper bound on the (constant) failure rate with no failures is given by $\lambda = -\ln(1-CL)/nT$, where CL is the confidence level, n number of device under test, and T the time the devices were under test [15]. If we restrict our estimate only to the ten devices with

$T_{CH} \geq 200°C$, we find $\lambda \leq$ 2.7%/K for CL = 99% (K ≡ 1000 hrs). In other words, we are 99% confident that the failure rate on these devices would be less than 2.7% for every 1000 hours. Note that this is an *upper bound* on the failure rate for given confidence level and can be used only as an estimate of the lower bound on the mean time to failure (MTTF). Without any failures and a very limited range of channel temperatures used in the test, we do not attempt to estimate the activation energy. However, an estimate of the MTTF upper bound under normal operating conditions may be obtained if we assume activation energy E_A = 1.7 eV published on devices that have the same epilayer design as our devices [16]. If the devices under test were to exhibit the same failure mechanism as the as-grown AlGaN/GaN/Si transistors, at T_{CH} = 150°C the acceleration factor would be 137. This would result in an upper bound on the failure rate of 192 FIT. The MTTF for 200°C operation is >4.3 years with confidence 99%. If activation energy 1.7 eV is assumed, then the MTTF at 150°C reaches 600 years (>5·10^6 hours) for same confidence level.

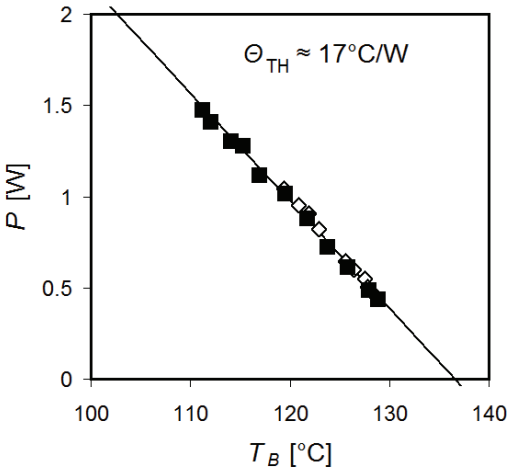


Figure 5 - Liquid crystal thermometry results at T_{LC} = 136°C.

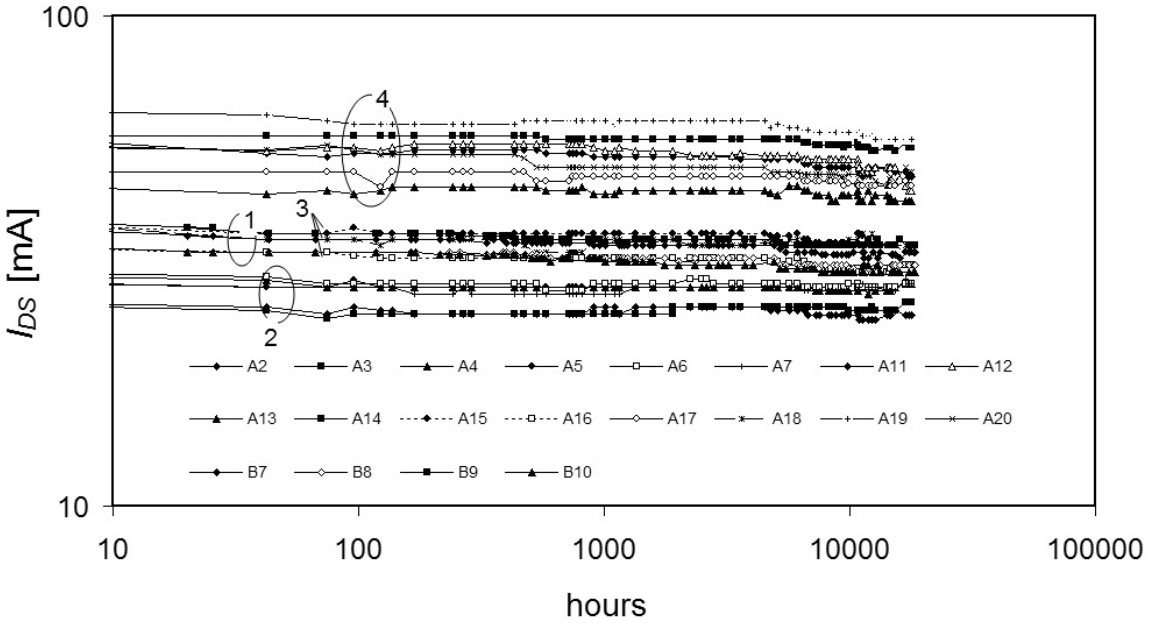


Figure 7 – I_{DS} versus time

I. Conclusions

GaN/Diamond technology using atomic attachment is progressing towards manufacturability. We have demonstrated that that transistors fabricated on these engineered wafers exhibit promising robustness: during 17,500 hours of operation the drift in drain current was less than 10% of the starting value in an experiment with 175,000 cumulative device-hours and estimated lower bound on MTTF of >5·10^6 hours.

Acknowledgment

The authors gratefully acknowledge the long-standing support of John Blevins (AFRL) and Chuck Pagel (NAVY). The authors further acknowledge the technical assistance provided by Ertugrul Sönmez at MicroGAN with wafer processing and Alex Schreiber in designing the mechanical fixtures used in this experiment.

REFERENCES

[1] S. Y. Park, C. Floreca, U. Chowdhury, J. L. Jimenez, C. Lee, E. Beam, P. Saunier, T. Balistreri, M. J. Kim, "Physical degradation of GaN HEMT devices under high drain bias reliability testing", *Microelectronic Reliability*, vol. 49, pp. 478-483, 2009.

[2] S. Singhal, T. Li, A. Chaudhari, A. W. Hanson, R. Therrien, J. W. Johnson, W. Nagy, J. Marquart, P. Rajagopal, J. C. Roberts, E. L. Piner, I. C. Kizilyalli, K. J. Linthicim, "Reliability of large periphery GaN-on-Si HFETs", *Microelectronic Reliability*, vol. 46, pp. 1247-1253, 2006.

[3] J. H. Leach, H. Morkoç, "Status of Reliability of GaN-based Heterojunction Field Effect Transistors", *Proc. IEEE*, vol. 98, no. 7, pp. 1127-1139, 2010.

[4] P.R. Hageman, J.J. Schermer, P.K. Larsen, "GaN growth on single-crystal diamond substrates by metalorganic chemical vapor deposition and hydride vapour deposition", *Thin Solid Films,* vol. 443, pp. 9-13, 2003.

[5] P.W. May, H.Y. Tsai, W.N. Wang, J.A. Smith, "Deposition of CVD diamond onto GaN", *Diamond and Related Materials* vol. 15, pp. 526-530, 2006.

[6] M. Oba, T. Sugino, "Oriented growth of diamond on (0001) surface of hexagonal GaN", *Diamond Relat. Mat.,* vol. 10, pp. 1343-1346, 2001.

[7] J. W. Zimmer, "Enhancing Semiconductors with Diamond," *Mater. Res. Soc. Symp. Proc.* 956, p. 2007.

[8] E. Piner, J. W. Zimmer, J.C. Roberts, G. Chandler, R.A. Sadler, "Epi-inverted N-Face GaN/Diamond for AlGaN/GaN/AlGaN FETs", Extended Abstracts 33rd WOCSDICE, Málaga, Spain, p. 18-21, 2009.

[9] A. Dussaigne, M. Maliverni, D. Martin, A. Castiglia, N. Grandjean, "(0001) GaN grown on (111) single crystal diamond substrate for high power electronics applications", *Proc. 18th European Workshop on Heterostructure Technology* (HETECH 2009), 2-4 November 2009, Ulm, Germany, p. 33 (2009).

[10] D. Francis, F. Faili, D. Babić, F. Ejeckam, A. Nurmiko, H. Maris, "Formation and characterization of 4-inch GaN-on-diamond substrates", *Diam. Rel. mat.*, vol 19, pp. 229-233, 2010.

[11] Q. Diduck, J. Felbinger, L. F. Eastman, D. Francis, J. Wasserbauer, F. Faili, D. I. Babić, F. Ejeckam, "Frequency performance enhancement of AlGaN/GaN HEMTs on diamond", *Electronics Lett.*, vol. 45, pp. 758-759, 2009.

[12] D. I. Babić, Q. Diduck, P. Yenigalla, A. Schreiber, D. Francis, F. Faili, F. Ejeckam, J. G. Felbinger, L. F. Eastman, "GaN-on-diamond field-effect transistors: from wafer to amplifier modules", *Proc. Symposium on Microelectronics, Electronics, and Electronic Technologies (MEET)*, MIPRO, Opatija, Croatia, May 20-24, 2010.

[13] D. I. Babić, Q. Diduck, F. Faili, J. Wasserbauer, F. Lowe, D. Francis, F. Ejeckam, "Laser machining of GaN-on-diamond wafers", *Diam. Rel. Mat.*, vol. 20, pp. 675-681, 2011.

[14] M. M. Minot, "Thermal Characterization of Microwave Power FETs Using Nematic Liquid Crystals", *IEEE MTT-S*, paper T-8, pp. 495-498, 1986.

[15] P. A. Tobias and D. C. Trinidade, *Applied Reliability*, Chapman & Hall/CRC, 1995.

[16] S. Singal, T. Li, A. Chaudhari, A.W. Hanson, R. Therrien, J.W. Johnson, W. Nagy, J. Marquart, P. Rajagopal, J.C. Roberts, E.L. Piner, I.C. Kizilyalli, K.J. Linthicum, "Reliability of large periphery GaN-on-Silicon HFETs", *Microelectronic Reliability*, vol. 46, pp. 1247-1253, 2006.

Low Power CMOS Sub-threshold Circuits

Branko Dokic* and Aleksandar Pajkanovic**
*Faculty of Electrical Engineering, Banja Luka, Republic of Srpska, Bosnia and Herzegovina
** Faculty of Technical Sciences, Novi Sad, Serbia
bdokic@etfbl.net

Abstract - Emerging new applications, such as wireless sensor nodes powered by energy harvesters, demand the change of the circuit design paradigm. Namely, for the past years, designers aimed at the minimum delay point in the energy-delay space, while the aforementioned applications require that the minimum energy point becomes the priority. One of the ways to achieve low energy consumption, while neglecting performance, is to operate the transistor in the sub-threshold (weak inversion) operation region. Since CMOS is the most widely accepted circuit design methodology, it would be the easiest and the most cost-effective to use it to design sub-threshold digital circuits. In this paper we provide an overview of sub-threshold CMOS digital logic circuits possibilities and limitations. We show that the same topology can be used, just as in strong inversion operation region. It is also shown that, comparing to the standard CMOS inverter, the speed and the temperature characteristic of the CMOS inverter with the dynamic threshold in the sub-threshold operation region is better.

I. INTRODUCTION

In the past decades the consumption decrease and speed increase problems were tackled by transistor scaling. Namely, decreasing the transistor size, the supply voltage V_{DD} is also decreased. Since the switching dissipation is proportional with V_{DD}^2, and the short-circuit dissipation with V_{DD}^3, it is obvious how the scaling influenced the energy consumption. Nevertheless, the manufacturers kept the supply voltage in the strong inversion region of transistor operation.

Besides the aforementioned operation region (which is the most frequently used), the transistor characteristics contain two more: the moderate inversion region and the weak inversion region (sub-threshold region) [1]. The moderate inversion region is sometimes considered the bottom of the strong inversion region and it is ignored in most of the literature [1]. On the other hand, the weak inversion region is of great interest. Namely, for the past decade or so, the attention of both the researchers and the integrated circuits manufacturers is focused in sub-threshold CMOS operation. In this region V_{DD} is less than the transistor threshold voltage ($V_{DDsub}<V_t$). Thus, the drain current is decreased by several orders of magnitude. This is the reason why short circuit dissipation is decreased by more than 10^5 times. It is also, the reason why the operation speed is decreased by 10^3 times, compared to the strong inversion region [1], [2], [3].

Even though such a significant operation speed decrease may seem as a limiting factor, there are applications in which it is not. Among others, one of these is energy harvesting – defined as "the conversion of ambient energy present in the environment into electrical energy" [4]. The basic idea is to use such a conversion system to create self powered microsystems. Energy can be created locally in several ways: making use of the kinetic energy of vibrations widely present in the environment, photovoltaic or thermoelectric effects provided there is sufficient incident light or temperature gradient, respectively, and so on [4]. Further, wireless sensor nodes can provide information from locations that are inaccessible to the systems that depend on cable connections – the idea is to use the energy harvester to capture small amounts of energy from the environment and use the generated electrical energy to power the nodes in wireless sensor networks [5].

The integrated circuits designer's aim changes in this new domain – the minimum consumption point is now of primary interest, not the minimum delay point (where points in delay-consumption space are defined in [3]). One of the ways to achieve the new goal is to operate the transistors in the aforementioned sub-threshold operation region. Compared to devising a new design approach for sub-threshold circuits, it would be both easier and more cost-effective to use an existing methodology. Since CMOS is the most widely accepted circuit design methodology, in this paper we provide an overview of sub-threshold CMOS digital logic circuits possibilities and limitations. We show that the same ideas and rules can be used, just as in strong inversion operation region. Analytic models are verified by Pspice analysis using 0.18 μm transistor model.

II. THE OPERATION REGIONS

There are three different transistor operation regions: weak inversion (sub-threshold), moderate inversion and strong inversion region. The name of an inversion region coincides with the level of inversion at the more heavily inverted channel end [1], [6].

The transfer characteristic is given in Fig. 1. Keeping V_{ds} fixed and varying the V_{gs} value, we can follow the transistors working point through all three operation regions. Convenient simplifications, such as these, are not known for the moderate region. Nevertheless, all-region models can be used [1]:

$$I_{ds} = \frac{W}{L}\mu C'_{ox}(2n)\varphi_t^2\left\{\left[ln\left(1+e^{(V_{gs}-V_t)/(2n\varphi_t)}\right)\right]^2 - \left[ln\left(1+e^{(V_{gs}-V_t-nV_{DS})/(2n\varphi_t)}\right)\right]^2\right\} \quad (1)$$

The moderate inversion region is often ignored in the literature (being considered the bottom of the strong inversion region), and we only mention its existence, but not further develop the subject.

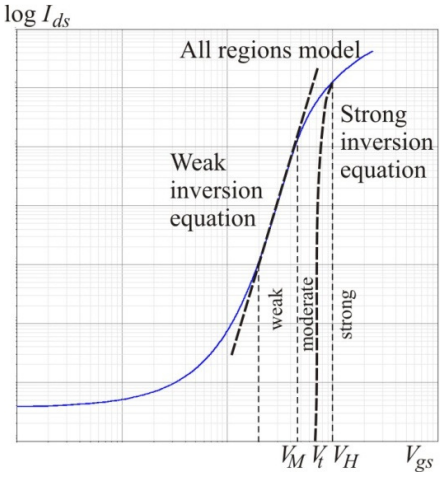


Figure 1. All transistor operation regions shown on the transfer characteristic

When operating in sub-threshold operation region, the drain current is given as [7]:

$$I_{Dsub} = \begin{cases} I_0 e^{\frac{V_{gs}-V_t}{n\varphi_t}} \left(1 - e^{-\frac{V_{ds}}{\varphi_t}}\right), V_{ds} < 3\varphi_t \\ I_0 e^{\frac{V_{gs}-V_t}{n\varphi_t}}, V_{ds} > 3\varphi_t, \end{cases} \quad (2)$$

where I_0 represents drain current when V_{gs}=V_t and is given by [7]:

$$I_0 = \mu_0 C_{ox} \frac{W}{L} (n-1)\varphi_t^2 \quad (3)$$

The parameters in Eqs. (2) and (3) are: μ_0 is the carrier mobility, $C_{ox} = \epsilon_{ox}/t_{ox}$ is the gate oxide capacitance (ϵ_{ox} is the dielectric constant and t_{ox} is the gate oxide thickness), W and L are channel width and length, V_t is the threshold voltage, $\varphi_t = kT/q$ is the thermal voltage (φ_t = 26mV at 300K), and n=1+C_d/C_{ox} is the sub-threshold slope factor.

According to Eq. (2), when V_{ds}> 3φ_t the drain current I_{Dsub} does not depend on the drain-source voltage, while this dependence is exponential when 0 ≤ V_{ds} ≤ 3φ_t, Fig. 2. Thus, analogous to the strong inversion region operation, in sub-threshold there are two regions of the I_{Dsub}= f(V_{ds}) characteristic: saturation region (V_{ds}> 3φ_t) and non-saturation region (0 ≤ V_{ds} ≤ 3φ_t). The difference is that the drain current is a quadratic function of the drain-source and gate-source voltages in strong inversion (V_{ds} in non-saturated region and V_{gs} in saturated region), while in sub-threshold this function is exponential (I_{Dsub} ~ $e^{V_{ds}}$in non-saturated and I_{Ddsub} ~ $e^{V_{gs}}$in saturated region). This shows that there is a complete analogy between CMOS circuits in strong and weak region analysis and synthesis. Some of the results published in this paper are derived in [8].

Transfer characteristics of MOS transistor are strongly temperature-dependent, Fig. 3. The main parameter responsible for this is the effective mobility, which decreases with temperature. Nevertheless, many more parameters depend on temperature, such as the thermal voltage and the intrinsic carrier concentration, n_i. Further, because of these, the Fermi potential, the threshold voltage, and so on, are temperature-dependent.

It has been reported that at high currents the decrease of the carrier mobility with temperature dominates and that at low currents, the decrease of the threshold voltage with temperature dominates. These effects are evident in the Fig. 3, where log I_d vs. V_{gs} for various temperatures is shown. We can also observe that the temperature affects the operation regions differently. The bottom, curved part of the curves is due to moderate and weak inversion. The curving of the upper part is due to the dependence of effective mobility on V_{gs}. In weak inversion, for certain V_{gs}, the drain current increases with temperature [1].

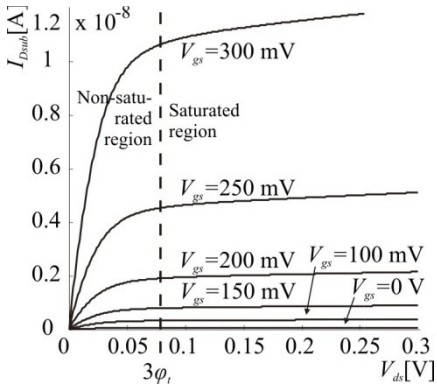


Figure 2. I_{DDsub} – V_{ds} characteristics in sub-threshold operation region

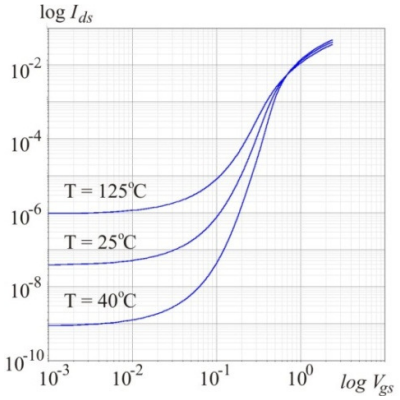


Figure 3. Temperature dependence on the transistor operation

III. STATIC CHARACTERISTICS

A. CMOS inverter

A CMOS inverter is shown in Fig. 4a. Operation region depends on the relation between supply voltage V_{DD} and NMOS and PMOS threshold voltages, V_{tn} and V_{tp}. In this paper, we assume V_{tn} = |V_{tp}| = V_t. The transistor operates in strong inversion when V_{DD}>V_{tn} + |V_{tp}| = 2V_t, while it operates in sub-threshold region when 3φ_t<V_{DD}<V_t, Fig. 4b. For low values of V_{DD}, the drain current I_{Dsub} is very low. Thus, the output voltage rise and

fall time intervals are extremely long. Sub-threshold region threshold voltage V_{tsub} can be defined as the voltage V_{gs} at which the drain current is equal to $0.01I_0$, Fig. 4b. Thus, based on Eq. (2):

$$V_{tsub} = V_t - 4.6n\varphi_t. \quad (4)$$

According to Eq. (4), supply voltage practical values in sub-threshold operation region are:

$$V_t - 4.6n\varphi_t < V_{DDsub} < V_t. \quad (5)$$

When transistor operates as a switch, it is considered turned off at $V_{gs} < V_{tsub}$ and turned on at $V_{tsub} < V_{gs} < V_{DD}$.

CMOS inverter operating in sub-threshold region transfer characteristics, both voltage and current, are shown in Fig. 5. Threshold voltage V_{Tsub} in sub-threshold is obtained, in the same way as in strong inversion, by equalizing currents $I_{Dnsub}=I_{Dpsub}$ in the saturated region of the characteristic, i.e.:

$$I_{0n}e^{\frac{V_i-V_t}{n\varphi_t}} = I_{0p}e^{\frac{-V_i+V_{DD}-V_t}{n\varphi_t}}. \quad (6)$$

Replacing $V_i = V_{Tsub}$ in Eq. (4), and solving by V_{Tsub}, we obtain:

$$V_{Tsub} = \frac{V_{DDsub}}{2} - \frac{n\varphi_t}{2} ln\left(\frac{I_{0n}}{I_{0p}}\right). \quad (7)$$

If the inverter is symmetric, i.e. if:

$$\frac{I_{0n}}{I_{0p}} = \frac{\mu_{0n}{}^{W_n}/_{L_n}}{\mu_{0p}{}^{W_p}/_{L_p}} = 1, \quad (8)$$

threshold voltage is given as $V_{Tsub} = V_{DDsub}/2$. Threshold voltage of the symmetric CMOS inverter operated in strong inversion region is also $V_T = V_{DD}/2$.

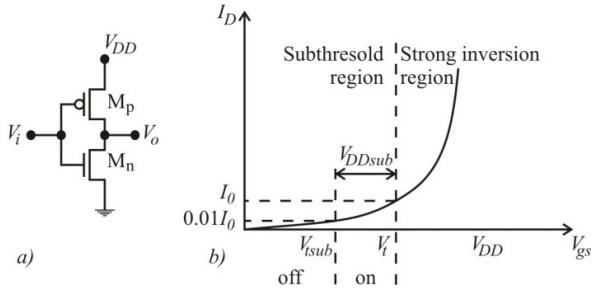


Figure 4. CMOS inverter (a) and MOS transistor I_D-V_{GS} characteristic (b)

Supply voltage current is given as:

$$I_{DDsub} = \begin{cases} I_{0n}e^{\frac{V_i-V_t}{n\varphi_t}}, & 0 \le V_i \le V_{Tsub} \\ I_{0p}e^{\frac{-V_i+V_{DDsub}-V_t}{n\varphi_t}}, & V_{Tsub} \le V_i \le V_{DDsub}. \end{cases} \quad (9)$$

Maximum value of I_{DDsub} is obtained at $V_i=V_{Tsub}$, so:

$$I_{DDMsub} = I_{0n}\sqrt{\frac{I_{0p}}{I_{0n}}}e^{\frac{V_{DDsub}/2-V_t}{n\varphi_t}}. \quad (10)$$

If the inverter is symmetric, Eq. (8), then:

$$I_{DDMsub} = I_{0n}e^{\frac{V_{DDsub}/2-V_t}{n\varphi_t}}. \quad (11)$$

Maximum short circuit currents ratio of the CMOS inverter, in strong [8] and weak inversion, Eq. (11) is given by:

$$\frac{I_{DDM}}{I_{DDMsub}} = \frac{(V_{DD}-2V_t)^2}{8(n-1)\varphi_t^2}e^{\frac{V_t-V_{DDsub}/2}{n\varphi_t}}. \quad (12)$$

For example, at V_t=500mV, V_{DD}=2V, V_{DDsub}=350mV and n=1.5, we obtain $I_{DDM}/I_{DDMsub}=0.22\times10^6$. In general, we can say that strong inversion maximum current I_{DDM} is greater than sub-threshold maximum current I_{DDMsub} by 10^4 to 10^7 times.

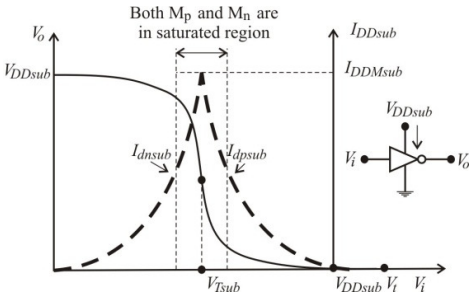


Figure 5. CMOS inverter in sub-threshold region $V_0=V_i$(a) and $I_{DDsub} - V_i$ (b) static characteristic

Eqs. (7) and (10) show that the threshold voltage and the maximum short-circuit current of the CMOS inverter while operating in sub-threshold region depend on the transistor geometry ratio $(W_n/L_n)/(W_p/L_p)$. Such claim is confirmed by static characteristics obtained by Pspice analysis, Fig. 6.

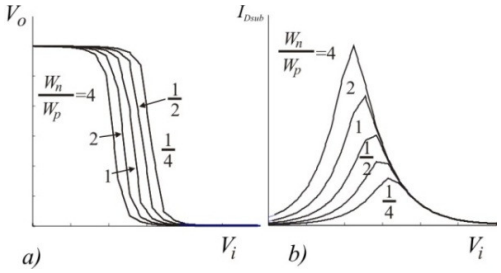


Figure 6.CMOS inverter operating in sub-threshold region voltage – results obtained by SPICE simulations: (a) and current (b) transfer characteristic, where the ratio W_n/W_p is varied for the same $L_n = L_p$

Threshold voltage V_T and current I_{DDM} in strong inversion region also depend on the ratio $(W_n/L_n)/(W_p/L_p)$.

B. NAND and NOR circuits

Complex CMOS logic circuits topology in sub-threshold operation region is the same as in strong inversion. Standard topologies consist of dual NMOS and PMOS transistor networks, Fig. 7, which, further, consist of m CMOS pairs, where m represents the number of inputs.

Just like in strong inversion region [9], the threshold voltages of NAND and NOR circuits in weak inversion region depend on the input number and on the active inputs number k, where $1 \le k \le m$. The difference is that this dependence is logarithmic in sub-threshold. Namely, m transistors in series can again be replaced by an equivalent transistor, channel length of which is $L_{eq}=mL$;

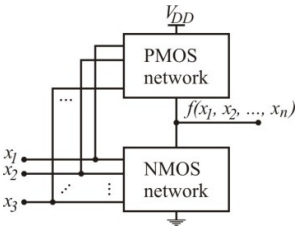


Figure 7. Standard CMOS topology

and k on transistors in parallel by an equivalent transistor, channel width of which is $W_{eq} = kW$. Based on these facts, we conclude that the threshold voltage of the m-input NAND circuit highest, and of the m-input NOR circuit the lowest, when all inputs are active (k=m). Considering Eqs. (10) and (11), the thresholds are given by:

$$maxV_{TsubNAND} = V_{TsubNOT} + k\varphi_t \ln m, \quad (13)$$

$$minV_{TsubNOR} = V_{TsubNOT} - k\varphi_t \ln m. \quad (14)$$

where $V_{TsubNOT}$ represents the threshold voltage of the inverter. Just like in the strong inversion region [9], the lowest threshold voltage $V_{TNANDmin}$ and the highest $V_{TNORmax}$ are obtained when only one input is active. In that case, these are equal to the inverter threshold voltage, i.e. $minV_{TsubNAND} = maxV_{TsubNOR} = V_{TsubNOT}$. Thus, e.g. according to the Eqs. (13) and (14), the difference between the maximum and minimum V_{Tsub}, for NAND2 and NOR2 is about 30mV; and for NAND3 and NOR3 about 44mV. This is confirmed by SPICE analysis and shown in Fig. 8.

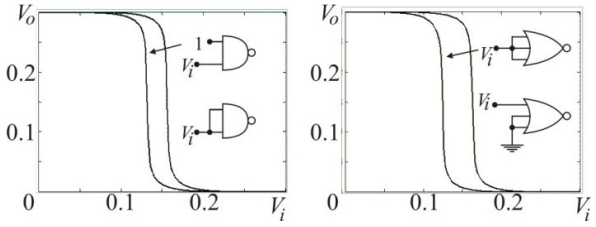


Figure 8. Transfer voltage characteristics of NAND2 and NOR3 circuits, obtained by SPICE simulations

IV. DYNAMIC MODE

In Fig. 9a CMOS inverter and its parasitic capacitances are shown and in Fig. 9b, its model for switching process analysis is shown. In [8] capacitances C_{gsp} and C_{gdn} are mapped to the input and to the output, but with capacitance values doubled:

$$C_I = C_{gsn} + C_{gsp} + 2(C_{gdn} + C_{gdp}),$$
$$C_O = C_{dsn} + C_{dsp} + 2(C_{gdn} + C_{gdp}), \quad (15)$$
$$C_L = C_O + NC_I,$$

where N is the fan-out number.

At step input voltage, when changing from 0 to V_{DD}, PMOS transistor is turned off (S_p off) and NMOS is turned on (S_n on), so the capacitor C_L is discharged by the current I_{dn}. The transistor M_n is first in saturated region and then in non-saturated region.

Capacitor C_L discharging I_{Dnsub} and charging I_{Dpsub} currents, are given by:

$$I_{Dnsub} = I_{0n}e^{\frac{V_{DDsub}-V_{tn}}{n\varphi_t}}, \quad I_{Dpsub} = I_{0p}e^{\frac{V_{DDsub}+V_{tp}}{n\varphi_t}} \quad (16)$$

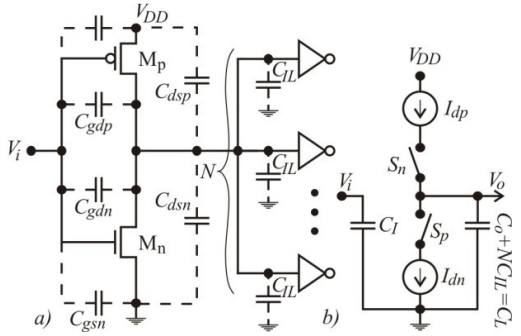


Figure 9. CMOS inverter with corresponding capacitances (a) and its dynamic model (b)

Since in the saturated region $V_{ds} < 3\varphi_t$, we can assume that the switching regime is finished when the working point enters the non-saturated region. The switching regime is defined by the C_L charge and discharge time intervals. If the inverter is symmetric I_{0n}=I_{0p}, V_{tn}=|V_{tp}|=V_t, the time intervals t_{pHL} and t_{pLH} are equal and are given by:

$$t_{pHLsub} = t_{pLHsub} = C_L \frac{V_{DD}}{I_0} e^{\frac{V_t - V_{DDsub}}{n\varphi_t}}. \quad (17)$$

Input pulses period has to satisfy the condition given by: $T_{sub} \geq t_{pHLsub} + t_{pLHsub}$.

Inverter maximum working frequency is given by:

$$F_{maxsub} = \frac{1}{T_{sub}} = \frac{I_0}{2C_L V_{DD}} e^{\frac{V_{DDsub} - V_t}{n\varphi_t}}. \quad (18)$$

When operating in weak inversion region, F_{max} is an exponential function of the supply voltage (Eqs. (18), Fig. 10). This function in strong inversion region is linear [8].

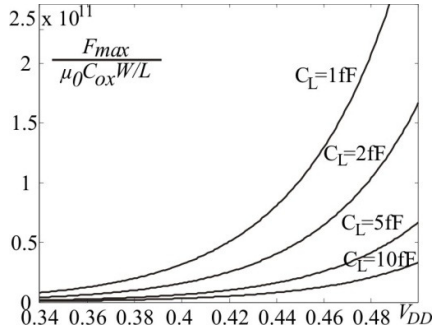


Figure 10. Normalized F_{max} as a function of V_{DD}, shown for different values of C_L

V. DISSIPATION

CMOS logic circuits have two energy consumption components: standby (static) and active (dynamic). The static consumption for both operation modes is given by:

$$P_S = V_{DD}(I_{DSS} + I_{Dleak}). \quad (19)$$

where I_{DSS} is the drain-bulk inversion current, and:

$$I_{Dleak} = I_0 e^{-\frac{V_t}{n\varphi_t}} \quad (20)$$

is sub-threshold current at $V_{gs} = 0$. Current I_{Dleak} increases exponentially when threshold voltage V_t decreases. It reaches its maximum at $V_t = 0$ ($I_{Dleak} = eI_0$).

Active (dynamic) dissipation also consists of two components: switching dissipation and transition dissipation. Switching dissipation is calculated as mean value of the transistor dissipation during one cycle of the step input voltage. It is a consequence of the C_L charging and discharging when in switching regime. The capacitor C_L is charged by the current I_{Dp} PMOS and discharged by the current I_{Dn} of the NMOS transistor. Since these currents are given by:

$$I_{Dpsub} = C_L \frac{dV_o}{dt},\ I_{Dnsub} = -C_L \frac{dV_o}{dt}, \quad (21)$$

we obtain:

$$P_{Dsub} = \frac{1}{T}\left[\int_0^{T/2} V_o(t) I_{Dn}(t)dt + \int_{T-T/2}^{T} \big(V_{DD} - V_o(t)\big) I_{Dp}(t)dt\right] = C_L {V_{DD}}^2 f, \quad (22)$$

where f=1/T represents the input pulses frequency.

Thus, switching dissipation, in both operating regions, is proportional to V_{DD}^2. Since V_{DD} in strong inversion region is several times greater than V_{DDsub} in sub-threshold, the switching dissipation is also greater by several tens when compared to the dissipation in sub-threshold region.

The above analysis is valid when ideal input voltage and ideal MOS transistors are considered. In practice, there are always input voltage rise and fall times. These are the reason why, at certain intervals, both transistors are turned on, causing additional dissipation. Thus, this dissipation is dubbed short circuit dissipation.

CMOS inverter short-circuit dissipation, when operating in strong inversion region and when the input voltage is a linear function of time, is given by [8]:

$$P_{dsc} = \frac{1}{3} I_{DDM} (V_{DD} - 2V_t)\big(t_r + t_f\big) f, \quad (23)$$

where t_r and t_f are the input voltage rise and fall times, and I_{DDM} is obtained in [8]. Since I_{DDM} is a quadratic function of V_{DD}, $P_{dsc} \sim {V_{DD}}^3$.

Supply voltage current response in sub-threshold region to an input voltage linear change is shown in Fig. 11. Dissipation mean value during one cycle is given by:

$$P_{dsc} = \frac{1}{T}\int_0^T V_{DDsub}\, i_{DDsub}(t)dt. \quad (24)$$

if the inverter is symmetric, then $I_{dnsub} = I_{dpsub}$. Thus, the dissipation is equal for all periods: 0-t_1, t_1-$T/2$, $T/2$-t_2 and t_2-T. Further, we have:

$$P_{dsc} = 4\frac{V_{DDsub}}{T} I_{0n} \int_0^{t_1} e^{\frac{St - V_t}{n\varphi_t}}, \quad (25)$$

where S=dV_i/dt represents the input voltage slope changes. When the input signal is symmetric (t_r=t_f), we have: S=V_{DD}/t_r, where t_r and t_f are rise and fall times of the input voltage. Solving Eq. (25), yields:

$$P_{dsc} = 4n\varphi_t I_{DDMsub} t_r f, \quad (26)$$

where I_{DDMsub} is given by Eq. (11). If the input voltage change is asymmetric ($t_r \neq t_f$), we obtain:

$$P_{dscsub} = 2n\varphi_t I_{DDMsub}\big(t_r + t_f\big) f \quad (27)$$

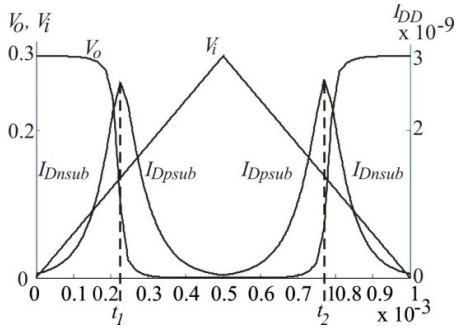


Figure 11. Supply voltage current response to an input voltage linear change when operating in sub-threshold

We have $P_{dsc} \sim V_{DD}^3$ while operating in strong inversion region and $P_{dscsub} \sim \varphi_t^3 e^{V_{DDsub}}$ in sub-threshold region. Thus, we obtain: $P_{dscsub} \ll P_{dsc}$. For example, if V_{DD}=2 V, V_t=0.5V, n=1.5 and V_{DDsub}=360mV, based on Eqs. (23) and (26) we calculate $P_{dsc}/P_{dscsub} \sim 4.7\times10^6$.

VI. DTCMOS INVERTER

Propagation time decreases exponentially as the threshold voltage V_t decreases, Eq. (18). However, as the V_t decreases, the static dissipation increases. A valid compromise can be achieved if dynamic threshold CMOS logic is used – DT-CMOS [6], [10]. Here, the transistor substrates are tied to the gates, Fig. 12.

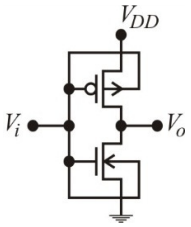


Figure 12. DT-CMOS inverter

The MOS transistor threshold voltage is given by:

$$V_t = V_{t0} + \gamma\left(\sqrt{2\Phi_F - V_{BS}} - \sqrt{2\Phi_F}\right), \quad (28)$$

where V_{BS} represents the bulk-source voltage, $V_{t0}$the threshold voltage at V_{BS} = 0, γ body-effect coefficient and Φ_F Fermi potential. Since V_{BSn}=V_i, and V_{BSp}=-V_{DD}, threshold voltages of the DTNMOS and DTPMOS transistors depend on the input voltage. Namely, at V_i=0, we have: V_{tn}=V_{t0}, and at V_i=V_{DD}, we see from Eq. (28) that it has to be: V_{tn}<V_{t0}. Therefore, when the transistor is turned off, the DTNMOS threshold voltage is the same as the NMOS threshold voltage, so the static dissipation is also the same. However, when the transistor is turned on, the DTNMOS threshold voltage is lower than the NMOS threshold voltage. Thus, in this case, the capacitor C_L

discharge current is greater and the discharge time shorter. The same conclusion is derived for the DTPMOS threshold voltage and the current, where $|V_{tp}|=V_{t0}$ at $V_i=V_{DD}$ and $|V_{tp}|<V_{t0}$ at $V_i=0$.

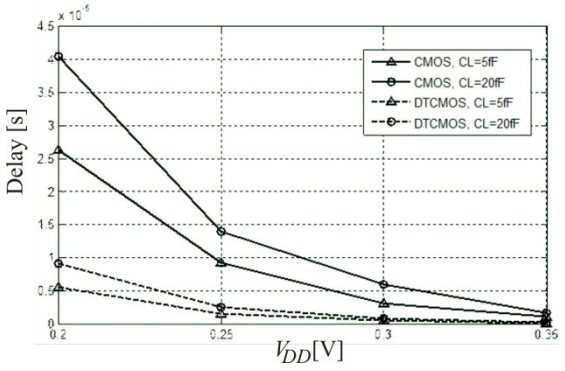


Figure 13. DTCMOS and CMOS symmetric inverters delay characteristics, obtained by SPICE.

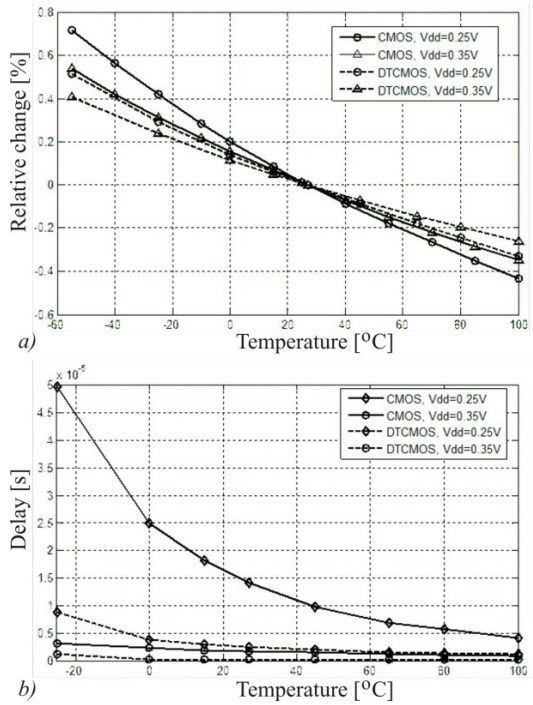


Figure 14. DTCMOS and CMOS inverter temperature characteristics, obtained by SPICE: a) relative changes of the threshold voltage, and b) logic delays.

Since the threshold voltage and the current I_{Doff}, in the off state, are equal to those when regular MOS is used, and, in the on state, the threshold voltage is lower, and the current I_{Don} is greater, the conclusion is that the logic delay of the DTCMOS inverter is shorter than that of the CMOS inverter, at the same static dissipation. In Fig. 13 DTCMOS and CMOS inverter logic delays are shown as functions of the supply voltage V_{DD} and the load capacitance C_L. At the same V_{DD} and C_L, the DTCMOS time delay is several times shorter, than that of the CMOS inverter.

Besides increased speed, DTCMOS inverter temperature characteristics are improved, Fig. 14. In Fig. 14a, relative threshold voltage changes are shown, and in Fig. 14b DTCMOS and CMOS logic delay is shown as the function of the temperature, and the supply voltage as a parameter, at C_L=20 fF.

We emphasize that the supply voltage V_{DD} of the inverter limited by the threshold voltage of the diode, created by the PN bulk-drain junction ($V_{DD}<V_{Df}\approx$0.6V). In other words, the connection in Fig. 12 works only in the sub-threshold operation region.

VII. CONCLUSION

The maximum short-circuit dissipation of the CMOS inverter is lower 10^4 to 10^7 times when operating in sub-threshold operation region than in strong inversion operation region. Switching dissipation is lower by several decades at the same working frequency. However, logic delay is 10^3 to 10^4 times greater in weak inversion region.

It is very important that there is a complete behavior analogy of digital circuits in both operation regions. The difference is in the functional dependence of the parameters. In strong inversion region, the dependence is quadratic, i.e. square root, and in sub-threshold exponential, i.e. logarithmic. The behavior analogy enables the usage of the same topologies for digital circuits in both operation regions.

At equal supply voltage and load capacitance values, the DTCMOS time delay is several times shorter, than that of the CMOS inverter. Besides the decreased time delay, the DTCMOS temperature characteristics are better.

REFERENCES

[1] Y. Tsividis and C. McAndrew, *Operation and Modeling of the MOS Transistor, 3rd Ed.* Oxford University Press, 2011.

[2] B. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in sub-threshold circuits," *IEEE Journal of Solid-State Circuits*, pp. 1778–1786, 2005.

[3] D. Markovic, C. Wang, L. Alarcon, L. Tsung-Te, and J. Rabaey, "Ultralow-power design in near-threshold region," *Proceedings of the IEEE*, pp. 237–252, 2010.

[4] D. Zhu, J. Tudor, and S. Beeby, "Strategies for increasing the operating frequency range of vibration energy harvesters: a review," *Measurement Science and Technology*, vol. 21, no. 2, pp. 1–29, 2010.

[5] T. Kazmierski, L. Wang, and M. Aloufi, "Energy efficient sensor nodes powered by kinetic energy harvesters design for optimum performance," *Electronics*, vol. 16, no. 1, pp. 65–76, 2012.

[6] Abhinov, Sanjeev Rai and Rajeev Tripathi: A robust Design for Ultra Low Power Operation using Dynamic Threshold SCL Logic, Int. Conf. on Comp. and Commun. Tech (ICCCT)-2011.

[7] A. Wang, B.H. Calhoun, A.P. Chandrakasan, *Sub-threshold Design for Ultra Low-power Systems*, Springer, 2006.

[8] B. Dokic, A. Pajkanovic, „Sub-thresholdOperatedCMOSAnalyticModel", *ProceedingsoftheIXInternationalSymposiumonIndustrialElectronicsINDEL 2012*, Banja Luka, November 1st-3rd, 2012, pp. 80-85;

[9] B. Dokic,: "Influence of Series and Parallel Transistors on DC Characteristic of CMOS Logic Circuits", *Microelectronics Journal*, Vol. 13, No. 2, 1982, pp. 25-30

[10] Hendrawan Soeleman Etal: Robust Ultra-Low Power Sub-threshold DTMOS Logic, ISLPED '00 Rapalloza, Italy

Low-power and high-speed 4-2 compressor

Ardalan Najafi*, Babak Mazloom-nezhad*, Amir Najafi**
* Dept. of Electrical and Computer Engineering, Shahid Beheshti University G. C., Tehran, Iran
** Dept. of Electrical Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
ar.najafi@mail.sbu.ac.ir , b-mazloom@sbu.ac.ir , amir.najafi@qiau.ac.ir

Abstract—Compressors are the most important component of multipliers. Multipliers themselves are important components that dictate the overall performance of arithmetic circuits. In this paper, a new 4-2 compressor architecture is proposed. This architecture uses Carry Generator Module (CGEN) which has been used for 5-2 and 7-2 compressors already. The proposed architecture is compared with the best architecture presented in the literature in terms of power and delay. Interestingly, analysis shows that the proposed compressor architecture outperforms the best existing architecture both in terms of power and delay. Architectures are simulated in 180nm and 130nm technologies. In both 180nm and 130nm technologies the proposed architecture reduces power-delay product (PDP) parameter by 17.01% and 39.04%, respectively.

Keywords—4-2 Compressor, High performance arithmetic circuits, Multiplier, Low-power design

I. INTRODUCTION

Making high performance multipliers, to consume less power and operate faster, has attracted lots of researchers' attention due to a rapid growth of using multipliers. Fast arithmetic circuits including multipliers are the most frequently used circuits in VLSI (Very Large Scale Integration) systems. Microprocessors and digital signal processors use multipliers for algorithms such as convolution and filtering. Since multipliers are constrained by power consumption and computation speed, they dictate the overall circuit performance in most of applications.

Multiplication process includes 3 steps: 1) partial product generation; 2) partial product reduction; 3) Final addition with carry propagating. The second step makes the worst-case delay, consumes the main part of power, and occupies the high fraction of silicon area. To decrease the latency of this step, compressors have been widely employed. Therefore, to have a high performance multiplier and subsequently a fast and low-power arithmetic cell, it's important to design low-power and high-speed compressors.

To achieve a better performance, use of higher order compressors have been considered. The 4-2 compressors were first proposed by Weinberger in [1] as an alternative to 3:2 counters. Several 4-2 compressors have been proposed for low-power applications [2-7]. By a survey on these structures, we found that 4-2 compressor proposed in [7] has the best result. In [7], the authors proposed a new 4-2 compressor and compared it with the previous existing one.

In this paper, we propose a 4-2 compressor architecture. We use the CGEN block which has been used in 5:2 and 7:2 compressor previously [6, 8]. By using Design Compiler software we demonstrate that our proposed architecture outperforms the best one proposed in [7], in terms of power and delay.

The paper is organized as follows: in section II, we describe previous works on 4-2 compressors, especially the one proposed in [7]. In section III, we explore our proposed compressor architectures. In section IV, the simulation results and comparisons are detailed. Finally, in section V, we conclude the paper.

II. PREVIOUS WORKS

A 4-2 compressor has five inputs and three outputs. Four inputs are primary inputs and the other is input carry which receives its value from previous stage of one bit lower in significance. The four primary inputs x_1 , x_2 , x_3 and x_4 , and the output Sum have the same weight. The output carry is weighted one binary bit order higher. The 4-2 compressor receives an input carry C_{in} from the preceding module of one binary bit order lower and produces an output C_{out} to next compressor module of higher significance. Every 4-2 compressor has to abide by the following equation:

$$x_1 + x_2 + x_3 + x_4 + C_{in} = Sum + 2(carry + C_{out}) \quad (1)$$

The conventional implementation of a 4-2 compressor is obtained by two serial full adders. This structure is shown in Figure 1.

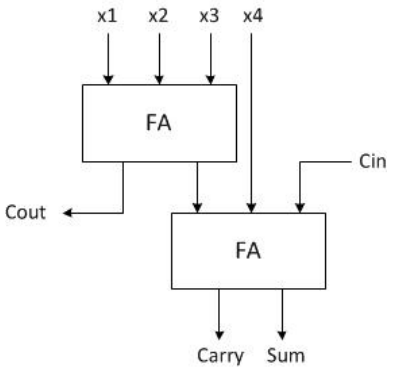


Figure 1. Conventional implementation of 4-2 compressor

Different structures of 4-2 compressors exist. A 4-2 compressor described with the following equations is shown in Figure 2 [6].

$$C_{out} = (x_1 \oplus x_2).x_3 + x_1x_2 = (x_1 \oplus x_2).x_3 + \overline{(x_1 \oplus x_2)}.x_1 \quad (2)$$

$$Sum = x_1 \oplus x_2 \oplus x_3 \oplus x_4 \oplus C_{in} \quad (3)$$

$$carry = (x_1 \oplus x_2 \oplus x_3 \oplus x_4).C_{in} + \overline{(x_1 \oplus x_2 \oplus x_3 \oplus x_4)}.x_4 \quad (4)$$

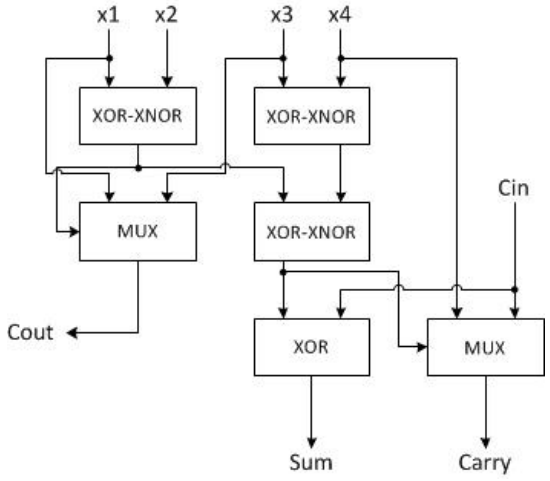


Figure 2. 4-2 compressor presented in [6]

In this structure XOR-XNOR modules generate XOR and XNOR signals simultaneously to other modules driven by them. Hence, the MUX modules don't need extra inverter block for their select inputs. The XOR outputs of XOR-XNOR modules are fed to select inputs and XNOR outputs are fed to complements of select inputs.

Another structure of 4-2 compressors which can be considered as the best one is proposed in [7]. This structure is the modified architecture of Figure 2. This architecture which is shown in Figure 3 has been used for research by several authors in recent years [9].

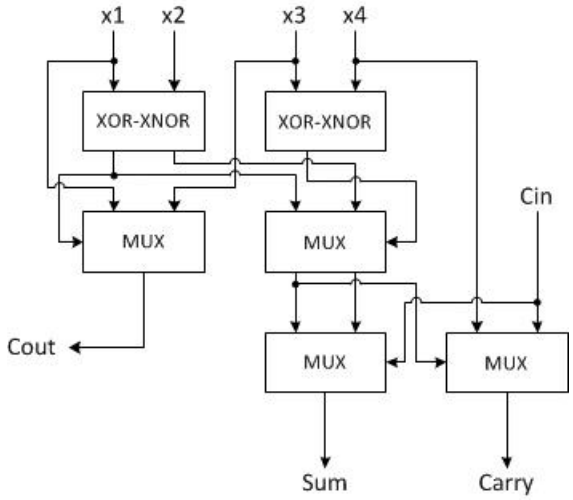


Figure 3. 4-2 compressor proposed in [7]

The equations governing the outputs in this architecture are the same as the equations mentioned above for previous version of 4-2 compressor.

This Architecture will be called ***Arch1*** in the rest of this paper.

To improve compressors' performance, carry generator (CGEN) blocks have been employed in some structures. A CGEN is a block which receives three inputs and produces carry signal according to TABLE I, that is the truth table of this block. Following equation describes the output of CGEN block:

$$carry = (a+b).C_{in} + a.b \quad (5)$$

TABLE I. TRUTH TABLE OF CGEN BLOCK

A	b	Cin	carry
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

CGEN block which has been used in the architecture of 5-2 and 7-2 compressors is shown in Figure 4 [6, 8].

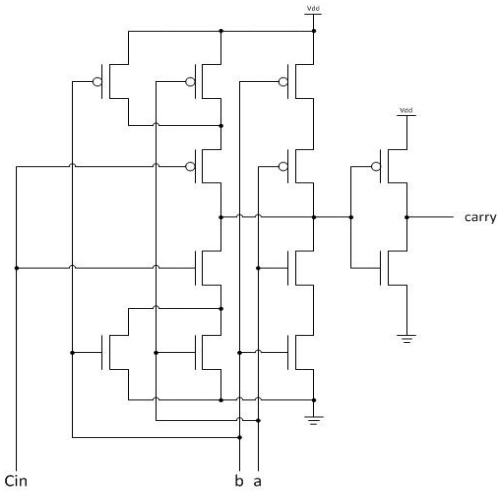


Figure 4. CMOS-CGEN Block

III. PROPOSED 4-2 COMPRESSOR

We proposed a new and faster compressor architecture, which will be called ***Arch2*** in the rest of this paper. This architecture, which uses CGEN block to produce outputs C_{out} and carry, is shown in Figure 5.

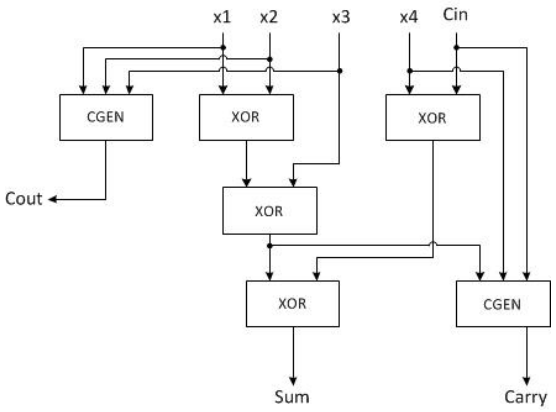


Figure 5. The proposed 4-2 compressor architecture

One of the differences between our proposed architecture and the previous version is different internal equations. In previous version, at first, two blocks produce $x_1 \oplus x_2$ and $x_3 \oplus x_4$ signals. Then outputs of these blocks are fed to inputs of another block to produce $x_1 \oplus x_2 \oplus x_3 \oplus x_4$ signal. Hence this signal is used to produce Carry signal by feeding to a MUX block. But, in our proposed architecture, an XOR gate produces $x_1 \oplus x_2$ signal. Then another XOR gate produces $x_1 \oplus x_2 \oplus x_3$ signal, which is required to feed to CGEN block to produce Carry signal. This difference is because of using CGEN block and is necessary for correct function of the architecture.

Using CGEN block to produce C_{out} signal, eliminates the need of waiting for $x_1 \oplus x_2$ signal, and hence lower delay to produce C_{out} signal has been achieved. Accordingly, the mechanism for generating carry output signals of Arch2 is different from those of existing architectures. The output functions of our proposed architecture are described by the following equations:

$$C_{out} = (x_1 + x_2).x_3 + x_1 x_2 \quad (6)$$

$$Sum = x_1 \oplus x_2 \oplus x_3 \oplus x_4 \oplus C_{in} \quad (7)$$

$$carry = (x_4 + C_{in})(x_1 \oplus x_2 \oplus x_3) + x_4.C_{in} \quad (8)$$

Although our proposed architecture uses different equations to generate carry outputs, it still abides by (1), the general equation of 4-2 compressors.

IV. Simulation Results

The simulations are performed by Design Compiler software and are targeted in 0.18 µm and 0.13 µm technologies. Accordingly, we divide this section into two parts: 0.18 µm simulation results and 0.13 µm simulation results.

A. 0.18µm Simulation Results

As the authors in [7] claimed the best performance of Arch1 among other existing architectures, we have selected this architecture to be compared to our proposed one. The simulation results are tabulated in TABLE II. As you can see in this table, Arch2 reduces both power and delay by 11.62% and 6.11% respectively. Hence, Arch2 reduces the power-delay product by 17.01%.

TABLE II. Comparison of Arch1 and Arch2 in 0.18µm technology

	power	delay	power-delay product
Arch1	84.4914 µw	0.6903 ns	58.32 *fj*
Arch2	**74.6746 µw**	**0.6481 ns**	**48.40 *fj***
Improvement percentage	11.62%	6.11%	17.01%

Design Compiler software let us to set some constraints. By applying some timing constraints we obtained a line chart. This power versus delay chart is shown in Figure 6.

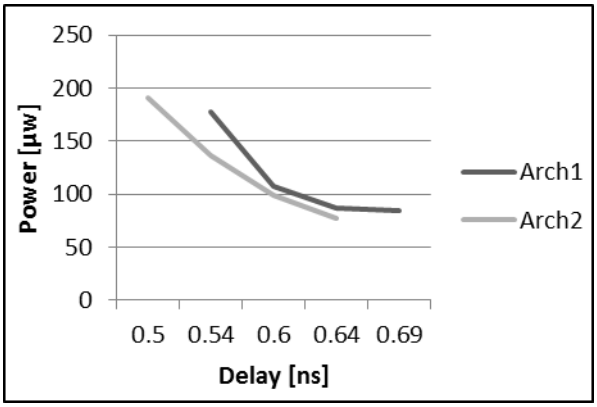


Figure 6. Power versus Delay in 180nm Technology

As it is obvious from Figure 6 and TABLE II, our proposed compressor architecture outperforms Arch1 both in terms of delay and power for 180nm technology.

B. 0.13µm Simulation Results

We have also compared Arch1 and Arch2 under 130nm technology. The results are tabulated in TABLE III As this table shows, power has been reduced by 25.52% and delay is reduced by 18.15% for Arch2 in comparison to Arch1. Hence, Arch2 reduces power-delay product by 39.04%.

TABLE III. COMPARISON OF ARCH1 AND ARCH2 IN 0.13µM TECHNOLOGY

	power	**delay**	**power-delay product**
Arch1	29.9805 µw	0.7671 ns	23 *f*j
Arch2	**22.3282 µw**	**0.6279 ns**	**14.02 *f*j**
Improvement percentage	25.52%	18.15%	39.04%

This table shows that Arch2 in 0.13µm technology outperforms Arch1 in terms of power and delay as in 0.18µm technology.

V. CONCLUSION

In this paper, we proposed a low-power and high-speed 4-2 compressor. The new compressor is proposed using CGEN blocks. We compared our novel architecture with its optimal counterpart. According to the results, a remarkable reduction in power and delay is achieved. The simulations have been performed by Design Compiler software in 180nm and 130nm technologies. As simulation results demonstrate, our proposed architecture reduces power by 11.62% and 25.52% in 180nm and 130nm technologies, respectively. It also reduces delay by 6.11% and 18.15% in 180nm and 130nm technologies, respectively.

[1] A. Weinberger, "4-2 carry-save adder module," *IBM Technical Disclosure Bulletin*, vol. 23, 1981.

[2] S. F. Hsiao, M. R. Jiang, and J. S. Yeh, "Design of high-speed low-power 3-2 counter and 4-2 compressor for fast multipliers," *Electronics Letters*, vol. 34, pp. 341-343, 1998.

[3] M. Margala and N. G. Durdle, "Low-power low-voltage 4-2 compressors for VLSI applications," in *Proc. IEEE Alessandro Volta Memorial Workshop Low-Power Design*, 1999, pp. 84-90.

[4] D. Radhakrishnan and A. P. Preethy, "Low-power CMOS pass logic 4-2 compressor for high-speed multiplication," in *Proc. of the 43th IEEE Midwest Symposium on Circuits and Systems*, 2000, pp. 1296-1298.

[5] K. Prasad and K. K. Parhi, "Low-power 4-2 and 5-2 compressors," in *Proc. of the 35th Asilomar Conf. on Signals, Systems and Computers*, 2001, pp. 129-133.

[6] C. H. Chang, J. Gu, and M. Zhang, "Ultra low-voltage low-power CMOS 4-2 and 5-2 compressors for fast arithmetic circuits," *IEEE Transactions onCircuits and Systems*, vol. 51, pp. 1985-1997, 2004.

[7] S. Veeramachaneni, K. M. Krishna, L. Avinash, S. R. Puppala, and M. B. Srinivas, "Novel Architectures for High-Speed and Low-Power 3-2, 4-2 and 5-2 Compressors," in *Proc. of 20th Int. Conf. on VLSI Design*, 2007, pp. 324-329.

[8] M. Rouholamini, O. Kavehie, A. P. Mirbaha, S. J. Jasbi, and K. Navi, "A New Design for 7:2 Compressors," in *Proc. IEEE/ACS Int. Conf. on Computer Systems and Applications*, 2007, pp. 474-478.

[9] J. Tonfat and R. Reis, "Low power 3–2 and 4–2 adder compressors implemented using ASTRAN," in *Proc. IEEE Third Latin American Symposium on Circuits and Systems (LASCAS)*, 2012, pp. 1-4.

High-Order Inductorless Elliptic Filter with Reduced Number of Capacitors Using Signal-Flow Graphs

Đani Glavinić * and Dražen Jurišić *
* University of Zagreb/Faculty of Electrical Engineering and Computing
Unska 3, HR-10000 Zagreb, Croatia
dani.glavinic@fer.hr; drazen.jurisic@fer.hr

Abstract—In this paper a method of realizing seventh-order elliptic filter using signal-flow graphs is presented. The elliptic filter having a minimum number of capacitors is compared with the filter having a higher number of capacitors. The version with minimum number of capacitors provides area savings in IC form. Both filters (i.e. filter with minimum and one with higher number of capacitors) have a low sensitivity to component tolerances in the pass band according to Orchard's theorem. The seventh-order elliptic filter has three parallel capacitors forming three parallel tanks, and therefore has three finite elliptic transfer-function zeros. The realizations of one and two parallel capacitors have already been presented elsewhere. The sfg derivations necessary to realize seventh-order filter having additional resistive network is presented, which is very complicated in the case of three zeros. Transfer function magnitudes are simulated using the PSpice program. Monte Carlo runs confirm the low sensitivity to component tolerances of both circuit types.

Keywords: Seventh-order elliptic filters, IC design, signal-flow graphs, small chip area, low sensitivity.

I. INTRODUCTION

This paper presents a method in that signal-flow graph (sfg) derivations are applied to the passive-LCR ladder filters, in order to derive active-RC filter circuits. In recent paper [1] the method was applied to the elliptic filters of low order (up to sixth order), whereas in this paper *high-order* (i.e. seventh-order) elliptic filter is derived. The active-RC filters obtained by this method have a low sensitivity to component tolerances according to Orchard's theorem, because they simulate passive-LCR ladder filter terminated with equal resistors in both ends [2]. When compared to allpole case, elliptic filters have additional capacitors to form finite zeros in transfer function. These zeros are realized by parallel LC tanks in series branches of ladder network. There are two main approaches how to realize those additional capacitors in the inductorless version of the filter (i.e. which is obtained by the simulation of passive-LCR ladder filter using signal-flow graphs). Two approaches are: (i) the filter with additional capacitive network (common approach), (ii) the filter with additional resistive network (new approach presented in this paper and in [1]). It will be confirmed by Monte Carlo runs using Cadence PSpice 16 [3] that both circuit types have low sensitivity to component tolerances.

Resistors are preferred over capacitors on a chip because they use less chip area and are easier to manufacture. The design method illustrated in [1] is for the case of low-order elliptic filters having *one* or *two* additional capacitors. In this paper the method is extended to the high-order filter having *three* capacitors. A systematic way is presented in which signal-flow graph is transformed to construct elliptic filter with canonic number of capacitors.

II. ELLIPTIC FILTERS WITH THREE TANK CAPACITORS

Consider the doubly terminated seventh-order low-pass elliptic passive-LCR ladder filter in Fig. 1(a) (as in [4][5]).

With Kirchhoff's laws and the voltage-current branch relations of the passive-LCR ladder filter in Fig. 1(a), we obtain the set of equations:

$$
\begin{aligned}
I_S &= \frac{1}{R_S}\cdot(V_S - V_1),\\
V_1 &= \frac{1}{sC_1}\cdot(I_S - I_2),\\
I_2 &= \left(\frac{1}{sL_2} + sC_2\right)\cdot(V_1 - V_3),\\
V_3 &= \frac{1}{sC_3}\cdot(I_2 - I_4),\\
I_4 &= \left(\frac{1}{sL_4} + sC_4\right)\cdot(V_3 - V_5),\\
V_5 &= \frac{1}{sC_5}\cdot(I_4 - I_6),\\
I_6 &= \left(\frac{1}{sL_6} + sC_6\right)\cdot(V_5 - V_7),\\
V_7 &= \frac{1}{sC_7}\cdot(I_6 - I_L),\\
I_L &= \frac{1}{R_L}V_{out}\\
V_{out} &= V_7.
\end{aligned}
\qquad (1)
$$

Note that:

$$I_2 = I_{L2} + I_{C2}, \quad I_4 = I_{L4} + I_{C4}, \quad I_6 = I_{L6} + I_{C6}. \quad (2)$$

Reformulating the system of equations (1) in order to obtain an appropriate sfg representation, we obtain:

$$\begin{aligned}
I_S &= \frac{1}{R_S}\cdot(V_S - V_1),\\
V_1 &= \frac{1}{sC_1'}\cdot(I_S - I_{L2}) + \frac{C_2}{C_1'}\cdot V_3,\\
I_{L2} &= \frac{1}{sL_2}\cdot(V_1 - V_3),\\
V_3 &= \frac{1}{sC_3'}\cdot(I_{L2} - I_{L4}) + \frac{C_2}{C_3'}\cdot V_1 + \frac{C_4}{C_3'}\cdot V_5,\\
I_{L4} &= \frac{1}{sL_4}\cdot(V_3 - V_5),\\
V_5 &= \frac{1}{sC_5'}\cdot(I_{L4} - I_{L6}) + \frac{C_4}{C_5'}\cdot V_3 + \frac{C_6}{C_5'}\cdot V_7,\\
I_{L6} &= \frac{1}{sL_6}\cdot(V_5 - V_7),\\
V_7 &= \frac{1}{sC_7'}\cdot(I_L - I_{L6}) + \frac{C_6}{C_7'}\cdot V_5,\\
I_L &= \frac{1}{R_L}V_{out},\\
V_{out} &= V_7.
\end{aligned} \quad (3)$$

where

$$\begin{aligned}
C_1' &= C_1 + C_2, & C_3' &= C_2 + C_3 + C_4,\\
C_5' &= C_4 + C_5 + C_6, & C_7' &= C_6 + C_7.
\end{aligned} \quad (4)$$

The sfg in Fig. 1(b) represents the system of equations (3).

In order to obtain voltage transfer functions for the paths of the sfg it is necessary to multiply the current nodes [e.g. in Fig. 1(b)] by a resistance, [e.g. R_0 [Ω] in Fig. 2(a)]. As already shown in [1] and repeated here multiplying a node by some factor (the resistor R_0 in our case) means that all *outgoing* paths from this node will be multiplied by that factor. Thus, all *incoming* paths to that same node must be *divided* by the same factor. This is shown for the sfg in Fig. 1(b), and results in Fig. 2(a) and Fig. 2(b). To obtain a one to one relationship between currents and voltages, we select a value of R_0=1Ω.

Two possible realizations of elliptic filters having three parallel tanks with C_2, C_4 and C_6 can be obtained from the system of equations (3); they are shown by sfgs in Fig. 2(a) and Fig. 4(a), respectively. In what follows we compare the two circuits resulting from these two sfgs with regard to complexity (number of resistors, capacitors, and chip area), and sensitivity.

TABLE I. ELEMENT VALUES OF LADDER-LCR FILTER IN FIG. 1(A)

Type	R_S	C_1	L_2	C_2	C_3	L_4
	1.0	1.38237	1.22404	0.12570	1.85021	0.91785
CC 07 25 50	R_L	C_4	C_5	L_6	C_6	C_7
	1.0	0.61906	1.66398	0.96224	0.43348	1.14130

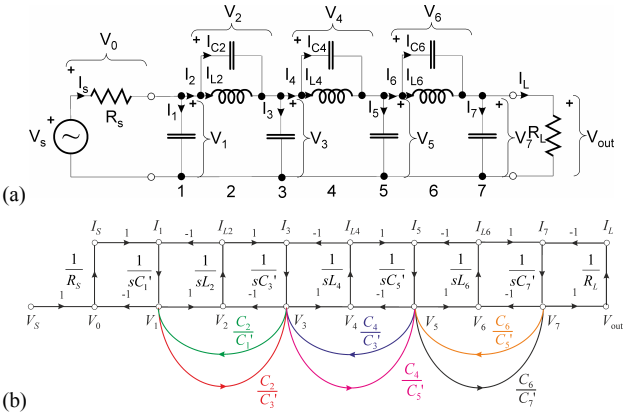


Figure 1. (a) Seventh-order passive-LCR ladder filter with designated voltages and currents. (b) Corresponding signal-flow graph [see (3)].

Consider a seventh-order elliptic (or Chebyshev-Cauer) filter with a reflection coefficient ρ=25% (corresponding to A_{max}=0.280287dB), and a normalized stop-band edge frequency ω_s=1.30541, which corresponds to the modular angle $\Theta=\sin^{-1}(\omega_c/\omega_s)$=50°. The pass-band edge frequency is normalized to unity (ω_c=1). Using filter tables (e.g. [4]) we obtain A_{min}=64.3615dB. This filter is referred to as CC 07 25 50 in most filter handbooks.

The resulting normalized voltage transfer function is given by:

$$T(s) = \frac{V_{out}}{V_{in}} = \frac{0.0044883\cdot(s^2+1.76)}{(s+0.3764)(s^2+0.5822s+0.3953)} \times \\ \times \frac{(s^2+2.397)(s^2+6.50)}{(s^2+0.2918s+0.8051)(s^2+0.08155s+1.027)}, \quad (5)$$

and the resulting normalized component values (all of which are available from filter design handbooks) are given in Table I.

A. Networks with Additional Capacitors

The conventional method of deriving elliptic filters with additional capacitors is given in [6] and [7]. For our example, the obtaining an active-RC filter from its sfg, in a conventional way, starts with Fig. 2(a). After multiplication by R_0 sfg has all voltage nodes as shown in Fig. 2(b). Note that some incoming branches, which perform addition of signals have gain –1; this will result in some negative components. To obtain all *positive* components, we have to further transform the sfg in Fig. 2(b). Appropriate nodes have to be multiplied, or scaled, by –1 [see Fig. 2(b)–(d)] which results in sfg in Fig. 2(e) having positive and negative integrators, and all positive incoming branches. After "loop reduction" rule was applied, final sfg in Fig. 2(f) is obtained.

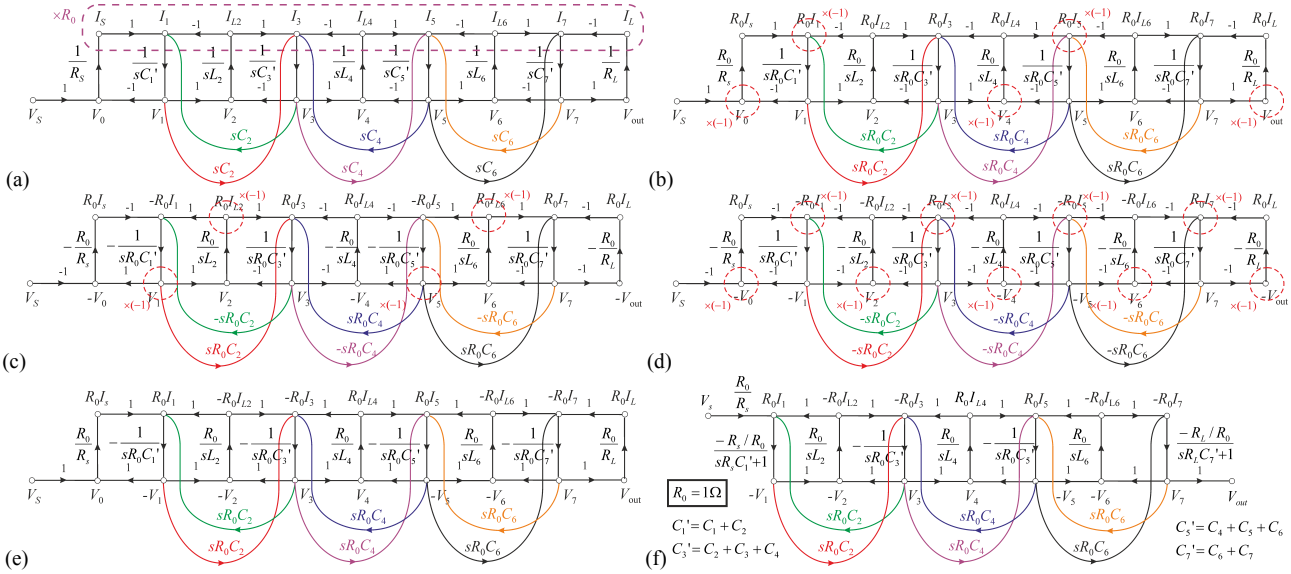


Figure 2. Deriving an active-RC simulation of a seventh-order passive-LCR ladder filter as in Fig. 1. (a) Sfg is obtained from sfg in Fig 1(b) by node shifting and scaling current nodes by R_0. (b)–(e) Multiply appropriate nodes by –1 in order to obtain only positive components with positive and negative integrators. (f) Loop reduction (outer integrators are "lossy" and belong to terminating resistors).

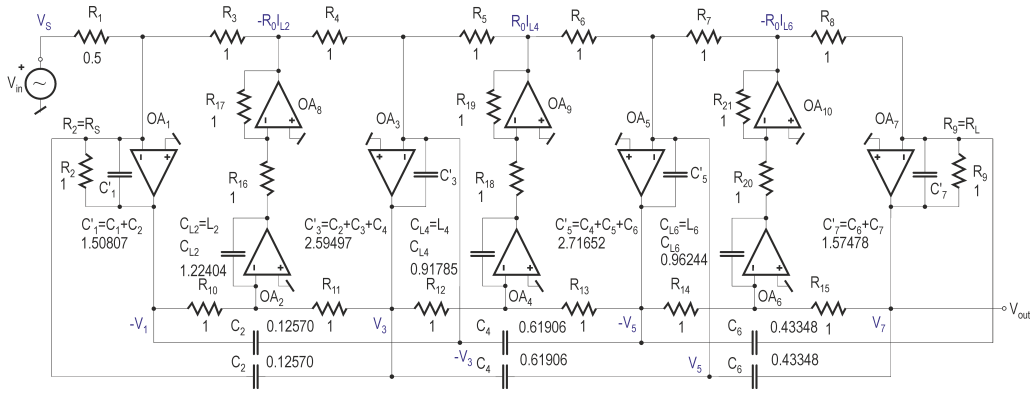


Figure 3. Final circuit with *additional capacitors* containing only positive components and positive and negative integrators (sfg is in Fig. 2(f)).

This results in the active RC circuit of Fig. 3, which is an inductorless active-RC simulation of the passive-LCR ladder network in Fig. 1(a). It is important to note that the low sensitivity to component tolerances, which are a characteristic of LCR ladder filters (see [2]), are carried over to their active-RC equivalents.

In this paper, all derivations will be carried out on single-ended designs; the conversion to a fully differential balanced version is straightforward and presented in [1].

B. Networks with Additional Resistors and Fewer Capacitors

As shown in [1], a third- to sixth-order elliptic filter has one or two parallel LC tank(s), which means *one* or *two* additional capacitor(s) C_2 (and C_4) when compared to the allpole filter of the same order. Seventh-order elliptic filter presented in this paper has *three* additional capacitors C_2, C_4 and C_6. According to the initial sfg in Fig. 4(a) [it is the same sfg as in Fig. 1(b)], which is shown again in Fig 4(c) (with only voltage nodes) the simulation of these capacitors resulted in six additional branches in the sfg with transmissions C_2/C_1' (from V_3 to V_1), C_2/C_3' (from V_1 to V_3), C_4/C_3' (from V_5 to V_3), C_4/C_5' (from V_3 to V_5), and C_6/C_5' (from V_7 to V_5), C_6/C_7' (from V_5 to V_7).

In order to derive the new circuit *with additional resistors and fewer capacitors*, we note that V_1, V_3, V_5 and V_7, are *output* nodes (because they are voltage outputs of opamps), therefore signal addition at these nodes is not possible. Thus, a sfg transformation must be performed in order to bring those signals to *input* nodes of opamps, where addition is possible. The sfg derivations necessary to do this are shown in consecutive steps in Fig. 4. These are the main contributions of this paper. The steps in Fig. 4(d)–(h) are typical sfg transformations i.e., node splitting, shifting a transmittance (shifting the termination point of an internal branch), and loop reduction (e.g. see [8] and [9]). In Fig. 4(i) a sfg with only negative integrator paths and adders having some incoming branches with negative gain is obtained. However, the corresponding active circuit would have some passive negative components. To eliminate these, additional nodes have to be multiplied by –1 as shown in Fig. 4(j)–(l), resulting in positive and negative integrators, and all-positive incoming branches. The final filter is shown in Fig. 5. Capacitor vales readily follow from (6), whereas resistor values are simply calculated as reciprocals of path transmission values in (7); both capacitor and resistor values are shown in Fig. 5 in normalized form.

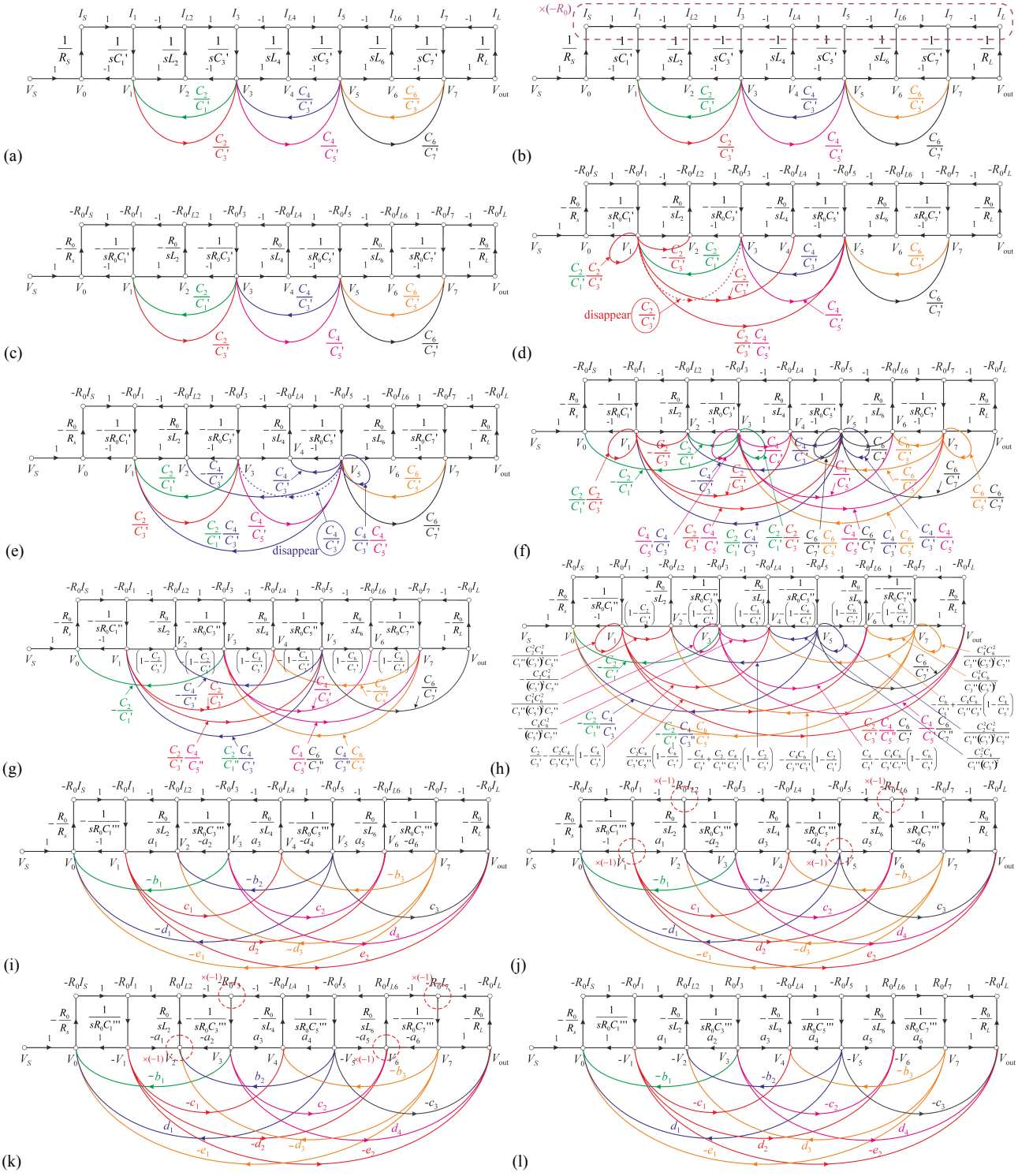


Figure 4. Deriving an active-RC simulation of a seventh-order passive-LCR ladder filter. (a) Initial graph (repeated graph from Fig. 1(b)). (b)–(c) Producing all voltage nodes; multiplication of current nodes by R_0. The reduction rules that apply follow: (d) Shifting the termination point of internal branches C_2/C_3' (the same is done with branches C_4/C_5' and C_6/C_7'). (e) Shifting the termination point of internal branches C_2/C_1' (and of C_4/C_3' and C_6/C_5' in the same way). The result of (d) and (e) is in (f). (g) Removal of self-loops and reduction of parallel branches. (h) Shifting the termination point of branches $C_2C_4/(C_1''C_3')$, $C_2C_4/(C_3'C_5'')$, $C_4C_6/(C_3''C_5')$, and $C_4C_6/(C_5'C_7'')$. (i) Removal of self-loops and reduction of parallel branches. (j)–(l) Multiply appropriate nodes by –1 in order to obtain only positive components with positive and negative integrators.

Capacitors in Fig. 4(l) are calculated from [use also (4)]:

$$C_1''=C_1'\left(1-\frac{C_2^2}{C_1'C_3'}\right),\quad C_3''=C_3'\left(1-\frac{C_2^2}{C_1'C_3'}-\frac{C_4^2}{C_3'C_5'}\right),$$

$$C_5''=C_5'\left(1-\frac{C_4^2}{C_3'C_5'}-\frac{C_6^2}{C_5'C_7'}\right),C_7''=C_7'\left(1-\frac{C_6^2}{C_5'C_7'}\right).$$

$$C_1'''=C_1''\left(1-\frac{C_2^2C_4^2}{C_1''C_3'^2C_5''}\right),C_3'''=C_3''\left(1-\frac{C_4^2C_6^2}{C_3''C_5'^2C_7''}\right),$$

$$C_5'''=C_5''\left(1-\frac{C_2^2C_4^2}{C_1''C_3'^2C_5''}\right),C_7'''=C_7''\left(1-\frac{C_4^2C_6^2}{C_3''C_5'^2C_7''}\right).\quad(6)$$

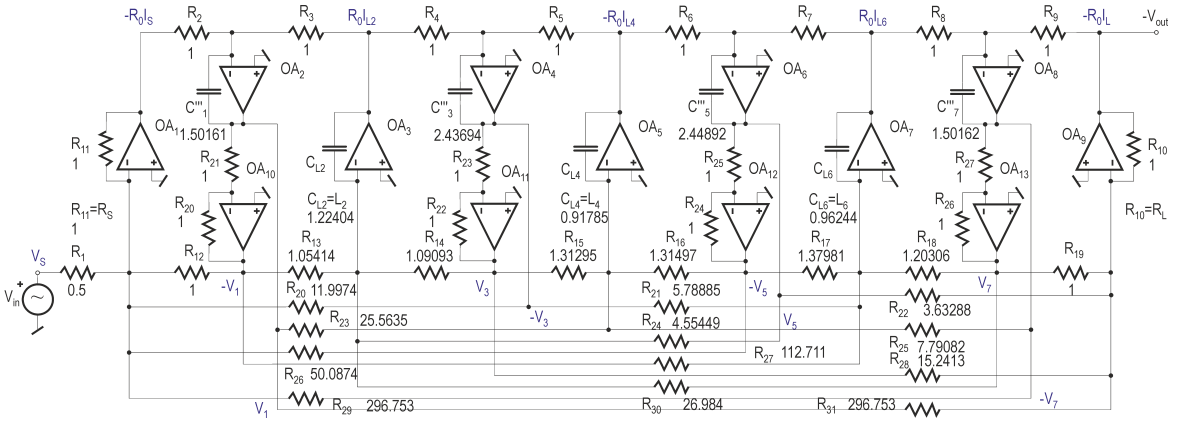


Figure 5. Final seventh-order active-RC ladder filter simulation with *minimum capacitors* (sfg is in Fig. 4(l)).

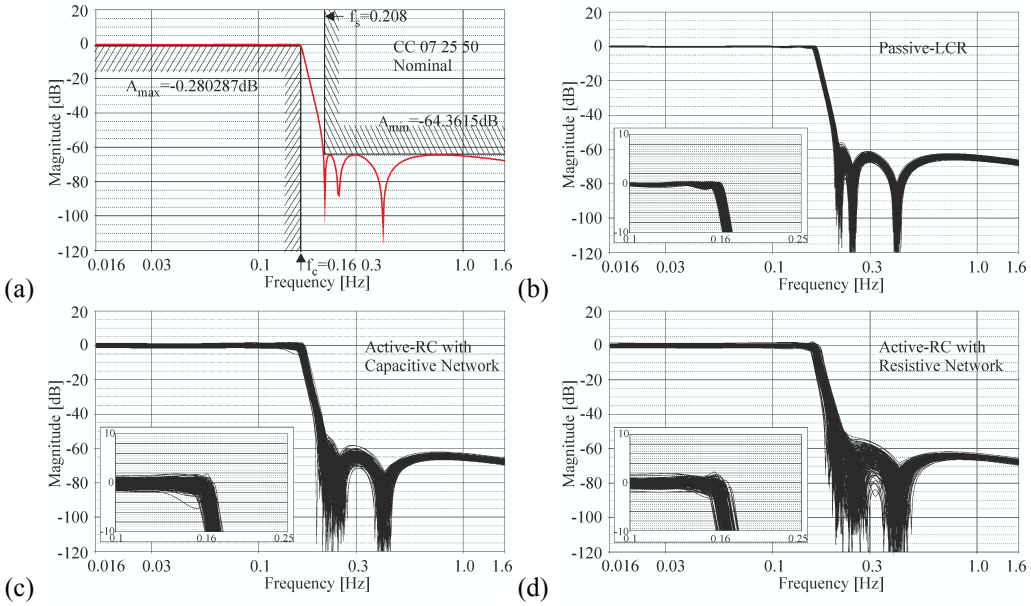


Figure 6. (a) Amplitude characteristics of a CC 07 25 50 filter. Monte Carlo runs of various realizations: (b) Passive ladder-LCR filter (Fig. 1(a)). (c) Active-RC simulation with additional capacitors (Fig. 3). (d) Active-RC simulation with additional resistors (Fig. 5).

Transmissions of paths in Fig. 4(l) are calculated from:

$$a_1 = 1 - \frac{C_2}{C_3'} - \frac{C_2 C_4^2}{C_3'^2 C_5''}, a_2 = 1 - \frac{C_2}{C_1'}, a_3 = 1 - \frac{C_4}{C_5'} - \frac{C_4 C_6^2}{C_5'^2 C_7''},$$

$$a_4 = 1 - \frac{C_4}{C_3'} - \frac{C_2^2 C_4}{C_1'' C_3'^2}, a_5 = 1 - \frac{C_6}{C_7'}, a_6 = 1 - \frac{C_6}{C_5'} - \frac{C_4^2 C_6}{C_3'' C_5'^2},$$

$$b_1 = 1 - \frac{C_2}{C_1'}, b_2 = \frac{C_4}{C_3'} - \frac{C_2 C_4}{C_1'' C_3'}\left(1 - \frac{C_2}{C_3'}\right),$$

$$b_3 = \frac{C_6}{C_5'} - \frac{C_4 C_6}{C_3'' C_5'}\left(1 - \frac{C_4}{C_5'}\right), c_1 = \frac{C_2}{C_3'} - \frac{C_2 C_4}{C_3' C_5''}\left(1 - \frac{C_4}{C_3'}\right),$$

$$c_2 = \frac{C_4}{C_5'} - \frac{C_4 C_6}{C_5' C_7''}\left(1 - \frac{C_6}{C_5'}\right), c_3 = \frac{C_6}{C_7'},$$

$$d_1 = \frac{C_2 C_4}{C_1'' C_3'}, d_2 = \frac{C_2 C_4}{C_3' C_5''}\left(1 - \frac{C_6}{C_7'}\right),$$

$$d_3 = \frac{C_4 C_6}{C_3'' C_5'}\left(1 - \frac{C_2}{C_1'}\right), d_4 = \frac{C_4 C_6}{C_5' C_7''},$$

$$e_1 = \frac{C_2 C_4 C_6}{C_1' C_3'' C_5'}, e_2 = \frac{C_2 C_4 C_6}{C_3' C_5'' C_7'}. \quad (7)$$

The conventional realization of the filter, with a capacitive network realizing the three tank capacitors as presented in Fig. 3, is not new (see [6] and [7]). The branches realizing the tank capacitors are obtained in the same way for all filter orders. On the other hand, the realization of finite zeros with resistive networks as presented in Fig. 5 is new, and requires a complicated sfg transformation leading to additional signal paths. Note also that new circuit has *more opamps and resistors* than the conventional one. This difference in complexity, i.e. in total number of passive and active components (but with canonic number of capacitors), increase with increasing order. In the latter circuit, large resistors such as R_{20}–R_{31} can be replaced by resistive T-networks with small values as shown in [1].

III. Sensitivity Analysis

In what follows, we examine the sensitivity to tolerances of passive component values of the original elliptic passive-LCR CC 07 25 50 filter whose amplitude characteristics are shown in Fig. 6(a) with that of the two active, simulated filters discussed above. Using the OrCAD PSpice 16 program [3] with Monte Carlo runs, we assume a zero-mean uniform distribution, and a 5%

standard deviation for all components. The obtained spread of responses for each filter is an indication of that filter's sensitivity to component tolerances. It is also an indication of the number of components in a given circuit; the fewer this number, the smaller will be the spread of responses resulting from the accumulated component tolerances.

The sensitivity of the passive-LCR filter (see Fig. 1(a)) is shown in Fig. 6(b). That of the active-RC simulation with a capacitive network (see Fig. 3) is shown in Fig. 6(c), and that of the active-RC simulation with a resistive network (see Fig. 5) in Fig. 6(d). Both active-RC filters have similarly low sensitivity, but larger sensitivity than the original passive-LCR filter (Fig. 6(b)). This is because they contain a larger number of components. As expected from Orchard's theorem [2], all three filters have low sensitivity in the pass band.

In this paper (as well as in [1]) the main contribution has been the derivation of low-sensitivity elliptic filters having canonic number of capacitors. Therefore, in all simulations using PSpice instead of real opamp models, ideal voltage-controlled-voltage sources (VCVS) denoted by "E" with frequency-independent gain set to 10^6 have been used.

IV. CONLUSION

In this paper (and in the companion paper [1]) we compare two realizations of active-RC elliptic ladder filters which are the inductorless equivalent of a given passive-LCR ladder filter. One is classical and well known; it features an additional capacitive network that realizes the series capacitors required to form parallel tanks. The second, which is new, uses additional resistive networks to replace the additional series capacitors of the first circuit. This reduces the total number of capacitors in the circuit. This reduction increases with the filter order.

In this paper it is shown that the realization of the newly proposed circuit becomes more complicated when the order is increased in terms of the increasing complexity of the sfg transformation. Consequently, the additional resistive network also becomes more complicated (see also [1]).

The sensitivity of the two circuits is compared, and it is concluded that no noticeable difference exists between them. Nevertheless, since the resistive network saves capacitors (which are generally more difficult to manufacture especially on-chip), the elliptic filter, with a resistive network for the realization of finite zeros, may be a very useful alternative to the conventional circuit.

ACKNOWLEDGMENT

This work has been financially supported by the bilateral project "*Integrated Systems for Analog and Mixed Signal Processing—Active-RC filters for ADSL Communication Systems*" between the Ministry of Science, Education and Sports of the Republic of Croatia and the Ministry of Science of the Republic of Monte Negro.

REFERENCES

[1] D. Jurisic, N. Mijat and G. S. Moschytz, "Inductorless elliptic filters with reduced number of capacitors using signal-flow graphs," in *Proceedings of the 22nd* Eurocon 2013*,* (Zagreb, Croatia), July 2013, in press.

[2] H. J. Orchard, "Inductorless filters," *Electron. Letters*, vol. 2, pp. 224-225, June 1966.

[3] PSpice Users Guide ver. 16.5, Cadence Design Systems, Inc. (Cadence), San Jose CA-USA, May 2011.

[4] A. I. Zverev, *Handbook of Filter Synthesis*, John Wiley and Sons, New York 1967.

[5] R. Saal, *Handbook of Filter Design*. Berlin: AEG-Telefunken, 1979.

[6] G. M, Jacobs, D. J. Allstot, R. W. Brodersen, P. R. Gray. "Design Techniques for MOS Switched Capacitor Ladder Filters," *IEEE Trans. Circuits and Syst.* vol. 25, pp. 1014–1021, 1978.

[7] M. Banu, Y. Tsividis, "An elliptic continuous-time CMOS filter with on-chip automatic tuning," *IEEE Journal of Solid-State Circuits*, vol. 20, pp. 1014–1021; 1985.

[8] S. J. Mason, "Feedback theory-Some properties of signal flow graphs," Proc. IRE, vol. 41, no. 9, pp. 1144-1156, Sept. 1953.

[9] S. J. Mason, "Feedback theory-Further properties of signal flow graphs," Proc. IRE, vol. 44, no. 7, pp. 920-926, July 1956.

Engineering Education: Where We Are And Where We Go from Here

B. Pejcinovic*
* Electrical and Computer Engineering Department, Portland State University, Portland, USA
brano@cecs.pdx.edu

Abstract - While MOOC-s (massively open on-line courses) are getting a lot of attention in the press, many other interesting developments have been happening in the field of engineering education. It can be argued that many of the problems, perceived or real, have had solutions for quite some time and that problem is not in the development of new, effective and efficient instructional techniques but that the problem lies in the implementation of the known techniques. This paper will attempt to briefly summarize the current state of engineering education in the US, perceived problems, and potential solutions. As an example of what can be accomplished by a switch to more research-based instructional strategies we will discuss our experience in introducing active learning technique in microwave circuit design course. It is hoped that this will in some small way contribute to the discussion on the direction of engineering education in Croatia.

I. INTRODUCTION

It seems that in the last 6 months it is impossible to avoid daily news regarding Massive Open Online Courses (MOOCs), including mainstream US press, see e.g. [1]. Arguably, many of the ideas behind MOOCs are not new and can be traced as far back as correspondence courses [2]. There are also full universities offering online education (including engineering degrees), such as DeVry University in the US. It is, therefore, somewhat surprising to see this much excitement. A debate is currently raging regarding the influence and ramifications of MOOCs, including their impact on university finances and pedagogy [3] [4].

The purpose of this paper, however, is not the discussion of MOOCs; instead, we will argue that we should use them as impetus and opportunity for rethinking and substantially changing our approaches to engineering education, including a possible inclusion of MOOCs into engineering curricula. As will be shown below, there is no dearth of excellent, research-based, proven instructional strategies that can be used in this endeavor. In a recent and very rigorous study [5] authors "… found increased student attendance, higher engagement, and **more than twice the learning** in the section taught using research-based instruction" (emphasis added). This was in spite of the fact that the control group was taught "… by an experienced highly rated instructor."

Despite all the research, long-term, sustained efforts at reforming engineering education in the US are still largely limited to a handful of universities and colleges. For many different reasons, engineering faculty seem to have ignored these developments but we believe that status-quo is quickly becoming unsustainable.

A. Barrier to Engineering Education Reform

Recent studies examined science and engineering faculty familiarity and use of various instructional strategies. For example, in [6] there is a list of 24 research-based instructional strategies (RBIS) in physics education (see references therein):

1. Active Learning Problem Sheets
2. Activity-Based Physics Tutorials
3. Context-Rich Problems
4. Cooperative Group Problem Solving
5. Experiment Problems
6. Interactive Lecture Demonstrations
7. Investigative Science Learning Environment
8. Just-In-Time Teaching
9. Modeling Physics
10. Open Source Physics
11. Open Source Tutorials
12. Overview, Case Study Physics
13. Peer Instruction
14. Physlets
15. Ranking Tasks
16. Real Time Physics for Scientific Thinking Labs
17. Scale-Up, Studio Physics
18. Socratic Dialog Inducing labs
19. Thinking Problems
20. TIPERS
21. Tutorials in Introductory Physics
22. Video-Based Labs
23. Workbook for Introductory Physics
24. Workshop Physics

Additional techniques are used in engineering and in [7] the following list is found (see references therein):

1. Active Learning
2. Think-Pair-Share
3. Concept Tests
4. Thinking-Aloud-Paired Problem Solving (TAPPS)
5. Cooperative Learning
6. Collaborative learning
7. Problem-Based-Learning (PBL)
8. Case-Based Teaching
9. Just-In-Time Teaching
10. Peer Instruction
11. Inquiry Learning
12. Service Learning

There are some obvious overlaps and several other RBIS could be added, such as Project-based learning and

Model Eliciting Activities. An in-depth discussion of six inductive teaching and learning methods can be found in [11]. Study reported in [7] concentrated on engineering while [6] dealt with physics instruction. However, they reached very similar conclusions. The main point is that almost all faculty have heard of at least one or more of these strategies. A very large percentage – 88% – of respondents in [6] reported using one or more of the listed strategies. However, about 1/3 of those who attempted to use RBIS gave up after trying them. Three main issues seem to be driving the faculty acceptance of RBIS:

1. time that it takes to develop them,
2. fear that their implementation would take away too much of the class time and, as a consequence, not all required material would be covered, and
3. questions about efficacy of various methods.

Despite item 3, it is clear that the problem of increasing effectiveness and efficiency of student learning has moved beyond the research stage. There are many well documented and effective instructional strategies, including over 30 listed above. Most faculty are familiar with some of them to some degree, but the barriers are now at the implementation stage. In other words, we have to find ways to implement these as efficiently and effectively as possible. This can take on many forms and variations which will be shaped by local environment and circumstances.

Normally, we would expect that many years would pass between any such call for reform (see [3]) and actual implementation. What is different now is the emergence of MOOCs as potential disruptive innovation. There is much hype surrounding their development in the last year but they could also be viewed as a necessary ingredient for change, i.e. they can help in "establishing a sense of urgency" which is the initial stage of any successful engineering education reform [8]. In a very short time we – engineering educators – will have to provide an answer regarding what effect MOOCs (if any) will have on our courses, curricula, and institutions. So far, MOOCs have been largely US-based but some well known European universities have recently joined them: Delft University of Technology and Ecole Polytechnique Federale de Lausanne, to name a few. It is very likely that MOOCs will become a global phenomenon.

How should engineering faculty respond to this development will vary across universities and countries. We believe that whatever the response there will have to be a significant change in our teaching strategies and that we will have to implement some of the well documented and recognized RBIS-s. We know many of the answers and it is time to implement them.

Reported below are results of one such attempt at implementation of active learning techniques in a two-quarter long (20 weeks) senior-level course sequence in Microwave Circuit Design.

II. Active learning in Microwave Circuit Design courses

We will use the following working definition of active learning: "any class activity that involves students in doing things and thinking about the things they are doing". This is very broad definition but it captures the essence of the approach: getting the students involved during the face-to-face time instead of listening to lectures. We have two microwave circuit design courses in our program and they cover topics ranging from passive circuits and devices in the first (reflections, matching, filters, combiners, couplers etc.) to active circuits and devices (LNAs, mixers, PA-s) in the second. Both courses rely on hands-on, lab-based exercises, simulation assignments and a culminating (group) project. Labs are done immediately following the lectures. Typical enrollment is 15-20 students with 2:1 ratio of undergraduates to graduates.

In 2012 we undertook a major revision of the courses with the following ideas as guidelines:

1. Doing should replace listening during face-to-face classroom time
2. Students should produce something tangible
3. Students should be engaged in their learning
4. Make immediate feedback a priority
5. Provide multiple ways to retrieve recently learned concepts
6. Push students into using higher cognitive functions, but
7. Provide appropriate scaffolding
8. Have multiple opportunities for design cycle: design → build → test → redesign

These are largely based on research findings summarized in [10].

We expect students to be able to:

- Design, build and test passive circuits (microstrip and SMD)
- Design, build and test active circuits (low-noise amplifier, mixer, power amplifier)
- Design circuits using simulation tools
- Measure real circuits at microwave frequencies and apply de-embedding and calibration
- Write good quality reports
- Read, comprehend and explain technical literature

We will first report on our experience with classroom response system from Learning Catalytics (https://learningcatalytics.com/) which can be readily implemented by any interested instructors.

A. *Classroom response system use*

For each lecture we looked into possible conceptual problems that students have exhibited in the past and tried to find appropriate questions and/or activities that would bring that up for discussion. One area of considerable difficulty is the concept of matching, reflections and how Smith chart is to be used. Figure 1 and Figure 2 show a composite of a question presented to students and student answers (which are not shared with students until after submission). It is immediately clear that several students are having difficulty and may even be completely lost. In such a case, we need to stop the lecture and:

a) have peer discussion, or
b) pose a different question, or
c) explain the material in more depth.

Some of the fundamental concepts can be tested in this way and class progress adjusted accordingly. We use a similar approach when practicing circuit design – instead of instructor demonstrating how it is to be done, students do it themselves. Designs are broken up into more manageable steps and students enter their results online.

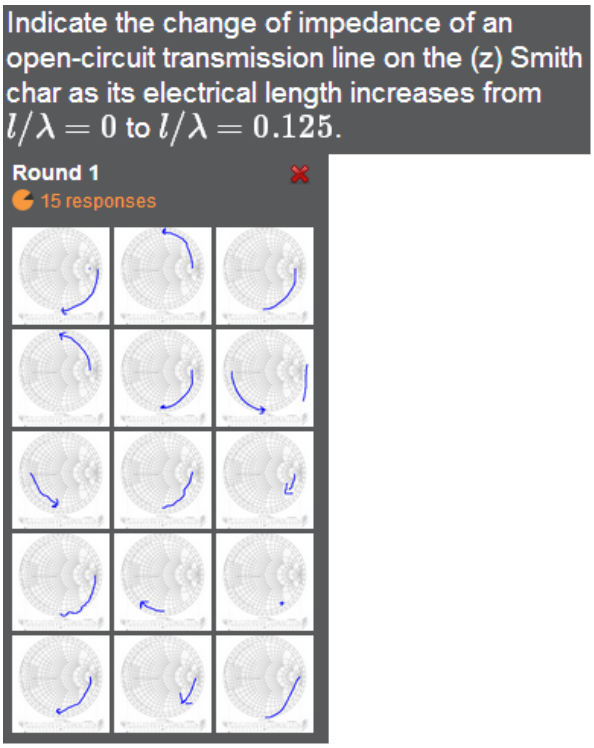


Figure 1. Example of in-class exercise testing students' conceptual understanding of transmission line behavior.

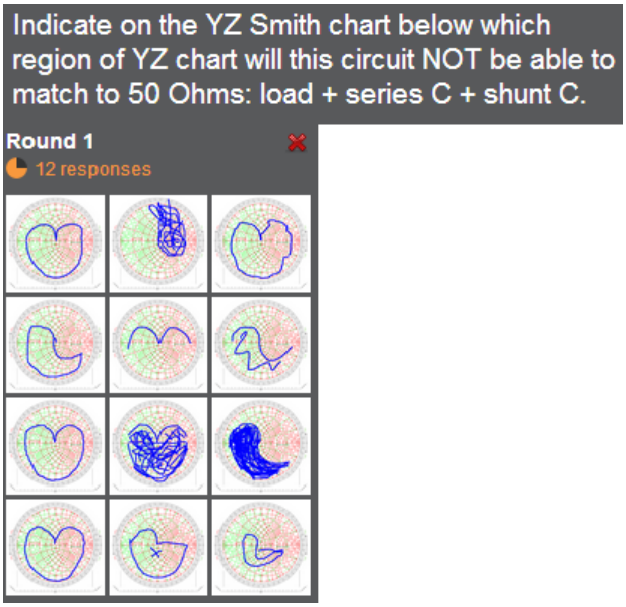


Figure 2. Another example of conceptual testing regarding impedance matching. Students have to provide graphical answer.

Some preliminary experiences regarding utilization of Learning Catalytics can be summarized:

Positive:

- Goes way beyond multiple choice questions
- Graphical and TeX capabilities
- Immediate formative feedback
- Available on wide range of devices (laptop, smartphones, tablets)
- Enables peer interaction and competition

Negative:

- Significant investment in time to build a "database" of questions or problems
- Inelegant handling of complex numbers
- It takes a couple of weeks for students to get used to it and regularly bring devices

B. Redesigned labs

Our intent was to reinforce student learning by having "lab" and "lecture" merge into one so that concepts can be immediately put to practice instead of waiting for assigned lab time. This means that as many mini-designs should be attempted during class/lab time so that instructor can provide immediate feedback.

One example is the first lab session: students measure surface-mount devices (SMD) and wire-lead components on LCR meter and we discuss how to deduce what the components are, what parasitics they exhibit and how to model them. This is followed up by a simulation assignment where students are asked to fit the model to measured data. We have eight weekly lab sessions covering microwave components, TDR response, measurement of characteristic impedance, transmission line design, matching circuits, VNA measurements and deembedding. What remains to be done is to integrate these into one common theme or project which would provide a larger objective for students to meet. While it is very labor intensive, observation of student performance in the labs is essential for understanding of their progress.

C. Self-efficacy Assessment

In addition to the usual items used for grading, such as exams, lab and project reports, at the end of each term we ran a survey with questions given below. The first four questions were designed to address research question on benefits of active learning. The rest of the questions addressed research question regarding the most appropriate and effective teaching technique. There were 14 responses out of 15 students after first term, and 8 out of 12 after the second term. Note that figures give raw numbers which will suffice for qualitative discussion below.

The first part of the survey consisted of the following questions:

"I am confident that I can (Scale: Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree):

1. Design microwave circuits
2. Build and test microwave circuits
3. Write good quality reports
4. Read and understand technical publications"

Results from Winter 2012 course are shown in Figure 3. In the follow on course in Spring 2012 all of the responses fell into strongly agree or agree category. We observed that students were less confident of their abilities

in the Writing and Reading categories. This will require more attention to these "soft" skills during the course.

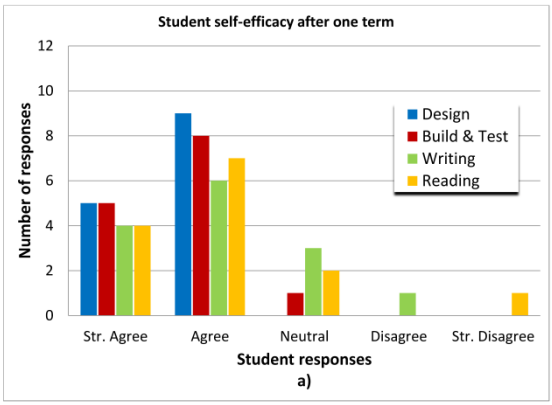


Figure 3. Student self-efficacy in designing circuits, building and testing, writing reports and reading technical literature after one term (Winter 2012).

D. Effectivness of instructional techniques

We also attempted to find out which of the various instructional techniques students find the most helpful in their learning. In the same survey they are asked the following:

"I found this (instructional) technique to be (Scale: Very helpful, Somewhat helpful, Neutral, Not helpful, Waste of time):

5. Building and testing circuits
6. Running circuit simulations
7. Doing in-class exercises and problems
8. Solving homework problems
9. Class project (filter design, simulation, build, test)
10. Listening to lectures in-class
11. Watching recorded videos of lectures (Echo 360)
12. Watching pre-recorded videos (special topics)
13. Working with teammates
14. Class project"

Results of the survey are presented in Figure 4.

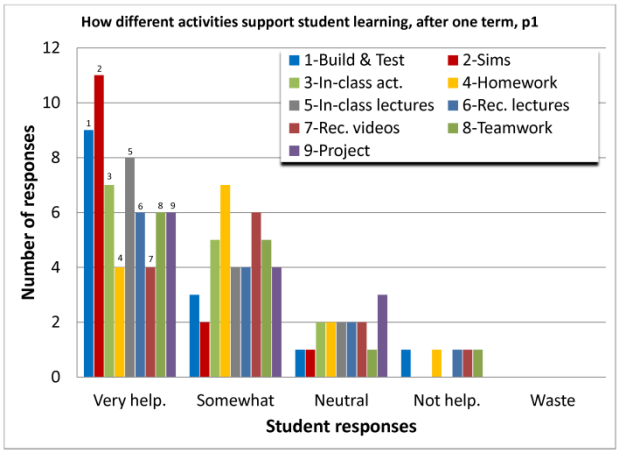


Figure 4. Student responses on of how well various class activities contributed to their learning after the first term (Winter 2012).

Clearly, "build and test" and, somewhat surprisingly, "simulations" were two biggest vote getters. This led us to modify the course so that simulations are introduced earlier. Building and testing was primarily done in the labs that followed lectures so these results look very encouraging for this approach to labs.

Homework was de-emphasized in the first run which can explain the relatively low score. However, we found that students need to spend more time getting fluent in some important tasks, such as utilizing Smith chart, which will require more of practice problems to be (re)introduced. Items related to lectures (3 and 5) scored very well and we know from 2013 mid-term evaluations that students find in-class activities to be most helpful in their learning. In order to improve the usefulness of the project we will have to make it more realistic and applied. A more complete set of data is presented in [9]. The sample size is not sufficient to draw firm conclusions yet or to generalize, but it is sufficient for course improvement. We will be collecting more data in 2013.

E. Mid-term course evaluations

Midway through the winter 2013 quarter students were asked to provide feedback on how well the course is going. For this survey students were divided into groups of 3 to 5 and were supposed to come up with consensus answers and recommendations. The only exception was the first question on intellectual challenge. The questions were:

1. Do you find the class intellectually stimulating (yes or no)? Why or why not? (majority vote)
2. What is helping you learn in this course? Explain. (consensus)
3. What aspects of the course detract from your learning? (consensus)
4. What changes or suggestions do you have for improving the class? (consensus)

Answers to the first question were unanimously yes. Active learning in class and in labs, use of Learning Catalytics and feedback were all mentioned as contributing factors to questions 1 and 2. For question 3 students mostly commented on the organization of material. For example, several textbooks and other materials were used in class which required students to spend considerable time reading it and prioritizing the information. This is valid criticism and we will organize the material better in the future. However, students do not realize that organizing their knowledge is helping their metacognition, i.e. it helps their understanding of their own learning. This is a very important component of student learning [10] but we should be careful not to turn the difficulty of knowledge organization into frustration. Finally, one common suggestion for course improvement was increasing the amount of lab time (currently two hours). We interpret this is an indication of students' real desire to learn, which would be a very positive sign.

III. CONLUSION

Early indications are that our experiment with introduction of active learning into lectures and redesign of labs has been very successful. This is demonstrated through student engagement and their performance. We

cannot do a pre- and post-analysis since this was the first time the course was taught by author after a long hiatus. None the less, as expected from the literature on this method, some dramatic improvements have been observed. For example, course evaluations are now significantly higher than the departmental average [1]. Through assessment we have been able to gain insight into which instructional methods or aids are most effective and what we need to improve.

This course redesign is only a very specific example of what can be accomplished, but many of the techniques used can be transferred to other courses. Eventually, we hope that in our EE program at PSU we will continue with even more ambitions reforms, such as introduction of project-based learning across the curriculum or introduction of flipped-classroom instruction. We believe that this process of experimentation, continuous improvement and implementation of research-based instructional strategies is essential for continued relevance of college or department based engineering education.

[1] We are in the middle of the process of changing departmental course evaluations as well.

REFERENCES

[1] T. Friedman, "The Professors' Big Stage" The Opinion Pages, The New York Times, March 5, 2013, http://www.nytimes.com/2013/03/06/opinion/friedman-the-professors-big-stage.html?ref=thomaslfriedman&_r=0

[2] Wikipedia, Massive Open Online Courses, Available online: http://en.wikipedia.org/wiki/Massive_open_online_course

[3] G.L. Mehaffy "Challenge and Change," Educause Review, pp. 25-42, September/October, 2012, Available online http://www.educause.edu/ero/article/challenge-and-change .

[4] R.J. Sternberg, "The End is Not Nigh for Colleges," The Chronicle of Higher Education, February 4, 2013, Available online http://chronicle.com.proxy.lib.pdx.edu/article/The-End-Is-Not-Nigh-for/136957/?cid=cr&utm_source=cr&utm_medium=en .

[5] L. Deslauriers, E. Schelew, C. Wieman "Improved Learning in a Large-Enrollment Physics Class," Science 332, 862 (2011); DOI: 10.1126/science.1201783

[6] C. Henderson, M. Dancy, and M. Niewiadomska-Bugaj, "Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process?," PHYSICAL REVIEW SPECIAL TOPICS - PHYSICS EDUCATION RESEARCH, vol. 8, 020104, 2012.

[7] M. Borrego, J. Froyd, M. Prince, C. Henderson "Faculty Use of Research Based Instructional Strategies," Proceedings of the 2011 AAEE Conference, Fremantle, Western Australia, 2011.

[8] R. Graham "Achieving excellence in engineering education: the ingredients of successful change," published by The Royal Academy of Engineering, London, ISBN 1-903496-83-7, March 2012.

[9] B. Pejcinovic "Application of active learning in microwave circuit design courses," in press, ASEE Annual Conference, Atlanta, USA, 2012 .

[10] S.A. Ambrose, M.W. Briges, M. DiPietro, M.C. Lovett, M.K. Norman, "How Learning Works, 7 Research-based Principles for Smart Teaching, John Wiley & Sons, 2010

[11] M.J. Prince and R.M. Felder, "Inductive Teaching and Learning Methods: Definitions, Comparisons, and Research Bases," J. Engr. Education, vol. 95, no. 2, pp. 123–138, 2006.

Concept for a SIL3 Middleware Encapsulating Safety-related Aspects of Applications for an 8051-based SIL3 Multi-Core System-on-Chip

Michael Schreiber, Emil Delic, Ali Hayek and Josef Börcsök
University of Kassel / Institute for Computer Architecture and System Programming, Kassel, Germany
michael.schreiber@uni-kassel.de, emil.delic@uni-kassel.de, ali.hayek@uni-kassel.de

Abstract – In this paper, a concept for a SIL3 middleware implementing safety-related aspects is proposed. The middleware is intended to be used by applications that are written for a recently developed safety system-on-chip. Each module implements a low-level driver. Each driver represents a specific functionality of the system-on-chip. Once being certified conforming to IEC 61508, the middleware would enable writing safety-related applications aimed at SIL3 almost as straightforward as non-safety-related applications. A multi-core SIL3 architecture for safety-related applications is explained. In addition, possible issues that can arise during the software development are identified. Furthermore, conformance arguments on meeting SIL3 are depicted.

I. INTRODUCTION

One of the major concerns in human society is time-constraints. Engineers, for example, are likely to having a tight schedule for accomplishing development. Nevertheless, developers of safety software applications are responsible for implementing safety requirements, e.g. conforming to IEC 61508-3 [1]. The norm provides a classification into safety integrity levels; from SIL1 (lowest) to SIL4 (highest).The norm provides guidelines on how to achieve an aimed SIL. For example, using formal methods is recommended for SIL3 but not for SIL1. The aim is avoiding and controlling systematic faults, i.e. mistakes remaining in the software that could lead to unintended behavior. The more complex the software, the higher the effort of applying recommended techniques and measures. To overcome this issue, a concept for a SIL3 middleware is proposed. The middleware is intended to encapsulating safety-related aspects. The overall aim is to allow writing safety-related applications almost as straightforward as non-safety-related software; thus, enabling safety application engineers to focus on the functional side of the application. The safety aspects are covered already by the middleware, as much as possible.

The proposed middleware is intended to be used on an 8051-based SIL3 multi-core System-on-Chip (SoC). The middleware would provide low-level drivers. Thus, using the middleware can also speed up application development. In order to achieve IEC 61508 certification, a safety application still needs to be developed conforming to IEC 61508. However, the effort for the certification process would be reduced, because the parts implemented in the middleware would already be certified. The safety application engineer would not need to worry about those parts. In [2] examples of safety-software products that have aimed a SIL certification but did not receive it are given. The reasons would be:

- Missing Requirements Specification
- Interaction between hardware and software is not good described
- Missing detailed descriptions and results from static code analysis
- Poor integration and unit testing
- Writing Acceptance Test Specification in the later stages of the software development

These concerns are to be considered at the very start of the middleware development to achieve IEC 61508 certification. It is very important to know how to correctly argue conformance to the IEC 61508 guidelines and procedures. To put it simple, you want to convince the certifying party, so you need to speak their lingo. Conformance arguments should suit to IEC 61508 requirements. Therefore, this paper presents a concept how to prove claims in the range of the certification process.

The middleware shall be written in a programming language which satisfies guidelines given in IEC 61508-7 (C.4.) [1]. Further, the programming language should allow efficient coding due to the target platform and memory restrictions on the SoC.

In the second section, the SIL3 multi-core SoC target platform will be depicted. In the third section, programming language and tools used for developing the middleware are explained. Furthermore, the concept for the SIL3 middleware is proposed. Essential parts of the middleware are presented as well as their interaction with other units of the SIL3 environment. Finally, in the last section, the planned middleware certification process is depicted.

II. MULTI-CORE ARCHITECTURE FOR SAFETY - RELATED APPLICATIONS

In this section, the safety-related System-on-Chip target platform is briefly explained. The chip has been developed at Institute for Computer Architecture and System Programming (ICAS) at the University of Kassel

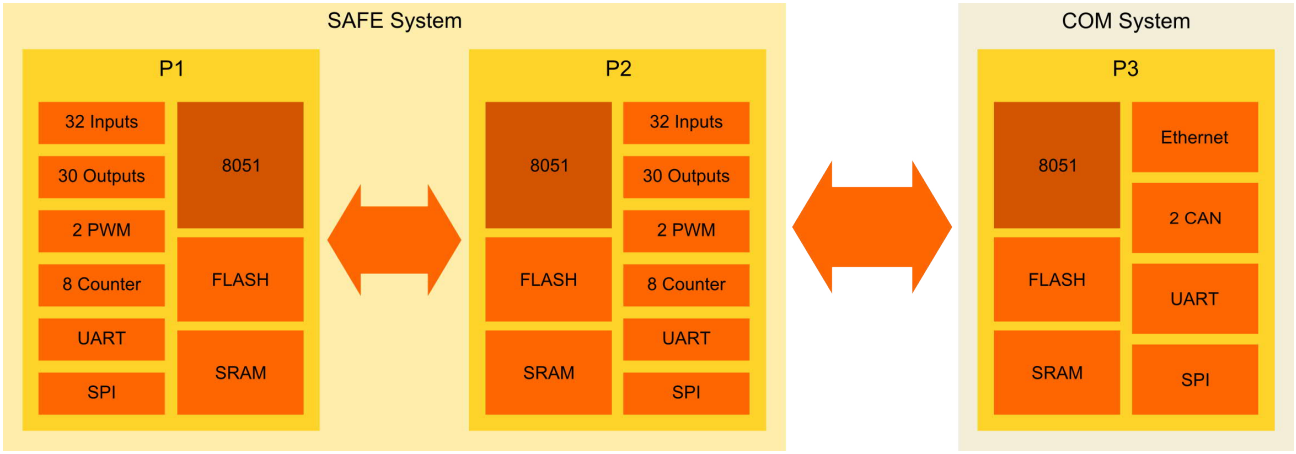


Figure 1. Multi-Core Safe System-on-Chip (SIL3) with Communication Single-Core System-on-Chip (not SIL3)

in collaboration with HIMA Paul Hildebrandt GmbH + Co KG. It is aimed at SIL3. The certification is in progress, currently. The concept has already been approved by TÜV Rheinland. The chip integrates all features of a programmable logic controller (PLC) into a single chip. For safety applications, it reduces number of components required and thus improves the system reliability by providing several on-chip functional safety features.

Fig. 1 shows an overview of the chip. It comprises a safe 1oo2D subsystem called SAFE System and a non-safe single-core subsystem called COM System. Both are connected only via a dual-ported RAM. The connection is interference-free. The COM System provides Ethernet and CAN interfaces. Using a so-called black channel approach, this non-safe part can serve as communication channel even for safety-related applications. It then would be up to the communication end-points implementing the safety layer, e.g. by using a safe protocol stack. This needs to be done by SW. Particularly, the operating system (OS) of HIMA – which we are using - already implements such a safe protocol stack. Both systems comprise 8051-based processor cores. These provide on-chip debugger units, multiple communication interfaces and data and program memories.

A. *Safety Features*

The SAFE System is a 1oo2D architecture with on-chip diagnosis. It also comprises:

- 32 Digital Inputs and 30 Digital Outputs (redundant)
- 8 Safe Counter Inputs
- 2 Safe pulse-width modulation (PWM) Outputs
- Redundant SPI Interface
- Timers, Interrupt Controller, and on-chip Debugger
- 256 Kbytes on-chip flash program memory for each channel
- 64 Kbytes on-chip SRAM data memory for each channel
- UART interface

B. *Communication Features*

The non-safe COM System is intended to realize communication between the chip and the external "world". The COM System can communicate with the SAFE System via a dual-ported RAM, thus establishing an interference-free connection of the safe and non-safe parts of the chip. The COM System comprises:

- 100 MBit/s Ethernet interface
- 2 CAN controller interfaces
- 512 Kbytes on-chip flash program memory
- 128 Kbytes on-chip SRAM data memory
- UART and SPI interfaces
- Timers, Interrupt Controller, Watchdog, and on-chip Debugger

III. PROGRAMMING LANGUAGE AND TOOLS

Safety software design starts with choosing a proper programming language and support tools. IEC 61508-3 states: "7.4.4.12: Programming languages for the development of all safety-related software shall be used according to a suitable programming language coding standard" [1]. Furthermore, Table A.3 and C-A.3 [1] highly recommend using strongly typed programming language with certified or proven-in-use tools. The language must be able to exclude language constructs (language subsets) that are error-prone or difficult to analyze. Examples on how to implement these requirements – in general, as well as for hardware design using VHDL, as used for designing the SIL3 target platform – are provided in [8].

The SIL3 middleware will be written using Embedded C++ which is a subset of C++ [3]. Some features have been omitted, such as multiple inheritance, runtime type identification (RTTI), exception handling. It is using less space and has better speed efficiencies than the Standard C++ language [3].

Fig. 2 shows guidelines provided in the IEC 61508 (claims) and their counterparts implemented in the SIL3 middleware (evidences). Further, fig. 2 shows how arguing conformance may be achieved. For example, all

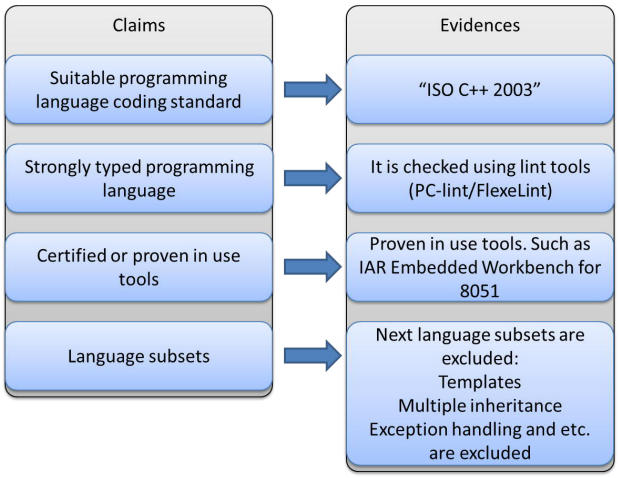


Figure 2. Guidelines given in IEC 61508 and corresponding implementations in the SIL3 middleware.

highly recommended procedures in IEC 61508 should have corresponding evidences in the SIL3 middleware.

Source code written in Embedded C++ will be checked according to [4] and [5]. The aim is to avoid systematic faults beforehand. Checking the code manually would be way too time-consuming and error-prone. Thus, PC-lint/FlexeLint will be used for this purpose. As compiler, IAR C/C++ Compiler for the 8051 Microcontroller Architecture will be used. The compiler is not certified for SIL3. However, it can be considered as proven in use. This facilitates the certification process. Otherwise, proving that it is reasonable and safe using the compiler for safety-related software development would be required. The PC-lint software can be used inside IAR Embedded Workbench.

The C/C++ compilers of IAR Systems provide an extended implementation of the Embedded C++, Extended Embedded C++. It allows: namespaces, cast syntax (static_cast and others but they will not be used), virtual functions and inline functions. Those features are omitted in Embedded C++. The middleware needs to use these features. Therefore, Extended Embedded C++ is chosen for writing the middleware.

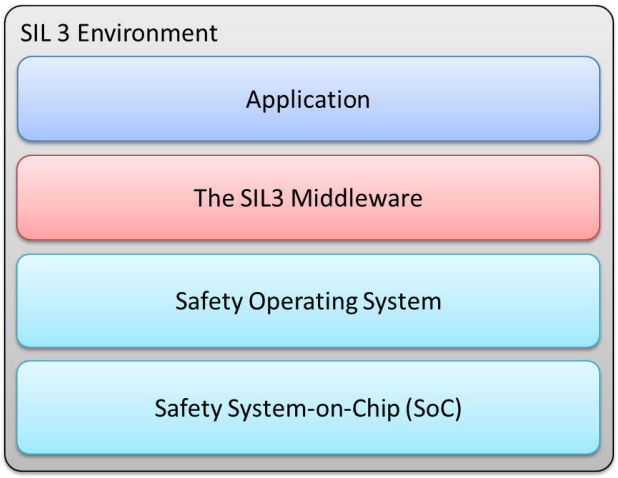


Figure 3. An example of using the SIL3 Middleware

Using IAR's Extended Embedded C++ prohibits using some standard C++ language features. Therefore, the middleware source code cannot make use of:

- Templates
- Multiple inheritance
- Exception handling
- Runtime type information (RTTI)
- reinterpret_cast, and const_cast
- The attribute "mutable".

All claims and evidences are documented for each software lifecycle phase. Those documents will be used by the certifying party. Well-written and structured arguing documents would simplify the certification process. Conformance arguments should be formulated explicit, rigorous, and structured, as explained in [6]. The middleware arguing documents will use arguing templates introduced in [6]. Also, remarks given in [2] will be taken into account.

IV. THE SIL3 MIDDLEWARE

The SIL3 middleware shall comprise all safety-related functionality of the target platform. It is planned to build up the middleware following a modular approach. Each module would represent a single functionality of the target platform. A library would contain all modules, made using IAR XLIB Librarian and XAR Library Builder [7]. This helps saving resources. Modules are linked only if used by the application. Thus, depending on which modules the application requires its footprint will be smaller. This is a very important capability, since, the SIL3 middleware will be used on an 8051-based, hence 8-bit, CPU core.

The middleware is planned to comprise the following low-level drivers:

- IODevice for general purpose, digital input and output ports;
- PWMDevice for pulse-width modulation outputs;
- CounterDevice for counter inputs;
- SPIDevice for SPI communication;
- TimerDevice module responsible for the implementing timers.

A. An example of using the SIL3 middleware

An example of using the middleware is shown on fig. 3. The middleware represents only a part of a safety-related application. Other parts would be the application code itself, the safety operating system and the SIL3 SoC as target platform. Those four parts represent a safe SIL3 platform. Safety application software developers need to take care only for getting their application code certified. The other three parts are certificated already. There may also be external components, such as sensors and actuators which are not regarded, here.

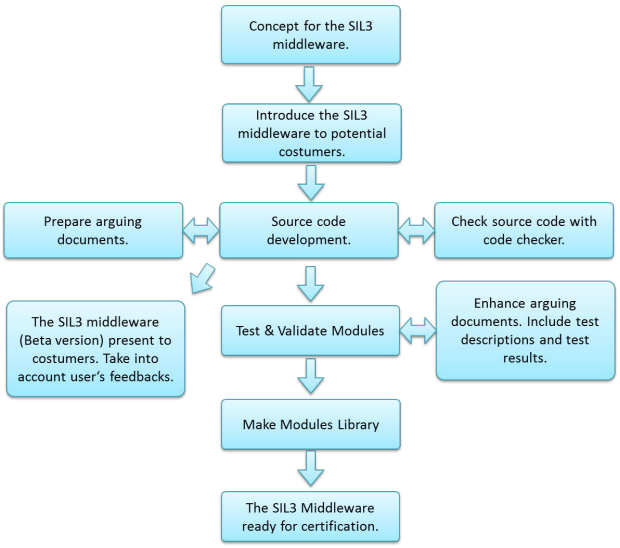


Figure 4. Certification plan

B. *Certification plan*

The first step is making a concept for the middleware, as depicted in fig. 4. That means a list of modules and their descriptions shall be prepared. Second, the concept would be introduced to the costumer. According to the customer feedbacks, the concept would be modified. Next, the source code development would start. Meanwhile, arguing documents are developed. The source code would be checked with the code checker PC-lint. Finished modules would be verified, i.e. tested. Verified modules would be presented to customers. This allows testing the middleware in rapid prototypes, enabling fast validation and ensuring that the customer gets what is needed for the respective application. Incorporating those feedbacks and finishing writing source code marks the beginning of the next phase: testing. Tests descriptions and results are added to arguing documents. After testing, all modules are combined in one library using IAR XLIB Librarian and XAR Library Builder. Thus, finally, the middleware, with well-prepared arguing documents, would be ready for the SIL3 certification process.

V. CONCLUSION

In this paper, the concept of a SIL3 middleware for an 8051-based SIL3 multi-core system-on-chip has been proposed. The multi-core SIL3 architecture for safety-related applications has been explained. The middleware would conform to the IEC 61508. Special care has been taken to identifying possible complications that could arise in the later stages of the software development. Such as:

- Missing Requirements Specification
- Interaction between hardware and software is not good described
- Missing detailed descriptions and results from static code analysis
- Poor integration and unit testing
- Writing Acceptance Test Specification in the later stages of the software development

Arguing conformance is considered as a key task. Thus, an example of writing the arguing conformance document has been given. It has been shown that IAR's Extended Embedded C++ satisfies IEC-61508 recommendations. The concept and structure of the middleware have been defined. Principal parts of the middleware are presented and their interactions with other parts of the SIL3 environment are shown. In the end, the Middleware's certification plan is provided.

REFERENCES

[1] IEC/EN 61508 Ed.2.0, "International Standard: 61508 Functional safety of electrical/electronic/programmable electronic safety-related systems Part 1-7," International Electrotechnical Commission, IEC, Geneva, 1999-2010

[2] M. H. Lloyd and P J Reeve, "IEC 61508 and IEC 61511 assessments - some lessons learned," 4th IET International Conference on Systems Safety 2009. Incorporating the SaRS Annual Conference.

[3] P.J. Plauger, "Embedded C++: An Overview," Embedded Systems Programming, December 1997,

[4] MIRA Limited, "MISRA C++:2008 Guidelines for the use of the C++ language in critical systems," MIRA Limited, 2008.

[5] ISO/IEC, "ISO/IEC 14882: Programming languages — C++," American National Standards Institute, 2003.

[6] P. Graydon, I. Habli, R. Hawkins, T. Kelly, J. Knight, "Arguing Conformance," IEEE Software, May-June 2012, pp.50-57.

[7] IAR System, "The IAR Linker and Library Tools Reference Guide," Ver 4.61, January 2009.

[8] A. Hayek, M. Schreiber, and J. Börcsök, "Basic VHDL Tests Conforming to IEC 61508," in IEEE Proceedings of Seventh International Conference on Networked Sensing Systems, INSS 2010, pp. 41–44.

[9] E. Delic, M. Schreiber, A. Hayek, J. Börcsök, "Validation of a SIL3 Middleware for Safety-related System-on-Chips," MIPRO 2013 – 36th International Convention: MEET - Microelectronics, Electronics and Electronic Technology, in press, 2013, Opatja, Croatia

[10] J. Börcsök, "Electronic Safety Systems. Hardware Concepts, Models, and Calculations," ISBN-13: 978-3778529447, Hütig Verlag, 2004, Heidelberg, Germany

[11] J. Börcsök, "Functional Safety: Basic Principles of Safety-Related Systems," ISBN-13: 978-3778529867, Hütig Verlag, 2007, Heidelberg, Germany

Validation of a SIL3 Middleware for Safety-related System-on-Chips

Emil Delic, Michael Schreiber, Ali Hayek and Josef Börcsök
University of Kassel / Institute for Computer Architecture and System Programming, Kassel, Germany
emil.delic@uni-kassel.de, michael.schreiber@uni-kassel.de, ali.hayek@uni-kassel.de

Abstract - The aim is reducing IEC 61508 certification effort for safety-related applications. Therefore, an object-oriented middleware has been implemented which, once being SIL3-certified, would allow transferring safety-related aspects from the application into an already-certified code layer. Safety-related applications using this middleware then could be written as straightforward as non-safety-related code. The SIL3 middleware is designed for an 8051-based microcontroller. Because of the 8-bit CPU architecture, several limitations have been encountered. These issues are explained in this paper, and solutions are proposed. The middleware will be used on the multi-more safety chip. SIL3 certified multi-core architecture for safety-related applications is described. Also, the SIL3 middleware architecture is given. The middleware is modular. Therefore, if the user does not want to use some modules they will not be included and code footprint will be smaller. In the SIL3 middleware numerous design patterns can be found. Patterns description and purpose are explained. The pseudo code for the singleton pattern is shown. Moreover, module implementation is depicted. The validation of the middleware is presented and the V-model for the middleware is provided. It conforms to the IEC 61508.

I. INTRODUCTION

Developers of safety software applications usually are in the unfair position. Applications developing cycle is becoming shorter and shorter. Additionally, developers of safety software applications should implement safety requirements, e.g. conforming to IEC 61508-3 [1]. To help them, a concept for a SIL3 middleware is proposed. The middleware is intended to encapsulating safety-related aspects. The aim of the SIL3 middleware is to reduce IEC 61508 certification effort for safety-related applications. Consider a company that would like to use the SIL3 SoC in their application. Developing new software for a safety-related application requires some effort. The software would need to be developed following state of the art as well as applicable norms, e.g. IEC 61508. Having such a software certified for SIL3 includes avoiding systematic faults in every software unit – and to prove having taken suitable techniques and measures, as recommended by IEC 61508. Furthermore, the middleware is implemented using MISRA C++ 2008 to avoid systematic errors. The middleware reduces implementation and certification effort by providing low-level drivers and safety measures, as much as possible, which already would be certified for SIL3. As a result, the middleware would allow writing safety-related applications as straightforward as non-safety-related software.

The second section describes the target platform. The SIL3 middleware shall be used on a safety-related System-on-Chip (SoC) developed by our Institute and HIMA Paul Hildebrandt GmbH + Co KG.

In the third section, the SIL3 middleware is depicted. Also, modules and their functions are specified. In the SIL3 middleware, numerous design patterns can be found.

The fourth section is highly influenced by the book "Design Patterns: Elements of Reusable Object-Oriented Software" written by "Gang of Four" [2]. The middleware is aimed at SIL3. Therefore, all pattern implementations used in the middleware should conform to the IEC 61508 as well. Using a modular approach and well known design patterns would enhance software quality and facilitate avoiding systematic faults.

The fifth section shows the implementation of the SIL3 middleware. Interactions between different software design patterns are depicted. Furthermore, the typical usage scenario of the middleware is shown as a sequence diagram. The last section deals with the validation of the SIL3 middleware.

II. MULTI-CORE ARCHITECTURE FOR SAFETY-RELATED APPLICATIONS

In this section, the target platform is briefly explained. The safety chip has been developed at Institute for Computer Architecture and System Programming (ICAS) at the University of Kassel in collaboration with HIMA Paul Hildebrandt GmbH + Co KG. It is aimed at SIL3. The certification is in progress, currently. The concept has already been approved by TÜV Rheinland. The chip integrates all features of a complete programmable logic controller (PLC) into a single chip. Using the SoC rather than a traditional PLC can enhance system reliability: by providing several functional safety features on-chip, thus reducing the number of components required. It comprises a safety-related 1oo2D subsystem called SAFE System and a non-safe single-core subsystem called COM System. Both are connected only via a dual-ported RAM. The connection is interference-free. The COM System provides Ethernet and CAN interfaces. Using a so-called black channel approach, this non-safe part can serve as communication channel even for safety-related applications. It then would be up to the communication end-points implementing the safety layer, e.g. by using a safe protocol stack. This needs to be done by SW.

Particularly, the operating system (OS) of HIMA – which we are using - already implements such a safe protocol stack.

Both systems comprise 8051-based processor cores. These provide on-chip debugger units, multiple communication interfaces and data and program memories. The two DP80390 CPU cores are used. DP80390 is high-level performance 8-bit microcontroller. Its core is compatible with the standard 8051 8-bit microcontrollers. DP80390 has Pipelined RISC architecture. Communication single-core has one DP80390 CPU. The non-safe COM System is intended to realize communication between the chip and the external "world".

III. The SIL3 Middleware Modules

The SIL3 middleware comprises all safety-related functionality of the target platform. The middleware is modular. Each module represents a single functionality of the target platform. Fig.1 lists all of these modules. The middleware interacts with the SIL3 multi-core SoC target platform and encapsulates all low-level access. So, a module can be regarded as a low-level driver. A library contains all modules, made using IAR XLIB Librarian and XAR Library Builder [3]. This helps saving resources. Modules are linked only if used by the application. Thus, depending on which modules the application requires its footprint will be smaller. This is a very important capability, since, the SIL3 middleware will be used on an 8051-based, hence 8-bit, CPU core.

IV. The SIL3 Middleware Design Patterns

Traditional software design patterns used for the middleware can be considered as well-defined and well-known. They have been widely used for several years, already. Therefore, they can be reasonably regarded as proven-in-use design methodologies. In this section, used patterns are explained. The middleware is aimed at SIL3.

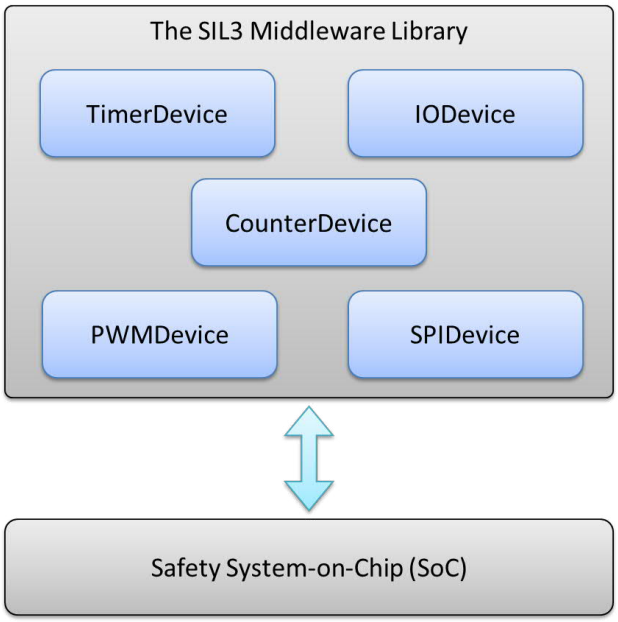


Figure 1. The SIL3 middleware modules

Therefore, the implementation of the used design patterns needs to conform to SIL3 as well. Thus, all safety integrity requirements need to be implemented, as specified by IEC 61508.

This section is highly influenced by [2]. This book describes accurate and appropriate programming solutions for common problems that can arise during software development. Those solutions will be used in the middleware. As shown in the previous section, the SIL3 middleware consists of modules. In addition, the middleware source code is divided into sections regarding the implemented design pattern. This facilitates writing and preparing documentation for the SIL3 certification; *"divide and conquer"*.

A. Singleton pattern for the SIL3 middleware drivers

In some cases, the user may want to have only one object of a specific class. Having different instances for the same SPI interface driver, for example, would be pointless. Every object of the driver class would be able to manage the same set of registers and the same physical interface at the same time. Thus, the risk of misuses can be very high. The singleton pattern allows having only one instance of the class. The middleware implements the singleton pattern for each driver module. For example, there is only one driver for all general purpose digital inputs and outputs (IOs), the "IODevice". This reduces the risk of having concurrent access to the same IOs. The singleton class is shown on the Fig 2. The following code shows a class definition making use of the singleton pattern. The code originates from [2].

```
class Singleton
{
public:
    static void Register( const char* name,
                          Singleton*);
    static Singleton* Instance();
protected:
    static Singleton* Lookup( const char*
name);
private:
    static Singleton* _instance;
    static List<NameSingletonPair>*
_registry;
};
```

The singleton class needs to register the singleton when it is instanced. The class shall have only one single instance. The exemplary class above does not prohibit creating objects via copy constructor and assignment operator. This class also has the issue that memory will be left allocated upon ending of the program, *memory leakage*. Thus, such an implementation cannot be used by

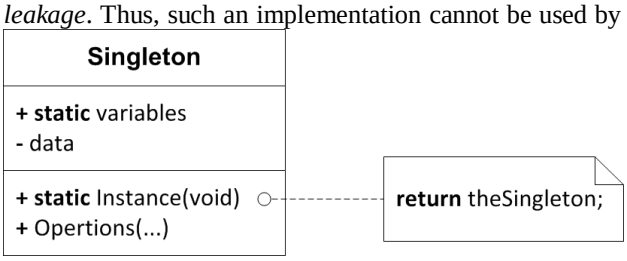


Figure 2. The Singleton class (UML class diagram)

the SIL3 middleware: It does not facilitate avoiding systematic faults reasonably. Therefore, it is unlikely to be in conformance with SIL3. Furthermore, it is not in conformance with the section C.2.6.6 Limited use of pointer "*To avoid the problems caused by accessing data without first checking range and type of the pointer"* [1]. Hence, it is highly recommended to avoid overuse of the pointers. Considering all mentioned facts, the Modified Meyers singleton will be used rather than the singleton pattern proposed in [2].

B. Modified Meyers singleton

The singleton implementation shall be pointer-free. Also, it should prohibit creating class objects using copy constructor and assignment operator. In [4] such a singleton is proposed. The first step is blocking object creation by constructors and assignment operator by declaring them as private within the class declaration. In addition, the singleton description and implementation must be in the middleware namespace. The middleware namespace combines its classes under a unique name. So, the middleware will not be defined within the global scope but in a separate scope, as demanded by MISRA C++ 2008 [5] . A class description implementing the Modified Meyers singleton would be as follows:

```
class Singleton : public Adapter
{
public:
  static Singleton& Instance(void);
private:
  Singleton(void);
  Singleton(Singleton const& );
  Singleton& operator = (Singleton const&
);
  ~Singleton();
}
```

Like discussed previously, all parts that are responsible for creating objects are defined in the private scope of the class. The method `Instance()` is responsible for creating the single object of the singleton class. Two important details of that method are worth to mention. First, singleton class object must be static. Static objects in a function are created during the first function call. The method `Instance()` uses this fact. Second, upon the first call of `Instance()`, the singleton is registered at the operating system (OS). In other words: The driver instance says *"Hello OS, the application wants to make use of me, so please remember keeping me up-to-date, thank you very much!"* To achieve this, the middleware uses the iterator design pattern. Moreover, the singleton class also implements the adapter pattern which is described in the following section.

C. Adapter pattern for the SIL3 middleware modules

The structural design pattern *"adapter"* is described in [2]. This pattern has the ability to access all middleware objects. The adapter pattern provides, as mentioned in [2]:

- Same interface for the all middleware drivers;
- Reusable classes;

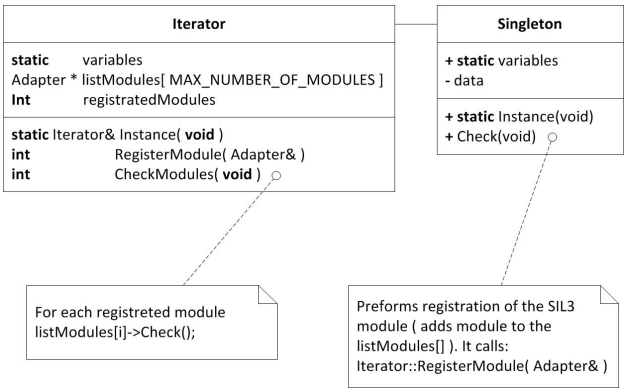


Figure 3. The iterator class and singleton class

The adapter pattern can be implemented as follows:

```
class Adapter
{
 public:
    virtual pl_type Check() = 0;
 protected:
    virtual ~Adapter(){};
};
```

The adapter class does not have any implementation. It is used as the interface to the middleware modules. The module classes are derived from the adapter class.

D. Iterator patern for the SIL3 middleware modules

The behavioral pattern *"iterator"* is described in [2]. The iterator pattern is used to performing the registration of the modules at the OS and to execute actions which are the common for all modules. A good candidate is the method that checks internal state of the modules. The iterator would call this method for the all modules. A class diagram is shown in fig. 3.

E. Facade patern for the SIL3 middleware modules

The SIL3 middleware is not a standalone application. It needs being able to communicate with other applications such as safety operating system and user application. The middleware is only one part of the Safety-Critical architecture. Other parts are: Safety

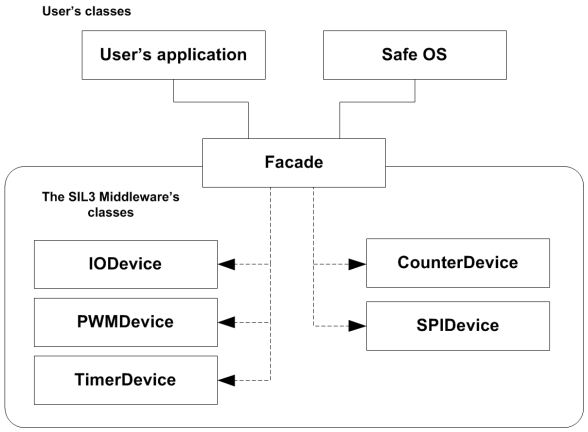


Figure 4. The Facade pattern is used to implement the SIL3 middlewares drivers

System-on-Chip, safe operating system and application code. Those four components would provide a base platform for SIL3 applications.

The SIL3 middleware interfaces the application and the safe operating system. The facade pattern can be implemented to providing a plain interface to the middleware. The middleware would be reusable and changes would be easier to implement. The facade acts as entry point for the middleware. Fig. 4 shows a basic representation of the facade pattern.

V. THE SIL3 MIDDLEWARE IMPLEMENTATION: OVERVIEW

In the previous section, the design patterns that are used for writing the middleware source code have been shown and explained. In this section, the implementation of the middleware is explained. Four different design patterns are used: singleton pattern, adapter pattern, iterator pattern and facade pattern. Modules of the middleware are coded with the singleton pattern. The module class attributes and methods imitate roles of the corresponding function of the safety-chip. The middleware uses the adapter pattern to offer the same interface for the all modules. The module classes are derived from the adapter class which is abstract. The pure virtual member function `Check()` in the adapter class is implemented in each module as a concrete member function. Additionally, the adapter has an array of pointers used by the iterator class to register modules used by the application. Additionally, the second iterator method performs common task for each registered module. The middleware needs only one iterator. Consequently, the iterator class is also a singleton.

The SIL3 middleware will be used by the safe operating system and the application. Application developers need not to worry about implementing low-level access to the target platform. The SIL3 middleware encapsulates all functionality implemented on the safety-chip and covers all low-level access. Hence, using the middleware can speed up development and ease the IEC 61508 certification process. The facade class represents the window that will be used for communications with middleware modules. The facade class, used as interface between application and middleware, provides a `Tick()` method. This method can be used to get all driver instances needed by the application. The facade class itself is instantiated within the safe OS automatically. The middleware design patterns and their interactions are depicted on fig. 5.

A. *Sequence diagram of the SIL3 middleware*

The sequence diagram of the middleware is shown in fig. 6. Object interactions are shown. The instantiation of middleware objects starts when the safe OS is booted. The facade class method `Instance()` is responsible for creating the iterator class object; again by calling the (iterator class) method `Instance()`.

In the first message, the application wants to check the internal states of the modules. It calls an appropriate method defined in the facade class. The facade method would delegate the request to the iterator. Because, none of the modules are created, yet, the iterator method would return immediately. Return values signalize that no modules are registered. Next, modules are created by calling `Instance()` of the singleton class in the `Tick()` method of the facade class. The message *"msg B"* demonstrates that: First, the `Tick()` method calls `Instance()` method of the singleton class. Then, this method performs the registration of the module. Pointers on the registered object are stored in the array of the iterator class. In the next step, under *"msg C"*, the application again calls the method for checking internal states of the modules. But this time, at least one module is

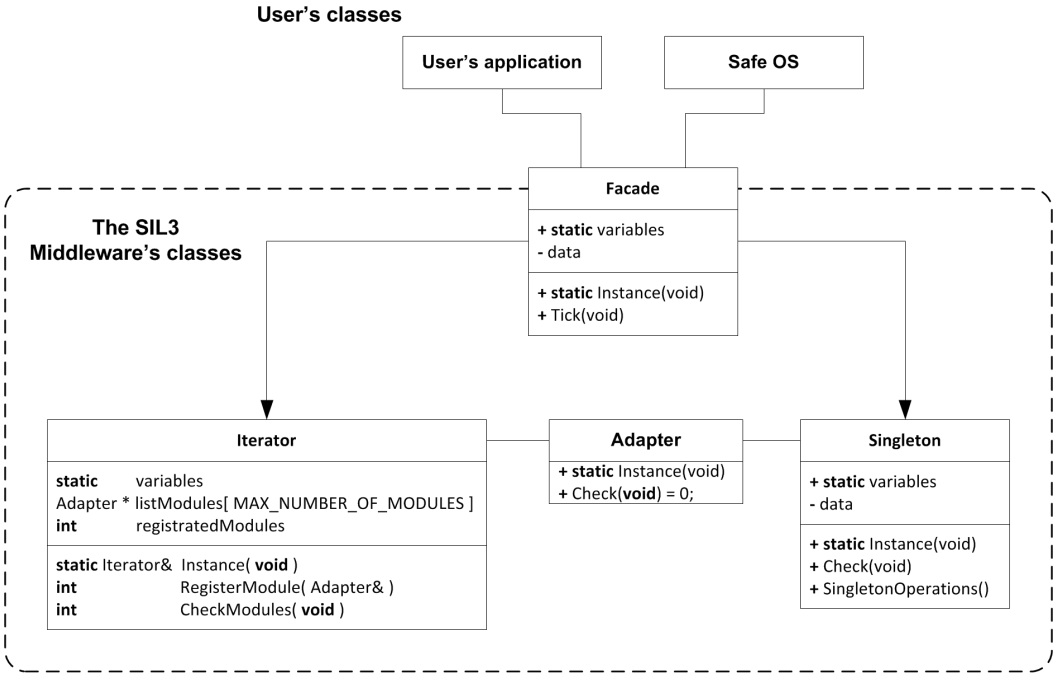


Figure 5. The SIL3 design patterns – Overview class diagram

registered and those modules perform internal checks. An appropriate message is delegated to the Facade object and then to the User application. The `Tick()` method of the Facade class is also able to call methods of registered modules, as demonstrated in *"msg D"* and *"msg E"*. An example would be calling the method of the IODevice driver that performs a safe change of a digital output port.

VI. V-MODEL OF THE SIL3 MIDDLEWARE

The V-model is used for developing the middleware. It enhances project clarity and project administration as stated in [4]. The V-Model is standardized. Fig. 7 shows the V-Model as used for the middleware development. It conforms to the "development lifecycle" shown in [1]. The objective has been to divide development of the software into well suited phases.

According to IEC 61508-3, the first step is to specify the safety integrity requirements and to determine the required SIL. Annex A of the IEC 61508-3 provides guidelines to select measures and techniques suitable for the aimed SIL. In the following step, the software architecture is determined. Conforming to [1], tools, programming language and compilers are selected in the software system design phase. The middleware modules are designed in the fourth step. The modules must satisfy requirements required by safety integrity level, SIL3. Dividing the middleware into modules facilitates future modifications and adaptations to different platforms. See also the explanations and guidelines given in section "7.4.5 Requirements for detailed design and development: Software System Design" [1]. After module design, code is implemented regarding section "7.4.6 Requirements for code implementation" in [1].

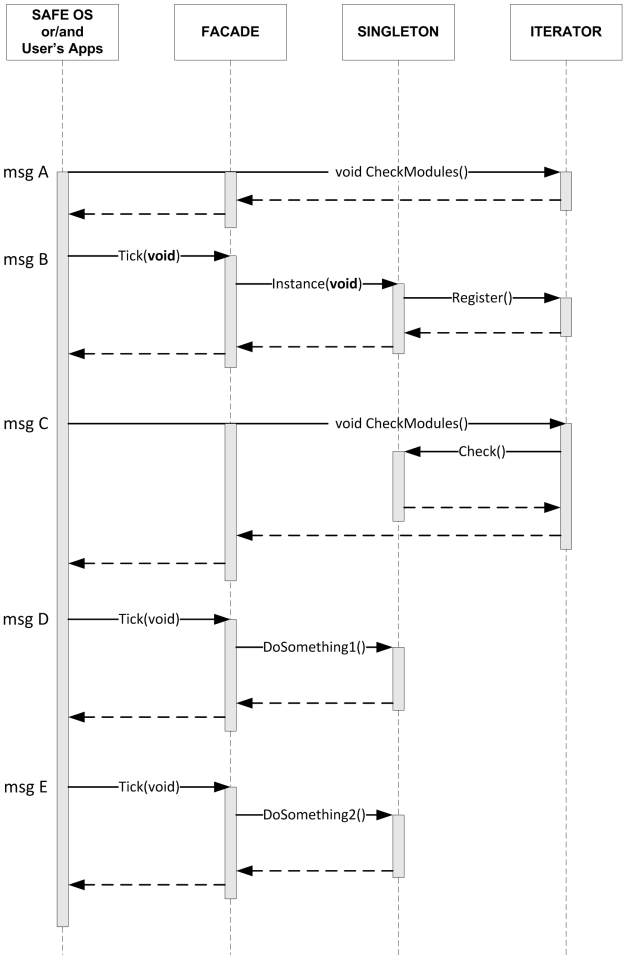


Figure 6. The sequence diagram of setting up and running the SIL3 middleware

The right part of the V-model is about verification and validation. However, verification measures, such as code review, code walkthrough, and testing, are out of the scope of this paper. For more information on this topic in general refer to [1]. For more information about verification in the range of chip development, refer to [7] .

VII. CONCLUSION

In this paper, the middleware target platform has been described. The SIL3 middleware and its modules have been depicted. A modular approach is recommended by IEC 61508, [1]. Hence, the middleware comprises different modules. Moreover, the used design patterns can be considered as proven in use design methods. This facilitates avoiding systematic faults and thus arguing SIL3 conformity. Each module serves as a low-level driver, encapsulating hardware access to a distinct functionality of the safety-related subsystem. Thus, safety-related low-level access needs not being implemented in the application code. This reduces the effort for application development and IEC 61508 certification for applications aimed at SIL1 to SIL3.

The middleware uses four design patterns: singleton, adapter, iterator, and facade. The iterator is responsible for acquiring states of the modules. While, the Facade is responsible for analyzing and detecting fail states of the modules. When the facade recognizes a fail state it requests the safe operating system to bring the whole system into a safe state. Modules are designed using the singleton pattern. Interactions between modules with other parts of the middleware are depicted on the fig. 5.

The V-model used for the SIL3 middleware development has been demonstrated. It is in conformance with IEC 61508. The typical usage scenario of the middleware has been shown in a sequence diagram. Verification is not presented in this paper. These topics are to be covered in future papers.

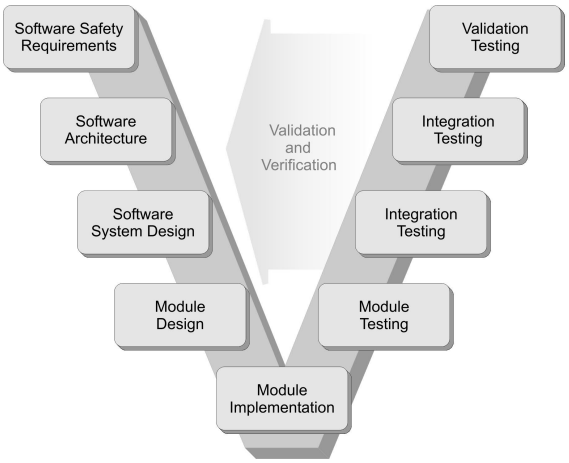


Figure 7. V-Model of the SIL3 Middleware

REFERENCES

[1] IEC/EN 61508, "International Standard: 61508 Functional safety of electrical/electronic/programmable electronic safety-related systems Part 1-7," 2nd ed., Geneva, Int. Elec. Com., 1999-2010

[2] E. Gamma, R. Helm, R. Johnson, J. Vlissides, "Design Patterns: Elements of Reusable Object-Oriented Software," USA, Addison-Wesley, 1994.

[3] IAR System, "The IAR Linker and Library Tools Reference Guide," Ver 4.61, January 2009.

[4] S. Meyers, "More Effective C++: 35 New Ways to Improve Your Programs and Designs," Addison-Wesley Professional, 1 edition, pp. 130-145, January 1996.

[5] MIRA Limited, "MISRA C++:2008 Guidelines for the use of the C++ language in critical systems," MIRA Limited, 2008.

[6] http://v-Model.iabg.de/index.php

[7] A. Hayek, M. Schreiber, and J. Börcsök, "Basic VHDL Tests Conforming to IEC 61508," in IEEE Proceedings of Seventh International Conference on Networked Sensing Systems, INSS 2010, pp. 41–44.

[8] M. Schreiber, E. Delic, A. Hayek, J. Börcsök, "Concept for a SIL3 Middleware Encapsulating Safety-related Aspects of Applications for an 8051-based SIL3 Multi-Core System-on-Chip," MIPRO 2013 – 36th International Convention: MEET - Microelectronics, Electronics and Electronic Technology, in press, 2013, Opatja, Croatia

[9] J. Börcsök, "Electronic Safety Systems. Hardware Concepts, Models, and Calculations," ISBN-13: 978-3778529447, Hütig Verlag, 2004, Heidelberg, Germany

[10] J. Börcsök, "Functional Safety: Basic Principles of Safety-Related Systems," ISBN-13: 978-3778529867, Hütig Verlag, 2007, Heidelberg, Germany

Weighted Least-Squares Design of Sharpened CIC Filters

Goran Molnar[1], Matija Glavinic Pecotic[2], and Mladen Vucic[1]

[1]University of Zagreb, Faculty of Electrical Engineering and Computing
Department of Electronic Systems and Information Processing
Unska 3, 10000 Zagreb, Croatia

[2]Siemens CMT d. d.
Heinzelova 70a, 10000 Zagreb, Croatia

E-mails: goran.molnar@fer.hr, matija.glavinic-pecotic@siemens.com, mladen.vucic@fer.hr

***Abstract*—Efficient multiplierless decimation filter is the cascaded-integrator-comb (CIC) filter. However, in many applications its magnitude response does not meet the requirement for selectivity. One technique for improving the response is sharpening. This technique results in filters having very low passband droop and high folding-band attenuations. The sharpening was first proposed by Kaiser and Hamming. They presented the polynomial amplitude-change function that is obtained by imposing flatness at the points (0,0) and (1,1). In this paper, we present a method for the design of sharpened CIC filters which is based on weighted least-squares error criterion. The method is given in a closed form which provides the optimum polynomial coefficients analytically. The filters presented are suitable for application in narrowband and wideband software radio receivers.**

I. INTRODUCTION

The simplest multiplierless decimation filter is the cascaded-integrator-comb (CIC) filter proposed in [1]. However, the CIC filter ensuring high folding-band attenuations has a high passband droop, which is usually intolerable. To meet the requirements for the passband and the folding bands simultaneously, a technique for improving the magnitude response might be used. The improvement can be made by modifying the basic CIC structure [2]–[10] or by connecting a finite-impulse-response (FIR) filter called CIC compensator in the cascade with the original filter [11]–[16]. The CIC filters with wide passbands usually have very high passband droops and, consequently, employ high-order compensators. Therefore, in the decimation of wideband signals, the modified structures are preferable.

Various modified structures are described in literature. They include the tapped interconnection of CIC filters [2], the filter with a nested FIR filter [3], the generalized CIC filter [4], the polyphase decomposition of the CIC filter [5], the filter obtained using the frequency sampling method [6], and the sharpened CIC filters [7]–[10]. In [17] and [18], compensation is combined with sharpening.

In this paper, we consider sharpened CIC filters. The sharpening of a symmetric FIR filter was introduced by Kaiser and Hamming [19]. They proposed the polynomial amplitude-change function that is obtained by imposing flatness at the points (0,0) and (1,1). The application of the Kaiser-Hamming polynomial to the CIC filter was first presented by Kwentus *et al.* [7]. The structure of the sharpened CIC filter was proposed by Saramäki and Ritoniemi [2]. In [10], Coleman incorporated the Chebyshev polynomial into the Saramäki-Ritoniemi structure.

The Kaiser-Hamming technique results in maximally flat sharpened CIC filters. Here, we present a method for the design of sharpened CIC filters based on the weighted least-squares approximation of the desired amplitude response. The desired response is specified only within the filter's passband and within the folding bands. The method is given in a closed form which provides the optimum polynomial coefficients analytically.

The paper is organized as follows. Section II describes the sharpened CIC filter. The method for the design of weighted least-squares sharpened CIC filters is presented in Section III. Section IV contains examples illustrating the features of the proposed decimation filters.

II. SHARPENED CIC FILTER

A. Amplitude Response of Sharpened CIC Filter

The transfer function of the CIC filter of the order N is given by [1]

$$H_{CIC}(z)=\left(\frac{1}{R}\frac{1-z^{-R}}{1-z^{-1}}\right)^N \tag{1}$$

where R denotes the decimation factor. The corresponding amplitude response is

$$H_{CIC}(\omega)=\left[\frac{1}{R}\frac{\sin\left(\frac{\omega R}{2}\right)}{\sin\left(\frac{\omega}{2}\right)}\right]^N \tag{2}$$

This paper was supported by the Ministry of Science, Education and Sports of the Republic of Croatia under grant no. 036-0362214-2217, and the ACROSS project funded by the Seventh Framework Programme of the European Union under grant no. 285939 FP7-REGPOT-2011-1.

The sharpening polynomial of the order M is given by

$$f(x) = \sum_{m=0}^{M} a_m x^m \tag{3}$$

By substituting $x=H_{CIC}(\omega)$ in (3), we obtain the amplitude response of the sharpened CIC filter. It is given by

$$H_{SCIC}(\omega) = \sum_{m=0}^{M} a_m H_{CIC}^m(\omega) \tag{4}$$

It is well known that the amplitude response of the original CIC filter has zeros placed at the central frequencies of the folding bands. Such a placement of zeros is preferable since it eliminates the DC component caused by the aliases. In the case of sharpened CIC filter, this property is kept for $a_0=0$. Therefore, we deal with the amplitude response

$$H_{SCIC}(\omega) = \sum_{m=1}^{M} a_m H_{CIC}^m(\omega) \tag{5}$$

B. Desired Amplitude Response

The CIC-based filter is usually followed by a lowpass filter whose passband corresponds to the band occupied by the signal of interest. Therefore, we specify the desired response only within the passband and within the folding bands. Furthermore, since the amplitude response of the decimation filter is an even function of ω, the desired response is specified over the positive band only. Such a response is given by

$$H_d(\omega) = \begin{cases} 1\,; & \omega \le \omega_p \\ 0\,; & \dfrac{2n\pi}{R} - \omega_p \le \omega \le \dfrac{2n\pi}{R} + \omega_p \\ 0\,; & \pi - \omega_p \le \omega < \pi \end{cases}, \; n = 1, \ldots, \frac{R}{2} - 1 \tag{6}$$

for an even R and by

$$H_d(\omega) = \begin{cases} 1\,; & \omega \le \omega_p \\ 0\,; & \dfrac{2n\pi}{R} - \omega_p \le \omega \le \dfrac{2n\pi}{R} + \omega_p \end{cases}, \; n = 1, \ldots, \left\lfloor \frac{R}{2} \right\rfloor \tag{7}$$

for an odd R, where ω_p is the edge of filter's passband, and $\lfloor u \rfloor$ denotes the greatest integer equal to or smaller than u.

III. Design Method

A. Problem Formulation

Our goal is to find a sharpening polynomial such that the amplitude response of the sharpened CIC filter minimizes the integral of the weighted squared error given by

$$\varepsilon_{wls}(\mathbf{a}) = \int_B w(\omega)\left[H_{SCIC}(\omega, \mathbf{a}) - H_d(\omega)\right]^2 d\omega \tag{8}$$

where B is the band of interest, $w(\omega)$ is a positive weighting function, and $\mathbf{a}$ is the vector of polynomial coefficients defined as

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_M \end{bmatrix}^T \tag{9}$$

The optimum polynomial coefficients are then found as

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \varepsilon_{wls}(\mathbf{a}) \tag{10}$$

We solve the problem in (10) analytically by using conventional procedure taken from the weighted least-squares design of the FIR filters [20].

B. Optimum Sharpening Polynomial

To find the optimum solution, we use the amplitude responses given on the uniformly spaced frequency grid $\{\omega_k;\ k=0, \ldots, K-1\}$ defined within the band B in (8). The integral in (8) is then approximated by a finite sum, and hence we use the error

$$\varepsilon(\mathbf{a}) = \sum_{k=0}^{K-1} w(\omega_k)\left[H_{SCIC}(\omega_k, \mathbf{a}) - H_d(\omega_k)\right]^2 \tag{11}$$

Equation (11) can be written in a matrix form as

$$\varepsilon(\mathbf{a}) = \left\| \mathbf{W}(\mathbf{H}\mathbf{a} - \mathbf{d}) \right\|^2 \tag{12}$$

where

$$\mathbf{W} = \begin{bmatrix} \omega_0 & 0 & \cdots & 0 \\ 0 & \omega_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_{K-1} \end{bmatrix} \tag{13}$$

$$\mathbf{H} = \begin{bmatrix} H_{CIC}(\omega_0) & H_{CIC}^2(\omega_0) & \cdots & H_{CIC}^M(\omega_0) \\ H_{CIC}(\omega_1) & H_{CIC}^2(\omega_1) & \cdots & H_{CIC}^M(\omega_1) \\ \vdots & \vdots & \ddots & \vdots \\ H_{CIC}(\omega_{K-1}) & H_{CIC}^2(\omega_{K-1}) & \cdots & H_{CIC}^M(\omega_{K-1}) \end{bmatrix} \tag{14}$$

and

$$\mathbf{d} = \begin{bmatrix} H_d(\omega_0) & H_d(\omega_1) & \cdots & H_d(\omega_{K-1}) \end{bmatrix}^T \tag{15}$$

The error in (12) can be expressed as

$$\varepsilon(\mathbf{a}) = (\mathbf{H}\mathbf{a} - \mathbf{d})^T \mathbf{W}(\mathbf{H}\mathbf{a} - \mathbf{d}) \tag{16}$$

and further as

$$\varepsilon(\mathbf{a}) = (\mathbf{a}^T\mathbf{H}^T - \mathbf{d}^T)\mathbf{W}(\mathbf{H}\mathbf{a} - \mathbf{d}) \tag{17}$$

After matrix multiplication, we arrive at

$$\varepsilon(\mathbf{a}) = \mathbf{a}^T\mathbf{H}^T\mathbf{W}\mathbf{H}\mathbf{a} - \mathbf{a}^T\mathbf{H}^T\mathbf{W}\mathbf{d} - \mathbf{d}^T\mathbf{W}\mathbf{H}\mathbf{a} + \mathbf{d}^T\mathbf{W}\mathbf{d} \tag{18}$$

and finally at

$$\varepsilon(\mathbf{a}) = \mathbf{a}^T\mathbf{H}^T\mathbf{W}\mathbf{H}\mathbf{a} - 2\mathbf{a}^T\mathbf{H}^T\mathbf{W}\mathbf{d} + \mathbf{d}^T\mathbf{W}\mathbf{d} \tag{19}$$

By taking

$$\frac{\partial \varepsilon(\mathbf{a})}{\partial \mathbf{a}} = \mathbf{0} \tag{20}$$

the optimum polynomial coefficients are given by

$$\hat{\mathbf{a}} = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}(\mathbf{H}^T\mathbf{W}\mathbf{d}) \tag{21}$$

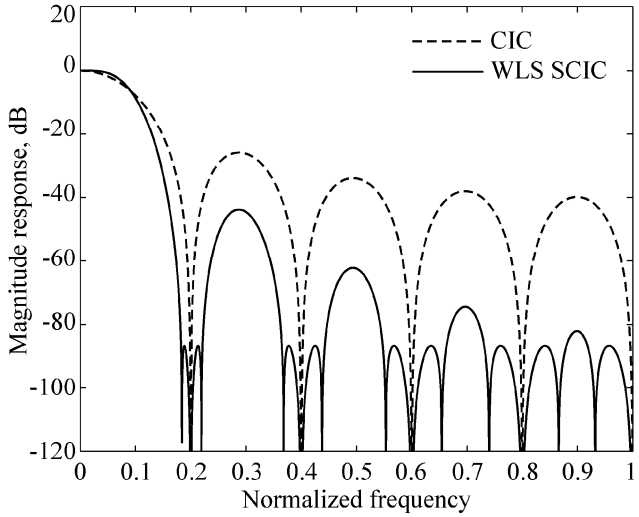


Figure 1. Magnitude responses of original CIC filter with N=2 and R=10, and weighted least-squares sharpened CIC (WLS SCIC) filter obtained for M=3. Passband is $|\omega| \le 0.02\pi$.

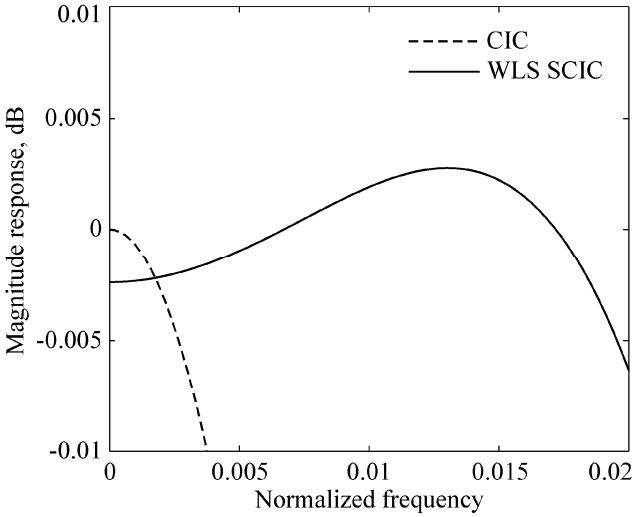


Figure 2. Passband magnitude responses of original CIC filter with N=2 and R=10, and weighted least-squares sharpened CIC filter obtained for M=3.

The minimizer in (21) is unique if the matrix $\mathbf{H}^T\mathbf{W}\mathbf{H}$ is positive definite. It is always met if $\mathbf{W}$ is the unity matrix. If such is not the case, an appropriate weighting function should be chosen.

It is clear that the calculation of the coefficients in (21) contains a matrix inversion and matrix multiplications. However, these operations do not cause problems in implementation since the number of coefficients is rather low.

IV. Design Examples

To illustrate the features of the proposed sharpened CIC filters, three examples are provided. The first example presents a narrowband sharpened filter. The second example describes a wideband filter. Finally, the comparison with a maximally flat filter obtained by the Kaiser-Hamming method is given in the third example. In all examples, the CIC filter with N=2 and R=10 is considered. Furthermore, the frequency grid with K=2000 points is used.

A. Narrowband Filter Design

In the decimation of narrowband signals, the sharpening based on low order polynomials is usually sufficient. Here, we design a narrowband sharpened CIC filter by using the third order polynomial. The design is worked out with the weighting function

$$w(\omega)=\begin{cases}1 & ;\ \omega \le \omega_p \\ 5 & ;\ \dfrac{\pi}{5}-\omega_p \le \omega \le \dfrac{\pi}{5}+\omega_p \\ 0.1 & ;\ \dfrac{n\pi}{5}-\omega_p \le \omega \le \dfrac{n\pi}{5}+\omega_p \ ,\ n=2,3,\text{and}\,4 \\ 0.1 & ;\ \pi-\omega_p \le \omega < \pi \end{cases} \tag{22}$$

and ω_p=0.02π. Using (21), the optimum polynomial coefficients are obtained as

$$\hat{\mathbf{a}}=\begin{bmatrix}-0.024199\\ 3.134081\\ -2.110156\end{bmatrix} \tag{23}$$

The magnitude responses of the original and the sharpened filter are shown in Figure 1. Their passband responses are shown in Figure 2. It is clear that the significant improvement is achieved despite the low order of the sharpening polynomial. The smallest folding-band attenuation of the CIC filter is 38 dB and the passband droop is 0.28 dB. The sharpened filter increases the folding-band attenuation to 75 dB and ensures the maximum passband deviation of 0.006 dB.

B. Wideband Filter Design

Wideband sharpening can be performed by employing high order polynomials. In this example, we design the filter by using the seventh order polynomial. The sharpening is made assuming the passband edge is ω_p=0.05π and the weighting function is given in (22). The optimum polynomial coefficients are obtained as

$$\hat{\mathbf{a}}=\begin{bmatrix}-0.002698\\ 0.323388\\ -9.636580\\ 89.50077\\ -195.1686\\ 165.8597\\ -49.87631\end{bmatrix} \tag{24}$$

Figure 3 shows the magnitude response of the filter obtained together with the response of the original filter. The corresponding passband responses are shown in Figure 4. It is clear that the sharpened filter exhibits a good selectivity what is the consequence of simultaneous widening of the passband and the first folding band. Furthermore, the CIC filter has the worst folding-band

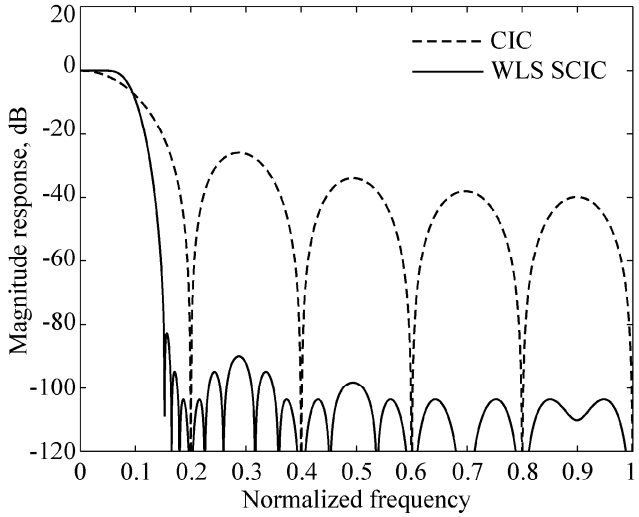


Figure 3. Magnitude responses of original CIC filter with N=2 and R=10, and weighted least-squares sharpened CIC filter obtained for M=7. Passband is $|\omega| \leq 0.05\pi$.

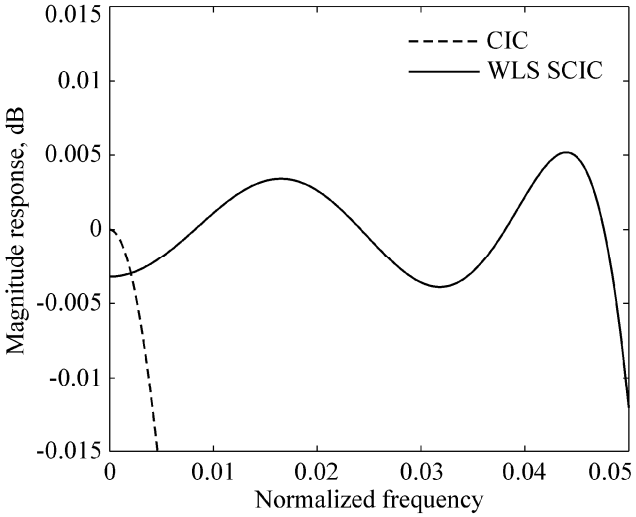


Figure 4. Passband magnitude responses of original CIC filter with N=2 and R=10, and weighted least-squares sharpened CIC filter obtained for M=7.

attenuation of 21 dB, whereas the sharpened filter has the attenuation of 74 dB. The passband droop of the original filter is 1.8 dB, whereas the improved response exhibits ripple with the largest error of 0.012 dB.

C. Comparison With Kaiser-Hamming Filters

Let us now compare the proposed sharpened CIC filter with the maximally flat counterpart. To obtain a maximally flat sharpened CIC filter, we use the Kaiser-Hamming method [19]. Their method constructs the polynomial which has an mth-order tangency at unity, nth-order tangency at zero, and passes through the points (1,1) and (0,0). These conditions can be formulated as

Passband	Stopband
$f(1)=1$	$f(0)=0$
$f'(1)=0$	$f'(0)=0$
⋮	⋮
$f^{(m)}(1)=0$	$f^{(n)}(0)=0$

where $m+n=M-1$. The polynomial satisfying the above conditions is given by [19]

$$f(x) = x^{n+1} \sum_{r=0}^{m} \frac{(n+r)!}{n!r!} (1-x)^r \tag{25}$$

The choice of m and n depends on application at hand. To obtain the sharpened CIC filter with low passband droop, m should be large. On the other hand, to ensure high folding-band attenuations, n should be large. A reasonably good magnitude is obtained if m is equal or somewhat smaller than n. For illustration we use the polynomial of the fifth order, obtained for m=2 and n=2. It is given by

$$f(x) = 10x^3 - 15x^4 + 6x^5 \tag{26}$$

For comparison, we design the filter using the same order of the sharpening polynomial, $\omega_p=0.03\pi$, and unity weighting function. The optimum coefficients are obtained as

$$\hat{\mathbf{a}} = \begin{bmatrix} 0.001694 \\ -0.342322 \\ 12.10196 \\ -18.21024 \\ 7.448995 \end{bmatrix} \tag{27}$$

Figure 5 shows the magnitude response of the maximally flat and the least squares sharpened filter. Their passband responses are shown in Figure 6. For the given band of interest, the CIC filter introduces the passband droop of 0.64 dB and attenuates aliases with more than 30 dB. The maximally flat filter ensures the droop less than 0.03 dB and the folding-band attenuations higher than 72 dB. However, the least squares filter improves the response better. It has the passband ripple of 0.003 dB and the smallest folding-band attenuation of 86 dB.

From the implementation point of view, the maximally flat filter is less complex than the least squares filter because it has two coefficients equal to zero. According to the Saramäki-Ritoniemi structure [2], the least squares filter employs 10 integrators, 10 combs, 5 multipliers, 20 delay elements, 5 downsamplers, and 4 interconnection adders. However, the structure of the maximally flat filter saves 2 multipliers, 14 delay elements, and 2 downsamplers.

V. Conclusion

The design of weighted least-squares sharpened CIC filters has been presented. The design is formulated as an unconstrained optimization problem. The optimum polynomial coefficients are obtained in a closed form. In comparison with the Kaiser-Hamming decimation filters, the proposed filters ensure better magnitude responses. However, for the same polynomial order, they have a larger number of non-zero coefficients, thus resulting in higher complexity.

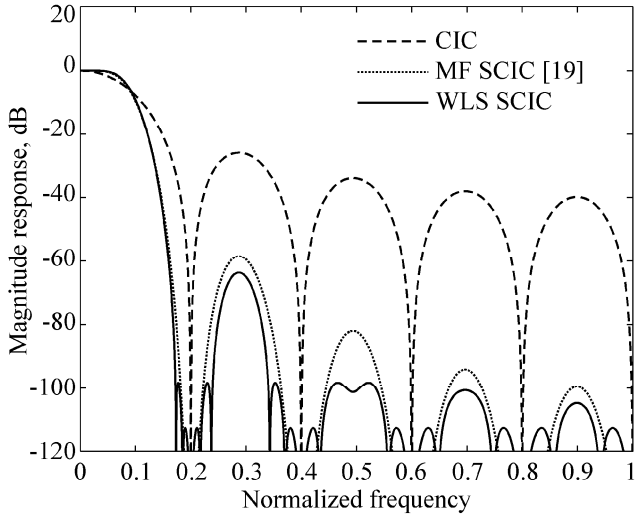


Figure 5. Magnitude responses of original CIC filter with N=2 and R=10, maximally flat (MF) [19] and weighted least-squares sharpened CIC filter obtained for M=5. Passband is considered within $|\omega| \leq 0.03\pi$.

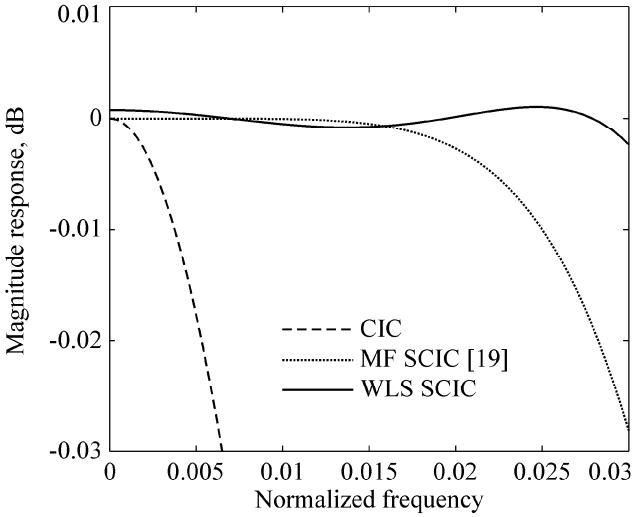


Figure 6. Passband magnitude responses of original CIC filter with N=2 and R=10, maximally flat [19] and weighted least-squares sharpened CIC filter obtained for M=5.

VI. REFERENCES

[1] E. B. Hogenauer, "An economical class of digital filters for decimation and interpolation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 2, pp. 155–162, Apr. 1981.

[2] T. Saramäki and T. Ritoniemi, "A modified comb filter structure for decimation," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 4, Hong Kong, Jun. 1997, pp. 2353–2356.

[3] V. Lehtinen and M. Renfors, "On CIC decimator variants - from shifting zeros to the sparse FIR-CIC structure," in *Proc. Int. Symp. Signal Process. Appl.*, Sharjah, United Arab Emirates, Feb. 2007.

[4] M. Laddomada, "Generalized comb decimation filters for ΣΔ A/D converters: Analysis and design," *IEEE Trans. Circuits Syst. I: Reg. Papers*, vol. 54, no. 5, pp. 994–1005, May 2007.

[5] G. Jovanovic Dolecek and S. K. Mitra, "Two-stage CIC-based decimator with improved characteristics," *IET Signal Process.*, vol. 4, no. 1, pp. 22–29, Feb. 2010.

[6] G. Jovanovic Dolecek and A. Fernandez-Vazquez, "Novel droop-compensated comb decimation filter with improved alias rejections," *Int. J. Electron. Commun. (AEÜ)*, vol. 67, pp. 387–396, 2013.

[7] A. Y. Kwentus, Z. Jiang, and A. N. Willson Jr., "Application of filter sharpening to cascaded integrator-comb decimation filters," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 457–467, Feb. 1997.

[8] G. Stephen and R. W. Stewart, "High-speed sharpening of decimating CIC filter," *Electron. Lett.*, vol. 40, no. 21, pp. 1383–1384, Oct. 2004.

[9] G. Jovanovic Dolecek and S. K. Mitra, "A new two-stage sharpened comb decimator," *IEEE Trans. Circuits Syst. I: Reg. Papers*, vol. 52, no. 7, pp. 1414–1420, Jul. 2005.

[10] J. O. Coleman, "Chebyshev stopbands for CIC decimation filters and CIC-implemented array tapers in 1D and 2D," *IEEE Trans. Circuits Syst. I: Reg. Papers*, vol. 59, no. 12, pp. 2956–2968, Dec. 2012.

[11] S. Kim, W. C. Lee, S. Ahn, and S. Choi, "Design of CIC roll-off compensation filter in a W-CDMA digital IF receiver," *Digit. Signal Process.*, vol. 16, no. 6, pp. 846–854, 2006.

[12] G. Jovanovic Dolecek and S. K. Mitra, "Simple method for compensation of CIC decimation filter," *Electron. Lett.*, vol. 44, no. 19, pp. 1162–1163, Sept. 2008.

[13] G. Jovanovic Dolecek, "Simple wideband CIC compensator," *Electron. Lett.*, vol. 45, no. 24, pp. 1270–1272, Nov. 2009.

[14] G. Molnar and M. Vucic, "Closed-form design of CIC compensators based on maximally flat error criterion," *IEEE Trans. Circuits Syst. II: Exp. Briefs*, vol. 58, no. 12, pp. 926–930, Dec. 2011.

[15] A. Fernandez-Vazquez and G. Jovanovic Dolecek, "Maximally flat CIC compensation filter: Design and multiplierless implementation," *IEEE Trans. Circuits Syst. II: Exp. Briefs*, vol. 59, no. 2, pp. 113–117, Feb. 2012.

[16] M. Glavinic Pecotic, G. Molnar, and M. Vucic, "Design of CIC compensators with SPT coefficients based on interval analysis," in *Proc. 35th Int. Convention MIPRO*, vol. MEET, Opatija, Croatia, May 2012, pp. 129–134.

[17] G. Jovanovic Dolecek and F. Harris, "Design of wideband CIC compensator filter for a digital IF receiver," *Digit. Signal Process.*, vol. 19, no. 5, pp. 1204–1213, 2000.

[18] C. Jeong, Y.-J. Min, and S.-W. Kim, "Double-sharpened decimation filter employing a pre-droop compensator for multistandard wireless applications," *ETRI J.* , vol. 33, no. 2, pp. 169–175, Apr. 2011.

[19] J. Kaiser and R. Hamming, "Sharpening the response of a symmetric nonrecursive filter by multiple use of the same filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, no. 5, pp. 415–422, Oct. 1977.

[20] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications*. Springer, New York, 2007.

Some Improvements of the Gaussian Elimination Method for Solving Simultaneous Linear Equations

S. Delić*, Ž. Jurić*
* Faculty of Electrical Engineering/Department of Automatic Control & Electronics, Sarajevo, Bosnia & Herzegovina
delic.sead@hotmail.com; zjuric@utic.net.ba

Abstract – Although it is known that Gaussian elimination method for solving simultaneous linear equations is not asymptotically optimal, it is still one of the most useful methods for solving systems of moderate size. This paper proposes some ideas how to speed-up the standard method. First, the trick which takes the advantage of the eventual symmetry of the system is presented, which speeds up the calculation by the factor slightly less than 2. Second, it is shown that by using some rearrangement of the calculation, it is possible to get additional speed-up, no matter whether the system is symmetric or not, although the eventual symmetry additionally doubles the execution speed. This rearrangement is performed using similar approach as in LU factorization, but retaining basic features of the Gaussian elimination method, like producing the triangular form of the system. As the required modifications in the original method are quite simple, the improved method may be used in all engineering applications where the original Gaussian elimination is used.

I. INTRODUCTION

The Gaussian elimination method and its close relatives like LU factorization are probably the most widely used methods for solving systems of linear equations [1]. Although it is known that its running time $\Theta(n^3)$ where n is number of the equations is not asymptotically optimal, other methods that have better asymptotical running time are mostly useless for systems that arise in the engineering practice. For example, Strassen in [2] demonstrates the algorithm that solves the system of n linear equations in with asymptotical running time $\Theta(n^{2,807})$. However, in practice, this method is faster than the Gaussian-like methods only for values of n greater than about 1000. Also, the algorithm given by Coppersmith and Winograd [3] is the excellent example for the extreme discrepancy between the theory and the practice. Namely, although their algorithm is asymptotically the fastest known algorithm for solving systems of linear equations, which has asymptotical running time $\Theta(n^{2,376})$, in reality it requires systems so large that they can not be processed by any actual hardware to be faster than Gaussian-like algorithms. In other words, it is completely useless in practice. Therefore, for solving systems that appear in engineering practice, Gaussian-like methods have no any real alternative.

The practical problems from the engineering often produce systems that are symmetric in nature, i. e. systems where the matrix of coefficients is a symmetric one. For example, either nodal analysis or loop current analysis of an electrical circuit without dependent sources [4] produces a symmetric set of equations. However, many known methods from the Gaussian family of methods in their original form can not take any advantage of the eventual symmetry of the systems. One exception is the Cholesky factorization method for symmetric systems [5], which is about twice faster than other methods from the same family of methods. Therefore, many books dedicated to the numerical methods usually recommend this method for solving symmetric systems. However, it may be applied only to the systems whose matrix of coefficients is positive definite, which is a big limitation. Otherwise, the method will fail. Moreover, this method requires the calculation of the square roots, so it is not possible to implement the method using the exact arithmetic. An alternative called symmetric indefinite factorization (or LDL^T factorization) that avoids calculation of the square root is also known [6]. Although this method may be applied even for many matrices that are not positive definite, it also can fail sometimes. Some possible improvements of this method are described in [7], but all of them are quite complicated. In this paper, we will show that it is possible to modify even the basic Gaussian elimination method to speed up the calculation when the system is symmetric. Moreover, the method remains very simple even after such modification.

Another problem with the classic Gaussian elimination method is that it performs too many accesses to the memory where the coefficients of the system are stored. This slows down the procedure, even with compilers that perform aggressive optimizations. We will show in the paper how the calculations may be rearranged to reduce accessing to the storage memory, and consequently to speed up the procedure. The expected gain in the speed is always above 1.3 – 1.4, although we sometimes get even larger improvements, up to nearly 2 (this depends of the used compiler and the programming language). The new method gives similar formulae as for the LU factorization, but it is still the Gaussian elimination and produces the same output as the ordinary Gaussian elimination (the triangular form of the system). Moreover, the procedure is more accurate and slightly faster than LU factorization.

Section 2 of the paper shows how the ordinary Gaussian elimination may be improved for the symmetric systems.

Section 3 shows how the Gaussian elimination may be rearranged to avoid too many accesses to the elements of storage matrix. Finally, Section 4 illustrates how the method developed in Section 3 may be improved further for symmetric matrices.

II. GAUSSIAN ELIMINATION FOR SYMMETRIC MATRICES

Although Gaussian elimination method is well-known, we must begin this section with the brief description of this method, just to introduce the terminology and the notations that will be used in the paper.

Suppose that we have the system of linear equations in matrix form $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, where $\mathbf{A}$ is known $n \times n$ matrix with coefficients $a_{i,j}$, $i = 1..n$, $j = 1..n$, $\mathbf{b}$ is known vector with coefficients b_i, $i = 1..n$ and $\mathbf{x}$ is the vector of unknowns x_i, $i = 1..n$. The Gaussian elimination method iteratively replaces the initial system $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ with equivalent systems $\mathbf{A}^{(q)}\,\mathbf{x} = \mathbf{b}^{(q)}$, $q = 1..n–1$ where matrices $\mathbf{A}^{(q)}$ are such that their coefficients $a_{i,j}^{(q)}$ satisfy the condition

$$a_{i,j}^{(q)} = 0\,,\ j = 1..q,\ i = j+1..n \tag{1}$$

After n–1 iterations, the matrix $\mathbf{A}^{(n-1)}$ is upper triangular matrix, so the last obtained system $\mathbf{A}^{(n-1)}\mathbf{x} = \mathbf{b}^{(n-1)}$ may be easily solved using the back substitution:

$$x_i = \frac{1}{a_{i,i}^{(n-1)}}\left(b_i^{(n-1)} - \sum_{j=i+1}^{n} a_{i,j}^{(n-1)}\,x_j\right),\ i = n..1 \tag{3}$$

Assume for the moment that the pivoting is not used in the procedure. Then, the coefficients of $\mathbf{A}^{(q)}$ and $\mathbf{b}^{(q)}$ can be calculated from the coefficients of $\mathbf{A}^{(q-1)}$ and $\mathbf{b}^{(q-1)}$ using formulae

$$a_{i,j}^{(q)} = a_{i,j}^{(q-1)}\ ,\ 1 \le i \le q \vee 1 \le j \le q \tag{4}$$

$$a_{i,q}^{(q)} = 0\,,\ i = q+1..n \tag{5}$$

$$a_{i,j}^{(q)} = a_{i,j}^{(q-1)} + \xi_{i,q}\,a_{q,j}^{(q-1)}\,,\ i = q+1..n,\ j = q+1..n \tag{6}$$

$$b_i^{(q)} = b_i^{(q-1)},\ i = 1..q \tag{7}$$

$$b_i^{(q)} = b_i^{(q-1)} + \xi_{i,q}\,b_q^{(q-1)},\ i = q+1..n \tag{8}$$

where

$$\xi_{i,q} = -\frac{a_{i,q}^{(q-1)}}{a_{q,q}^{(q-1)}}\,,\ i = q+1..n \tag{9}$$

Here, we have $\mathbf{A}^{(0)} = \mathbf{A}$ and $\mathbf{b}^{(0)} = \mathbf{b}$, i. e. $a_{i,j}^{(0)} = a_{i,j}$ and $b_j^{(0)} = b_j$. Of course, only (6), (8) and (9) should really be implemented. Namely, (4) and (7) just say which coefficients stay the same as in the previous iteration, and coefficients (5) need not to be calculated, as we know that they are zeros.

Note also that auxiliary coefficients $\xi_{i,q}$ need not to be stored between iterations, and that all of them may be kept in the same variable if the processing is performed row by row, as the same coefficient $\xi_{i,q}$ is used for the whole row in (6). However, storing all of coefficients $\xi_{i,q}$ may be useful if later the vector $\mathbf{b}$ is changed, because it is then enough to perform recalculation only using (8), as all $\xi_{i,q}$ are known. This does not need any extra space, because coefficients $\xi_{i,q}$ may be stored instead of elements $a_{i,q}^{(q)}$, as they are zeros anyway, and need not to be stored at all. Note also that the back substitution formula (3) may be also expressed as

$$x_i = \frac{1}{a_{i,i}^{(i-1)}}\left(b_i^{(i-1)} - \sum_{j=i+1}^{n} a_{i,j}^{(i-1)}\,x_j\right),\ i = n..1 \tag{10}$$

This simply follows from (3), (4) and (7).

So far, nothing in the procedure suggests how we can exploit the eventual symmetry of $\mathbf{A}$, i. e. the property $a_{i,j} = a_{j,i}$, $i = 1..n$, $j = 1..n$. However, it will be shown in the paper that assuming that the pivoting is not performed, this condition implies

$$a_{i,j}^{(q)} = a_{j,i}^{(q)}\,,\ i = q+1..n,\ j = q+1..n \tag{11}$$

This means that nearly one half of coefficients in (6) need not to be calculated at all (later we will explain how to incorporate pivoting in this scheme). In other words, instead of (6), we should perform calculations

$$a_{i,j}^{(q)} = a_{i,j}^{(q-1)} + \xi_{i,q}\,a_{q,j}^{(q-1)}\,,\ i = q+1..n,\ j = i..n \tag{12}$$

This will save a lot of calculations. However, then coefficients $\xi_{i,q}$ can not be calculated using (9), as it requires $a_{i,q}^{(q-1)}$ for $i = q+1..n$, which are not calculated. Fortunately, we can use the symmetry property (11) to conclude that $a_{i,q}^{(q-1)} = a_{q,i}^{(q-1)}$ for $i = q+1..n$. Therefore, $\xi_{i,q}$ may be calculated using

$$\xi_{i,q} = -\frac{a_{q,i}^{(q-1)}}{a_{q,q}^{(q-1)}}\,,\ i = q+1..n \tag{13}$$

Using (12) (together with (13)) instead of (6) saves about one half of calculations of $a_{i,j}^{(q)}$. More precise, it is easy to see that exact saving factor tends to 1/2 as n increases. Note that the amount of other calculations remains the same even in the symmetric case. However, calculations (6) or (12) require $\Theta(n^3)$ unit operations and all other calculations require only $\Theta(n^2)$ unit operations, so the speed up factor also tends to 2 as n increases. Our measurements show that the expected gain in speed is about 1.4 for $n = 10$ and about 1.8 for $n = 50$.

Now, we need to explain how to perform pivoting with described optimized algorithm. It is known that the pivoting (at least partial one) is absolutely necessary for the accuracy and the numerical stability of the Gaussian elimination, and that without the pivoting the algorithm may even fail (when a zero is encountered on the main diagonal). However, performing the pivoting with the described optimized method is a bit tricky, but possible. This is probably the main reason why the possibility of the optimization of the Gaussian elimination for symmetric matrices was unnoticed in the literature.

The main problem with the pivoting is the fact that the pivoting (either partial or full) destroys the basic symmetric property (11). To solve this problem, we must perform such kind of the pivoting that (11) remains valid. This means that we must search for the pivot element only along the main diagonal below the element $a_{q,q}^{(q-1)}$, i. e. among elements $a_{i,i}^{(q-1)}$, $i = q..n$. When the pivot element is found, say $a_{s,s}^{(q-1)}$, $q \le s \le n$, we need to exchange both q-th row with s-th row in matrix $\mathbf{A}^{(q)}$ and in vector $\mathbf{b}^{(q)}$, and q-th column with s-th column in matrix $\mathbf{A}^{(q)}$. Exchanging of both rows and columns is necessary to keep the symmetry property (11) valid even after the pivoting. Note that exchanging of columns of $\mathbf{A}^{(q)}$ also requires the exchanging of rows of $\mathbf{x}$, otherwise the equation $\mathbf{A}^{(q)}\mathbf{x} = \mathbf{b}^{(q)}$ will not stay valid. This effectively requires prenumeration of the unknowns from vector $\mathbf{x}$, so it is necessary to keep track of such prenumerations in the program code, similarly as when the full pivotization is performed in the ordinary Gaussian elimination.

At first, it seems that the necessity of swapping both rows and columns slows slightly the procedure in comparison to the partial pivoting that is usually performed with the ordinary Gaussian elimination. However, due to the symmetric property (11), it is easy to show that only q swaps are necessary, exactly as with the partial pivoting used with the ordinary Gaussian elimination.

One potential problem with the pivoting along the diagonal is the possibility that all elements on the main diagonal below $a_{q,q}^{(q-1)}$ are zero (or very small) although the matrix $\mathbf{A}$ is invertible. This situation is not likely for matrices that occur in the engineering practice. For example, this will never happen for the diagonally dominant matrices. However, this can still happen. In such case, the described procedure fails. Then, to resolve this situation, we can pick some column (say, r-th column) in right of q-th column that has non-zero value in q-th row and add this column to the q-th column of $\mathbf{A}^{(q)}$. To keep the symmetry, we also need to add r-th row to the q-th row of $\mathbf{A}^{(q)}$, although due to the symmetry property, only one half of elements really need to be updated. As such modification effectively requires a linear transformation of some unknown in $\mathbf{x}$ (to keep the validity of $\mathbf{A}^{(q)}\mathbf{x} = \mathbf{b}^{(q)}$), it is necessary to keep track of such transformations in the program code. Alternatively, if we want to sacrifice the speed, we can use (11) to recover elements of $\mathbf{A}^{(q)}$ which are not calculated, and afterwards perform ordinary partial pivoting (which will destroy the symmetry) and continue with the ordinary Gaussian elimination.

We will give a simple induction proof of validity of (11). First, we will prove that (11) is true for $q = 1$. Really, from the assumption that $a_{i,j} = a_{j,i}$ and from (6), (9) and $\mathbf{A}^{(0)} = \mathbf{A}$, we have

$$\begin{aligned} a_{i,j}^{(1)} &= a_{i,j}^{(0)} + \xi_{i,1}\, a_{1,j}^{(0)} = a_{i,j}^{(0)} - \frac{a_{i,1}^{(0)}}{a_{1,1}^{(0)}}\, a_{1,j}^{(0)} = \\ &= a_{i,j} - \frac{a_{i,1}}{a_{1,1}}\, a_{1,j} = a_{j,i} - \frac{a_{1,i}}{a_{1,1}}\, a_{j,1} = \\ &= a_{j,i} - \frac{a_{j,1}}{a_{1,1}}\, a_{1,i} = a_{j,i}^{(0)} - \frac{a_{j,1}^{(0)}}{a_{1,1}^{(0)}}\, a_{1,i}^{(0)} = \\ &= a_{j,i}^{(0)} + \xi_{j,1}\, a_{1,i}^{(0)} = a_{j,i}^{(1)}, \quad i = 2..n,\ j = 2..n \end{aligned} \tag{14}$$

Suppose now that (11) is valid for some $q = k$, where $1 \le k \le n-2$. We will show that (11) is then also valid for $q = k+1$. Really, from the inductive hypothesis and from (6) and (9), we have:

$$\begin{aligned} a_{i,j}^{(k+1)} &= a_{i,j}^{(k)} + \xi_{i,k+1}\, a_{k+1,j}^{(k)} = a_{i,j}^{(k)} - \frac{a_{i,k+1}^{(k)}}{a_{k+1,k+1}^{(k)}}\, a_{k+1,j}^{(k)} = \\ &= a_{i,j}^{(k)} - \frac{a_{k+1,j}^{(k)}}{a_{k+1,k+1}^{(k)}}\, a_{i,k+1}^{(k)} = a_{j,i}^{(k)} - \frac{a_{j,k+1}^{(k)}}{a_{k+1,k+1}^{(k)}}\, a_{k+1,i}^{(k)} = \\ &= a_{j,i}^{(k)} + \xi_{j,k+1}\, a_{k+1,i}^{(k)} = a_{j,i}^{(k+1)}, \quad i = k+2..n,\ j = k+2..n \end{aligned} \tag{15}$$

Finally, using the principle of the mathematical induction, we can conclude that (11) is valid for all $q = 1..n-1$.

The natural question is how the proposed method compares with the other known methods for symmetric systems. The LDL^T factorization is based on formulae

$$d_i = a_{i,i} - \sum_{k=1}^{i-1} l_{i,k}^{\,2}\, d_k, \quad i = 1..n \tag{16}$$

$$l_{j,i} = \frac{1}{d_i}\left(a_{j,i} - \sum_{k=1}^{i-1} l_{i,k}\, l_{j,k}\, d_k \right), \quad i = 1..n,\ j = i+1..n \tag{17}$$

The naive implementation based directly on (16) and (17) requires too many multiplications and accesses to the storage memory in the innermost loops. Using the approach described in [8], which is based on caching values of $l_{i,k}\, d_k$ in an auxilary vector, it is possible to implement LDL^T factorization in about $n^3/6$ multiplications. However, the exact total number of multiplications that are necessary to solve the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ using LDL^T is $n^3/6 + 3n^2/2 - 5n/3$, in opposite to $n^3/6 + n^2 - 7n/6$ in our method, which is smaller by $n(n-1)/2$ (the number of other arithmetical operations is the same). Also, our method does not need any auxiliary storage. Moreover, it is quite complicated to implement pivotization in LDL^T scheme, even along the diagonal. Various approaches for doing so are described in detail in [7] and [9]. The stability of so-called diagonal pivoting method is proven in [10], even for quite ill-behaved matrices. However, it is very involved and requires about $O(n^3)$ comparisons to implement pivoting, which slows down considerably the method. The

method described here is much easier both to implement and to teach, which also may be interesting from the pedagogical aspect of view, and works equally well for well-behaved matrices (without too small elements on the diagonal) that usually arise in the engineering practice. Even if the matrix is not so well-behaved, the described linear transformation (i. e. change-the-variable) trick that is easy to implement can solve problems that arise if all elements on the diagonal become too small during the process.

III. COMPACTED GAUSSIAN ELIMINATION (GAUSS X_p METHOD)

Now, we will present the modification of the standard Gaussian elimination, which may be called compacted Gaussian elimination or Gauss X_p method (in opposite to the standard Gaussian elimination, which can be called Gauss X_q method). To see the difference, it is useful to represent the standard Gaussian elimination in the matrix form. From (4), it is easy to conclude that the matrix $\mathbf{A}^{(q)}$ that is obtained after q-th iteration of the Gaussian elimination may be represented in the following form:

$$\mathbf{A}^{(q)} = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,q} & a_{1,q+1} & \cdots & a_{1,n} \\ 0 & a_{2,2}^{(1)} & a_{2,3}^{(1)} & \cdots & a_{2,q}^{(1)} & a_{2,q+1}^{(1)} & \cdots & a_{2,n}^{(1)} \\ 0 & 0 & a_{3,3}^{(2)} & \cdots & a_{3,q}^{(2)} & a_{3,q+1}^{(2)} & \cdots & a_{3,n}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_{q,q}^{(q-1)} & a_{q,q+1}^{(q-1)} & \cdots & a_{q,n}^{(q-1)} \\ 0 & 0 & 0 & \cdots & 0 & a_{q+1,q+1}^{(q)} & \cdots & a_{q+1,n}^{(q)} \\ 0 & 0 & 0 & \cdots & 0 & a_{q+2,q+1}^{(q)} & \cdots & a_{q+2,n}^{(q)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & a_{n,q+1}^{(q)} & \cdots & a_{n,n}^{(q)} \end{pmatrix} \quad (18)$$

Of course, the transformation from $\mathbf{A}^{(q-1)}$ to $\mathbf{A}^{(q)}$ is performed using (6) (together with trivial relations (4) and (5)). However, it is useful to look at this transformation as a matrix multiplication. Namely, it is easy to see that we have $\mathbf{A}^{(q)} = \mathbf{X}_q \mathbf{A}^{(q-1)}$, where

$$\mathbf{X}_q = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \xi_{q+1,q} & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \xi_{q+2,q} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \xi_{n,q} & 0 & 0 & \cdots & 1 \end{pmatrix} \quad (19)$$

Really, (6) is then nothing more than the rule for the matrix multiplication, and coefficients in (9) are selected such that (5) holds.

The matrix $\mathbf{X}_q$ is an elementary lower triangular matrix in which only q-th column is nontrivial, i. e. different from an unit vector. The compacted Gaussian elimination is based on the multiplication with slightly different elementary lower triangular matrix $\mathbf{X}_p$ in which only p-th row is nontrivial:

$$\mathbf{X}_p = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \xi_{p,1} & \xi_{p,2} & \cdots & \xi_{p,p-1} & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \quad (20)$$

The process is again iterative, i. e. we will iteratively perform $\mathbf{A}^{(p)} = \mathbf{X}_p \mathbf{A}^{(p-1)}$ (and, of course, $\mathbf{b}^{(p)} = \mathbf{X}_p \mathbf{b}^{(p-1)}$). As $\mathbf{X}_1$ is in fact the identity matrix, the iterations will start from $p = 2$, and we will set that $\mathbf{A}^{(1)} = \mathbf{A}$.

Note that the left multiplication of $\mathbf{A}^{(p-1)}$ and $\mathbf{b}^{(p-1)}$ with $\mathbf{X}_p$ changes only the p-th row of $\mathbf{A}^{(p-1)}$ and $\mathbf{b}^{(p-1)}$. As a consequence, we have $a_{i,j}^{(p)} = a_{i,j}^{(p-1)}$ and $b_i^{(p)} = b_i^{(p-1)}$ for $i \neq p$, because only p-th row is changed in p-th iteration. Moreover, each row is changed only in one iteration. This means that after p-th iteration, each row from second to p–1-th is affected only once in previous iterations, and the first row and rows from p+1-th to n-th are not affected at all. In other words, we have $a_{i,j}^{(p)} = a_{i,j}^{(2)}$ and $b_i^{(p)} = b_i^{(2)}$ for $2 \le j \le p-1$, and also $a_{i,j}^{(p)} = a_{i,j}$ and $b_i^{(p)} = b_i$ for $j = 1$ or $p+1 \le j \le n$. Now, we will select $\xi_{p,j}, j = 1..p-1$ to provide the condition

$$a_{p,j}^{(p)} = 0\,,\ j = 1..p-1 \quad (21)$$

In other words, we want that $\mathbf{A}^{(p)}$ has the following shape:

$$\mathbf{A}^{(p)} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,p-1} & a_{1,p} & \cdots & a_{1,n} \\ 0 & a_{2,2}^{(2)} & \cdots & a_{2,p-1}^{(2)} & a_{2,p}^{(2)} & \cdots & a_{2,n}^{(2)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{p-1,p-1}^{(2)} & a_{p-1,p}^{(2)} & \cdots & a_{p-1,n}^{(2)} \\ 0 & 0 & \cdots & 0 & a_{p,p}^{(2)} & \cdots & a_{p,n}^{(2)} \\ a_{p+1,1} & a_{p+1,2} & \cdots & a_{p+1,p-1} & a_{p+1,p} & \cdots & a_{p+1,n} \\ a_{p+2,1} & a_{p+2,2} & \cdots & a_{p+2,p-1} & a_{p+2,p} & \cdots & a_{p+2,n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,p-1} & a_{n,p} & \cdots & a_{n,n} \end{pmatrix} \quad (22)$$

Using the matrix multiplication rule, we see that we can satisfy (21) if we take that

$$\xi_{p,j} = -\frac{1}{a_{j,j}^{(2)}} \left(a_{p,j} + \sum_{k=1}^{j-1} \xi_{p,k}\, a_{k,j}^{(2)} \right),\ j = 1..p-1 \quad (23)$$

Afterwards, the matrix multiplication rule gives:

$$a_{p,j}^{(2)} = a_{p,j} + \sum_{k=1}^{p-1} \xi_{p,k}\, a_{k,j}^{(2)}\,,\ j = p..n \quad (24)$$

$$b_p^{(2)} = b_p + \sum_{k=1}^{p-1} \xi_{p,k}\, b_k^{(2)} \quad (25)$$

Therefore, in each iteration, we need to calculate $\xi_{p,j}$, $j = 1..p–1$. Then, $\mathbf{A}^{(p)}$ and $\mathbf{b}^{(p)}$ should be calculated using (24) and (25). After n–1 iterations, i. e. when $p = n$, we have that $\mathbf{A}^{(n)}$ is triangular, so the final system $\mathbf{A}^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$ may be solved easily using the backward substitution. In fact, this system is exactly the same system that is obtained using the ordinary Gaussian elimination.

Someone may object that (24) and (25) are much more complicated that (6) and (8), as they perform much more calculations than one multiplication and one summation. However, note that in each iteration, say p-th, (24) is executed only $p–n+1$ times (for $j = p..n$) and (25) is executed only once, but (6) is executed $(n–q)^2$ times (for $i = q+1..n$ and $j = q+1..n$) and (7) is executed $n–q$ times (for $i = q+1..n$) in q-th iteration. By careful counting, we can see that the total number of performed arithmetical operations in Gauss X_p method is exactly the same as in the ordinary Gaussian elimination (concretely, $(n^3–n)/3$ multiplications and additions, and $(n^2–n)/2$ divisions are necessary to obtain the upper triangular form).

Note that it is necessary to store values $\xi_{p,j}$, $j = 1..p–1$ to to be able to calculate $a_{p,j}^{(2)}$ and $b_p^{(2)}$. These values may be stored instead of $a_{p,j}^{(p)}$, $j = 1..p–1$ as the values $a_{p,j}^{(p)}$, $j = 1..p–1$ are zeros anyway. This is a good option if we want later to change vector **b**. However, when **b** is fixed, even better alternative is to store these values into an one-dimensional vector, as values of $\xi_{i,j}$ from previous iterations (i. e for $i = 2..p–1$) are not used in the p-th iteration. This may considerably speed up the calculation, as the access to the one-dimensional vector is usually much faster that the access to the two-dimensional matrix, especially in programming languages that does not allow usage of pointer arithmetic to speed up access to the elements of the matrix (like FORTRAN, BASIC, Pascal, Java, etc.).

Now we can see why the X_p method is faster than the ordinary Gaussian elimination. In the ordinary Gaussian elimination, the innermost loop consists of calculations (6) and (8). Assuming that $\xi_{i,q}$ is kept in a single variable during the execution of the innermost loop, (6) requires three accesses to the storage memory (one reading and one writing of $a_{i,j}^{(q)}$ and one reading of $a_{q,j}^{(q)}$). In total, we have exactly $n^3–n^2/2–n/2$ accesses to the elements of the matrix **A** and $3(n^2–n)/2$ accesses to the elements of the vector **b** in ordinary Gaussian elimination before producing the upper triangular form. From the other side, the innermost loops in the X_p method are loops that calculate sums in (23), (24) and (25), not loops that calculate whole expressions (23), (24) and (25). Therefore, the innermost loops contain only two accesses to the storage memory (say, $\xi_{p,k}$ and $a_{k,j}^{(2)}$ in (24)). In total, we have $2n^3/3+2n^2–8n/3$ accesses to the elements of the matrix **A** and $n^2/2+3n/2–2$ accesses to the elements of the vector **b** to produce the upper triangular form when Gauss X_p method is used. Moreover, one of accesses to the elements of the matrix (i. e. access to $\xi_{p,k}$) may be implemented through an one-dimensional array, which additionally speeds up calculation in many programming languages. With such modification, the number of accesses to the elements of the matrix **A** is reduced to $n^3/3+3n^2/2+5n/6$, but the number of accesses to the elements of the vectors is increased to $n^3/3+n^2–2n–2$ (the total number of memory accesses remains the same). Of course, the calculation of the sum requires an accumulator variable, but nearly all modern compliers will allocate such variable in a register. Even if we work in a programming language which does not perform such optimization, the access to the single variable is much faster than the access to the indexed variable, especially when two-dimensional arrays are used for the storage and when the programming language does not allow pointers and pointer arithmetic to speed up the access.

We tested the described approach in Visual Basic, Pascal and C (using the straightforward programming approach), and we get the speed up factor from 1.3 to 1.8, depending of the actual language and the size of the system). Of course, it is not possible to compare the speed of the method with the built-in procedures for equation solving that are implemented in software packages like MATLAB, as they are written in assembly language or using highly-optimizing C compilers. However, we tried the described approach in the MATLAB's internal interpreted/pcode-compiled language, and we also detected the speed up factor of about 1.5.

Someone can object that the X_p method is quite similar to the Crout's method for the LU factorization. Really, these two methods are close relatives, as Crout's method calculates elements $l_{i,j}$ and $u_{i,j}$ of matrices **L** and **U** in the LU factorization of matrix **A** using formulae that are quite similar to (23) and (24):

$$u_{i,j} = a_{i,j} - \sum_{k=1}^{i-1} l_{i,k}\, u_{k,j}\ ,\ i = 1..n,\ j = i+1..n \quad (26)$$

$$l_{j,i} = \frac{1}{u_{i,i}}\left(a_{j,i} - \sum_{k=1}^{i-1} l_{j,k}\, u_{k,i} \right),\ i = 1..n,\ j = i+1..n \quad (27)$$

There are various orders in which the coefficients in (26) and (27) may be calculated. However, in Crout's method all of these coefficients need to be stored in the matrix, as whole matrices **L** and **U** are necessary later to solve the system $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ using both the forward and the backward substitution. Contrary, in the Gauss X_p method there is no forward substitution, as it produces exactly the same output (the upper triangular form of the system) as the ordinary Gaussian elimination, which may also be a good point from the pedagogical aspect of view. In addition, our experiments show that the Gauss X_p method is much more accurate than LU factorization when the matrix is large. Probably the main reason for such behavior is that in LU factorization all errors are accumulated on the left side of the equation $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, as **b** stays intact during the whole procedure. From the other side, in the Gauss X_p method errors are distributed on both sides of the equation.

Finally, we need to say about the pivoting in the Gauss X_p method. It is known that the implementation of the pivoting in the LU factorization methods is a bit involved and slows down the procedure. Contrary, it is quite easy to implement partial column-wise pivoting in the Gauss X_p method. Really, all we need to do is to swap p–1-th column with the column in which $a_{p-1,j}^{(2)}$, $j = p..n$ has the

largest magnitude. Of course, column swapping requires prenumeration of the unknowns from vector **x**.

IV. GAUSS X_p METHOD FOR SYMMETRIC MATRICES

One interesting feature is that the just demonstrated Gauss X_p method is especially useful when the matrix **A** is symmetric. Namely, it can be shown that under this condition the coefficients $\xi_{p,j}$ may be calculated using much simpler formula

$$\xi_{p,j} = -\frac{a^{(1)}_{j,p}}{a^{(1)}_{j,j}}\ ,\ j = 1..p{-}1 \tag{28}$$

instead of using (23). This will speed up the complete procedure by the factor slightly less than 2. Really, for $p = 2$, (23) simply reduces to (28). Suppose now that (28) holds for some $p = k$, where $2 \le k \le n-1$. We will show that (28) also holds for $p = k+1$. Really, from the the assumption that $a_{i,j} = a_{j,i}$, the inductive hypothesis and from (23) and (24), we have:

$$\begin{aligned}
\xi_{p,k+1} &= -\frac{1}{a^{(2)}_{l+1,l+1}}\left(a_{p,l+1} + \sum_{k=1}^{l}\xi_{p,k}\,a^{(2)}_{k,l+1}\right) = \\
&= -\frac{1}{a^{(2)}_{l+1,l+1}}\left(a_{l+1,p} + \sum_{k=1}^{l}\xi_{p,k}\,a^{(2)}_{k,l+1}\right) = \\
&= -\frac{1}{a^{(2)}_{l+1,l+1}}\left(a_{l+1,p} - \sum_{k=1}^{l}\frac{a^{(2)}_{k,p}}{a^{(2)}_{k,k}}a^{(2)}_{k,l+1}\right) = \\
&= -\frac{1}{a^{(2)}_{l+1,l+1}}\left(a_{l+1,p} - \sum_{k=1}^{l}\frac{a^{(2)}_{k,l+1}}{a^{(2)}_{k,k}}a^{(2)}_{k,p}\right) = \\
&= -\frac{1}{a^{(2)}_{l+1,l+1}}\left(a_{l+1,p} + \sum_{k=1}^{l}\xi_{l+1,k}\,a^{(2)}_{k,p}\right) = -\frac{a^{(2)}_{l+1,p}}{a^{(2)}_{l+1,l+1}}
\end{aligned} \tag{29}$$

Therefore, using the principle of the mathematical induction, we can conclude that (28) really holds for all $p = 2..n$.

The described optimization for the symmetric matrices looks quite simple. However, the pivotization becomes quite tricky when such optimization is used. At first, we must search for the pivot element only along the main diagonal, as in the ordinary Gaussian elimination optimized for the symmetric matrices, as described in Section 2. Suppose that in the p-th iteration the pivot element is $a_{r,r}$, $p \le r \le n$. Then we need to replace both the p–1-th column with the r-th column and the p–1-th row with the r-th row to keep the symmetry of the rest of the matrix. However, we must not overwrite the p–1-th row of the matrix, because this will destroy the part of the matrix that is already reduced to a triangular shape. So we need to find a different strategy.

To solve this problem, we will slightly change the order of calculations. Instead of calculating $\xi_{p,j}$, $j = 1..p{-}1$ using (28) in the p-th iteration, we should calculate $\xi_{i,p-1}$, $i = p{-}1..n$ using the analogous formula

$$\xi_{i,p-1} = -\frac{a^{(1)}_{p-1,i}}{a^{(1)}_{p-1,p-1}}\ ,\ i = p{-}1..n \tag{30}$$

and store them in the place where coefficients $a_{i,p-1}$, $i = p{-}1..n$ were stored. This is possible, because elements of **A** below the main diagonal are irrelevant, due to the symmetry of **A**. It is easy to show that such approach guarantees that all necessary values will be calculated before we apply (24). Assume now that we want to perform the pivotization at the beginning of the p-th iteration, and that the pivot element is $a_{r,r}$. Then, we will swap both the p–1-th column with the r-th column and the p–1-th row with the r-th row in the whole modified matrix **A**, together with stored coefficients $\xi_{i,j}$ below the main diagonal. It can be proved easily that such approach produces the correct values. In a rare occasion when all elements on the main diagonal below $a^{(2)}_{p-1,p-1}$ are zeros, we can apply similar approach as in the Gaussian elimination optimized for symmetric matrices described in Section 2.

V. CONCLUSION

This paper proposes some idea for speeding up the Gaussian elimination method for solving systems of linear equations, especially for the case when the system is symmetric, which is common case in the engineering practice. So far, there are no methods for solving systems of linear equations that have better asymptotic complexity than $\Theta(n^3)$ which are also practicable for the systems of moderate size. Therefore, we are usually stick to the Gaussian family of methods, so we need to try to improve the constant factors in such methods as much as possible. The approach given in this paper is particularly useful on the classic computer architectures and using the classic programming languages.

REFERENCES

[1] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, Numerical Recipes (The Art of Scientific Computing), 3rd ed., Cambridge University Press, 2007.

[2] V. Strassen, "Gaussian Elimination is Not Optimal", Numerische Mathematik, vol. 13, pp. 354-356, 1969.

[3] D. Coppersmith and S. Winograd, "Matrix Multiplications via Arithmetic Progressions", Journal of Symbolic Computation, vol. 9, pp. 251-280, 1990.

[4] H. W. Hayt and J. E. Kemmerly, Engineering Circuit Analysis, 5th ed., McGraw Hill, New York, 1993.

[5] G. H. Golub and C. F. van Loan, Matrix Computations, 3rd ed., Baltimore: Johns Hopkins Unversity Press, 1996.

[6] D. Watkins, Fundamentals of Matrix Computations, Wiley, New York, 1991.

[7] J. R. Bunch and B. N. Parlett, "Direct methods for Solving Symmetric Indefinite Systems of Linear Equations", SIAM Numerical Analysis, pp. 639-655, 1971.

[8] R. L. Burden, J. D. Faires, Numerical Analysis, 8th ed., Thomson Learning Inc., 2005.

[9] J. R. Bunck and L. Kaufman, "Some Stable Methods for Calculating Inertia and Solving Symmetric Linear Systems", Mathematics of Computation, Vol. 31, No. 137, pp. 163-179, 1977.

[10] J. R. Bunch, "Analysis of the diagonal pivoting method", SIAM Numerical Analysis, pp. 656-680, 1971.

LED Powered Identification Tag - Energy Harvesting and Power Consumption Optimization

G. Horvat*, D. Vinko* and T. Švedek*
* University of Osijek, Faculty of Electrical Engineering, Department of Communication, Osijek, Croatia
goran.horvat@etfos.hr , davor.vinko@etfos.hr , tomislav.svedek@etfos.hr

Abstract - This paper presents an ID tag concept which is LED powered. The photovoltaic effect of the LED array is used to harvest energy from ambient light or from light source in reader device in order to power the tag. Harvested energy is stored in a capacitor that is used to supply ID tag logic which transmits ID information using one diode of LED array which is also used for powering. The powering circuit, i.e. LED energy harvester is tested for various commercially available LEDs, to determine the most suitable LED type with respect voltage gain and current drivability. The power consumption of presented LED ID tag is optimized using Response Surface Methodology with respect to the following input parameters: frequency of operation, frequency of data transmission and the data packet length with respect to the consumed power and the transmission distance. The optimization proposes optimal input parameters for minimum power consumption and maximum operating distance, presenting a compromise between the aforementioned parameters in reference to the input variables.

Keywords: LED, Photovoltaic effect, Identification, Tag, Optimization, Response Surface Methodology, RSM.

I. INTRODUCTION

Identification is the method of identifying an item using various technologies. Before advancement in technologies identification was often performed by another person in the manners of optical recognition. Nowadays, advanced technologies are used to safely identify a user or an item, most commonly by using a Radio Frequency IDentification (RFID). Another advantage of RFID is the self powering of the tags, having no need for power supplies (i.e. batteries) [1], [2]. The common uses of RFID Tags alongside with advanced readers have enabled the widespread of the technology resulting in unwanted side effects and security breaches. For instance, a standard 125 kHz card can be read from several dozens centimetres, without contact, resulting in an unsafe way of identifying [3].

In order to explore other options of identifications this paper proposes a new concept of identification tags using LED as both the power source (photovoltaic effect) and data transmitter. Such concept combines the benefits of RFID system (self powering) and smaller dimensions of the tag as a result of LED powered concept as opposed to RFID system which uses rather large coil.

This work was sponsored by the Ministry of Science, Education and Sports of the Republic of Croatia under projects 165-0361360-3049 and 165-0362027-1479.

II. LED ID TAG – CIRCUIT DESIGN

Basic functional concept of LED ID tag consists of energy harvesting LEDs, power management circuit and microcontroller unit, Fig. 1.

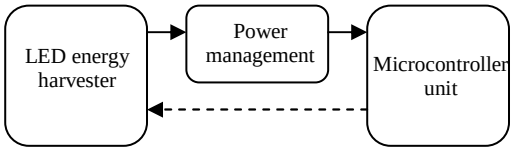


Figure 1. Functional diagram of LED ID tag

LED energy harvesting circuit consists of serially connected LEDs and a storage capacitor. LEDs, when illuminated, generate voltage which is then stored in storage capacitor. In the LED powered circuit, a LED type has a significant impact on the circuit performances and on the amount of energy that can be harvested. To evaluate powering ability of LEDs, five common LED types were tested: 5 mm orange, red and green diode, and 3 mm red and transparent/red. All LEDs were exposed to the same light source (10 W white LED) at the same distance (approx. 1 cm). By changing the resistive load parallel to LED, the current-voltage characteristics are measured, Fig. 2.

The short-circuit current of the LED photovoltaic harvester is very low. With maximum of this current around order of magnitude of 10 µA, the power consumption of the tag must be reduced as much as possible. To achieve needed voltage level for the microcontroller unit, at least five LEDs are serially connected; since a single LED can generate open circuit voltage lower than 1.6 V.

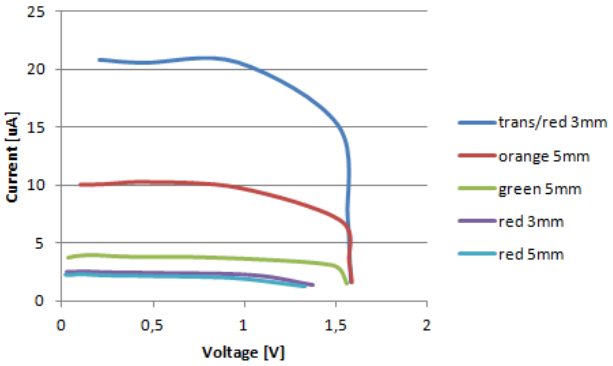


Figure 2. LED energy harvester voltage-current characteristics

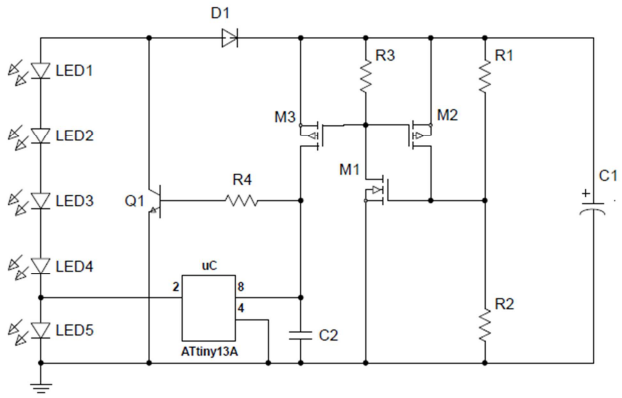


Figure 3. LED ID tag circuit

As mentioned before, with very low short-circuit current from LED energy harvester, a power management circuit is needed. Such low instantaneous power is common in energy harvesting circuits and presented power management circuit can be applied to other types of energy harvesters [4]-[7]. Fig. 3 shows the schematic of the tag. Majority of the design is power management circuit called power-on circuit, together with energy harvester (LED1 to LED5, *D*1 and *C*1) and microcontroller unit (ATtiny13A and *C*2).

Power-on circuit has the following function: it waits until voltage across storage capacitor *C*1 reaches predefined value and then it connects microcontroller unit to the power supply, i.e. storage capacitor [8]. Energy stored in capacitor *C*1 is then used to supply the microcontroller which uses LED5 from energy harvester to transmit data back to interrogation device. By doing so, the charge in storage capacitor is depleted and the charging process starts over.

With non-constant voltage value of power supply, i.e. storage capacitor voltage, a standard Schmitt trigger cannot be used as power-on circuit. Therefore, a new design shown in Fig. 3 is used. Additional problem with power-on circuit is that its power consumption slows down charging of the storage capacitor, thus degrading the system performance. To cope with aforementioned restrictions and demands on the power-on circuit the presented design uses MOS transistors to switch on the microcontroller unit. Altogether with resistive voltage divider (*R*1 and *R*2 - both in MΩ region), the circuit consumes under 2 µA.

Five LEDs, when illuminated, generate voltage which charges, through diode *D*1, the storage capacitor *C*1. Voltage divider comprised of resistors *R*1 and *R*2 sets the desired turn-on voltage. When the voltage across resistor *R*2 reaches threshold voltage of *M*1, transistor *M*1 starts conducting and it forces the gates of *M*2 and *M*3 to ground potential, thus turning them on. *M*2 short circuits resistor *R*1, further pushing *M*1 into conduction. At the same time, *M*3 connects microcontroller unit to the storage capacitor, thus turning the microcontroller on. With limited energy stored in capacitor *C*1, the microcontroller power consumption was further optimized as described in section III.

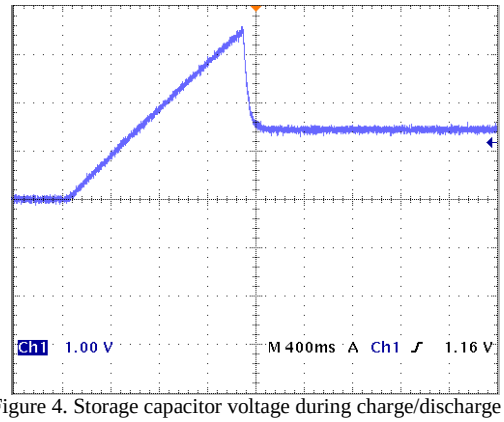


Figure 4. Storage capacitor voltage during charge/discharge cycles without NPN transistor *Q*1

NPN transistor *Q*1 is used to short-circuit LED array when the microcontroller is turned on. This is due to power-on circuits tendency to stabilize at some very low quiescent current after first charge/discharge cycle, Fig. 4.

Without NPN transistor *Q*1, the power-on circuit is able to perform just one charge/discharge cycle. With added *Q*1 which turns on at the same time as microcontroller, the low current provided from LED energy harvester is shut off, and the microcontroller can drain more charge from storage capacitor before *Q*1 turns off. This way, the voltage across storage capacitor at the end of the discharge phase is lower than voltage level required for the quiescent state. Power-on circuit can now perform continuous charge/discharge cycles, Fig. 5.

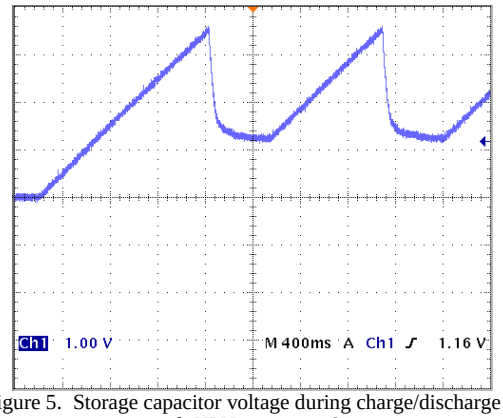


Figure 5. Storage capacitor voltage during charge/discharge cycles with NPN transistor *Q*1

III. RSM Modelling and Optimization

Response surface methodology (RSM) is a set of techniques from mathematics and statistics used for empirical model building and model optimization. RSM is mostly used for process modelling and optimization in process engineering, food technologies and mechanical engineering [9],[10] . In electrical engineering application of RSM can be found in modelling motors in electromechanical engineering but is rarely found in the area of power consumption optimization. Found usage of RSM include optimizing protocol interactions [10], optimizing call capacity in VoIP networks [11], modelling queuing systems in [12] and modelling power consumption in Wireless HART TDMA systems. Related work on power consumption is seen in [13] where energy harvesting circuit is optimised for minimum power

consumption using RSM. Guided by this work in this paper we have optimised the power consumption of a microcontroller for use in light tag identification. The parameters investigated in the means of optimising power consumption of a microcontroller are as follows:

- f_{data}: Data stream frequency: Represents the frequency of the data stream that is encoded with Manchester coding and is fed to the output LED,
- N_{bits}: number of bits: Number of bits that are transmitted and
- t_{delay}: Inter frame delay: The delay between two consecutive data packets.

The aforementioned parameters are treated as input parameters whereas the model response parameter is Power consumption represented as electrical current in µA.The chosen method of RSM optimization was the Box-Behnken (BBD) design primarily due to the choice of un-coded parameters. Other methods like Central composite design (CCD) has the disadvantage of needing input un-coded parameters that exceeds the central-point runs by factor alpha, resulting in decimal input parameters. Decimal input parameters are uneasy to handle in digital systems whereas the parameters are often integers. To avoid this shortcoming of CCD we have used the BBD as a reference design.

Box-Behnken design (BBD) was applied for determining optimal parameters for minimum power consumption. It requires N number of experiments according to the following equation:

$$N = 2k(k-1) + c_p \quad (1)$$

where the k is the factor number and c_p is the replicate number of the central point. Three levels of design exist (-1, 0, +1), with equally spaced intervals between levels [9]. The design parameters from the model must be expressed in their corresponding physical units and must be converted into dimensionless quantities with zero mean and the same standard deviation before proceeding with the RSM modelling. These quantities are called coded variables. The coding of the variables is performed by the transformation process described by the equation (5) [13].

$$x = \frac{a-[a_{min}+ a_{max}]/2}{[a_{max}+ a_{min}\ ]/2} \quad (2)$$

where the a is the natural variable, x is the coded variable and a_{min} and a_{max} are the maximum and minimum values used for the natural variable, respectively. Now the approximate function y is expressed in term of coded variables, and determining a model for y determines the success of the RSM model [13]. Most of the engineering data can be fitted with a second order response surface model [10]. Also, upon statistical analysing the measurement data the fit summary suggested a second order polynomial, with p value of 0.0001. Consequently, the second order polynomial of the following form was chosen for the purpose of analysis:

$$\hat{y} = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} {x_i}^2 + \sum\sum_{i<j} \beta_{ij} x_i x_j \quad (3)$$

where y is the response, β_0, β_i, β_{ii} and β_{ij} are constant coefficients of intercept, linear, quadratic and interaction terms, respectively. x_i and x_j are the coded independent variables. Determining of the coefficients is performed using commercial software Design-Expert® v. 8.0.7.1. [14]. Analysis of variance (ANOVA) was used to evaluate the quality of the model. The test of statistical difference was based on total error criteria with confidence level of 0.95 (or 95%).

A. Modelling and optimizing the power consumption of a LED power identification tag

This analysis was chosen primarily to model and optimize power consumption of energy harvesting tag based on LED energy harvester. Input factors and their ranges are shown on Table I.

TABLE I. Factors and values of levels for each factor

Factor	Low (-1)	Middle (0)	High (+1)
f_{data} [kHz]	2	5	8
N_{bits} [bit]	32	64	96
t_{delay} [ms]	50	200	350

The ranges are chosen according to the specification of similar RFID communication protocols, i.e. EM4100 where the range of data frequency was selected from available clock division factors in [15]. For number of bits, Middle factor was selected as a reference factor from [15] and High and Low factors are chosen as adjacent available values.

For model measurement responses we have chosen average power consumption [µA] measured by Agilent U1273A True RMS Multimeter. For the set of input factors the generation of experiment matrix is performed by Design-Expert software. From (1) it is evident that 17 measurement runs are needed in order to generate the model. These parameters alongside with measurement results are shown in design matrix on Table II.

TABLE II. Experimental matrix and the measurement values

Run	f_{data} [kHz]	N_{bits} [bit]	t_{delay} [ms]	I[µA]
1	3,5	64	200	251
2	3,5	96	50	523
3	5	96	200	277
4	3,5	64	200	246
5	5	64	350	190
6	2	64	350	223
7	3,5	32	50	344
8	5	64	50	436
9	5	32	200	180
10	2	32	200	207
11	3,5	32	350	159
12	3,5	64	200	249
13	3,5	96	350	231
14	2	96	200	340
15	3,5	64	200	252
16	3,5	64	200	251
17	2	64	50	501

The effect of linear, quadratic or interaction coefficients on the response was tested for significance

by analysis of variance (ANOVA). Regression coefficients of intercept, linear, quadratic, and interaction terms of the model were calculated using the least square method. The degree of significance of each factor is represented in Table III by its *p*-value.

TABLE III. *P* values of the response alongside with *p*-values.

Source	Sum of Squares	Mean Square	*F* Value	*p*-value Prob > F
Model	187944,6	20882,74	331,2842	< 0.0001
A-f_{data}	**4418**	**4418**	**70,08725**	**< 0.0001**
B-N_{bits}	**28920,13**	**28920,13**	**458,7895**	**< 0.0001**
C-t_{delay}	**125250,1**	**125250,1**	**1986,971**	**< 0.0001**
AB	324	324	5,139943	0.0577
AC	256	256	4,06119	0.0837
BC	**2862,25**	**2862,25**	**45,4068**	**0.0003**
A^2	725,3289	725,3289	11,50663	0.0116
B^2	431,6447	431,6447	6,847622	0.0346
C^2	**24560,59**	**24560,59**	**389,6298**	**< 0.0001**
Residual	441,25	63,03571		
Lack of Fit	**339,25**	**113,0833**	**4,434641**	**0.0921**
Pure Error	102	25,5		

As seen from Table III the most contributing factors of the power consumption al all the input parameters! This was expected since the input parameters were chosen as parameters with most impact on power consumption. This can be concluded from the *p*-value in the sense that when the p-value is less than 0.05, that factor has a significant influence on the response (for confidence level of 95%).

Further on, the Lack of fit value larger than 0.05 represent non-significant lack-of-fit, as seen from Table III. Accordingly, the surface plots of the responses against Data frequency - f_{data} [kHz], Inter Frame delay - t_{delay} [ms] and Number of bits - N_{bits} [bit] are shown on Fig. 6 - 7.

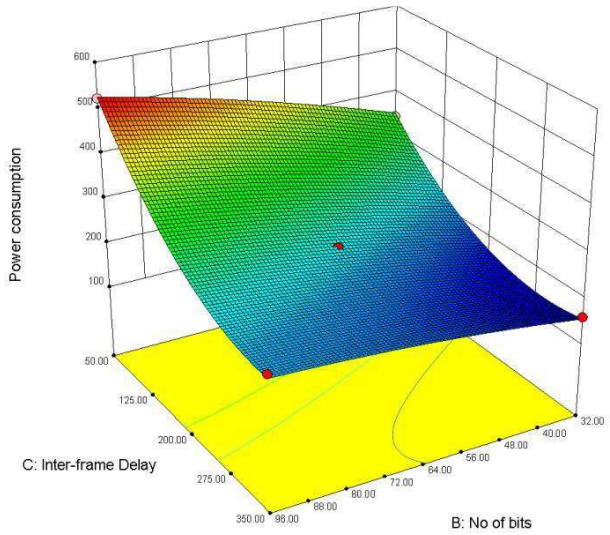


Figure 6. Surface plot of Power consumption [µA] against Inter frame Delay [ms] and No of bits [bit] for Data Frequency = 3.5 kHz.

From ANOVA it is possible to model the power consumption of a microcontroller using analysis of Variance (ANOVA) and least square method. The resulting mathematical model of the microcontroller power consumption is as follows:

$$I\ [\mu A] = 427.71 - 51.61 \cdot f_{data} + 4.92 \cdot N_{bits} - 1.96 \cdot t_{delay} - 0.19 \cdot f_{data} \cdot N_{bits} + 0.036 \cdot f_{data} \cdot t_{delay} - 5.57 * 10^{3} \cdot N_{bits} \cdot t_{delay} + 5.83 \cdot f_{data}^{2} - 9.89 \cdot 10^{3} \cdot N_{bits}^{2} + 3.39 \cdot 10^{3} \cdot t_{delay}^{2} \quad (4)$$

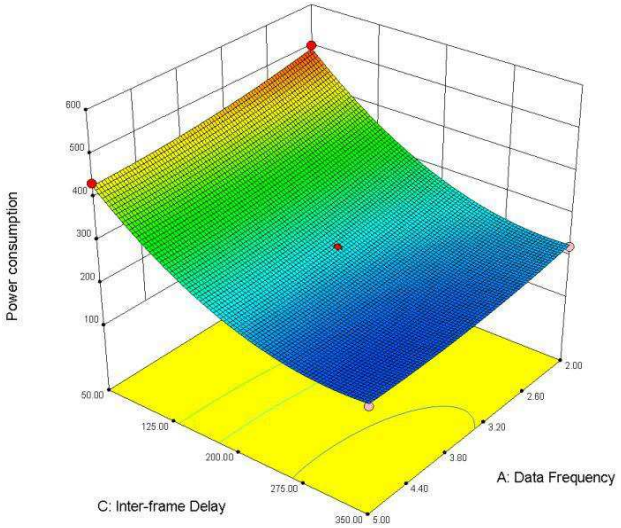


Figure 7. Surface plot of Power consumption [µA] against Inter frame Delay [ms] and Data Frequency [kHz] for No of bits = 64.

The optimization of the power consumption is performed by maximising the desirability function. The surface plot of the desirability function is shown on Fig. 8. From Fig. 8 it is clear that the optimal parameters for:

- minimum Power consumption - I
- minimum Inter frame delay - t_{delay}
- maximum Number of bits - N_{bits}

are obtained when the desirability function reaches the value of 0.705.

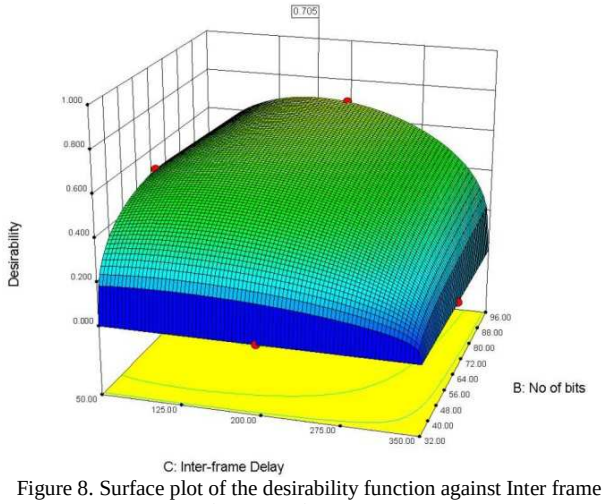


Figure 8. Surface plot of the desirability function against Inter frame Delay [ms] and No of bits [bit] for Data Frequency [kHz] = 5kHz.

The reason for setting these criteria is the following: the power consumption needs to be minimised in order to power the microcontroller for longest possible time; Inter frame delay needs to be minimised in order to send as much data packets as possible (to eliminate the possibility of a corrupted data packet); Number of bits needs to be as high as possible in order to ensure the high security of the tag. The optimized values of input parameters are shown on Table IV, alongside with the predicted responses.

TABLE IV. Optimal parameters for minimum power consumption, maximum Number of bits and Minimum inter frame delay.

Optimized parameters	**Values**
Data frequency -t_{data}	5 kHz
Number of bits - N_{bits}	96 bit
Inter Frame delay - t_{delay}	169 ms
Estimated response	**Value**
Power consumption – I	311.6 µA

IV. MEASUREMENTS

In order to validate the estimated microcontroller consumption response, the measurements of LED ID tag operation were performed. The power consumption measurements reveal the average electrical current consumption to be in the range of 366 µA, representing a deviation of approximately 15%. If compared to maximum values of 525 µA and minimum values of 159 µA, the measured optimized value of electrical current is within the range of the estimated value. Next on, a series of measurements is shown for voltage values of energy harvesting circuit.

As seen in Fig. 9, for charge/discharge cycles with 10 µF storage capacitor, the voltage across storage capacitor (dark blue trace) and voltage across microcontroller unit (light blue trace) are measured. Due to protective reverse biased diodes on pin 2 of microcontroller which is connected to LED5, the negative voltage value of approximately 300 mV appears across microcontroller. This can also be noticed on Figs. 10 to 12. The reference point (0 V) in Figs. 9 to 12 is set on the central horizontal line for both traces, as can be seen at the beginning of the plot on Fig. 9.

During discharge cycle microcontroller is using one LED (LED5 from Fig. 3) to transmit data. Transmitted data stream (light blue trace on Fig. 10) consists of 96 bits of information which are Manchester coded, Fig. 10.

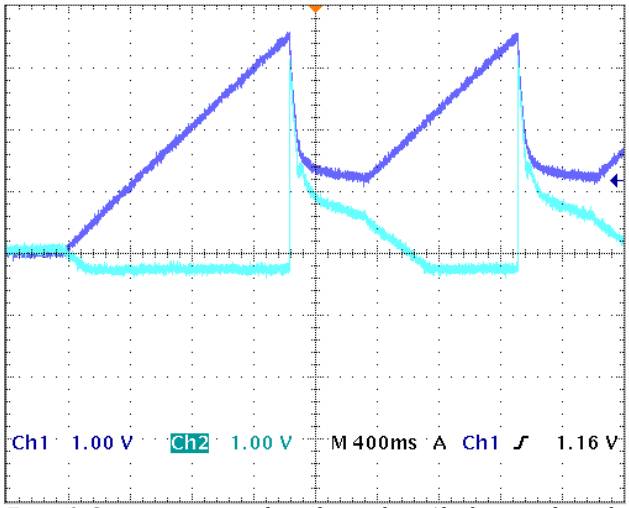


Figure 9. Storage capacitor voltage during charge/discharge cycles with microcontroller voltage plot (light blue trace), $C1$ = 10 µF

The impact of storage capacitor discharging can also be seen on transmitted data stream trace which is actually a voltage applied to LED5.

One method to ensure the complete data transmission without data stream degradation due to storage capacitor voltage drop is to increase the capacitance of storage capacitor. Fig. 11 shows the date stream (voltage across LED5) with 100 µF storage capacitor. When compared to Fig. 10 it can be seen that beside the data stream degradation, the energy in storage capacitor was not sufficient to transmit the entire set of 96 bits.

In addition, with 100 µF storage capacitor, stored energy is sufficient to transmit a several consecutive data stream sequences, Fig. 12. When the data is transmitted, the consumption of the tag is higher due to operating LED5. This can be seen as slightly steeper slope of storage capacitor voltage in time intervals when the data stream is transmitted.

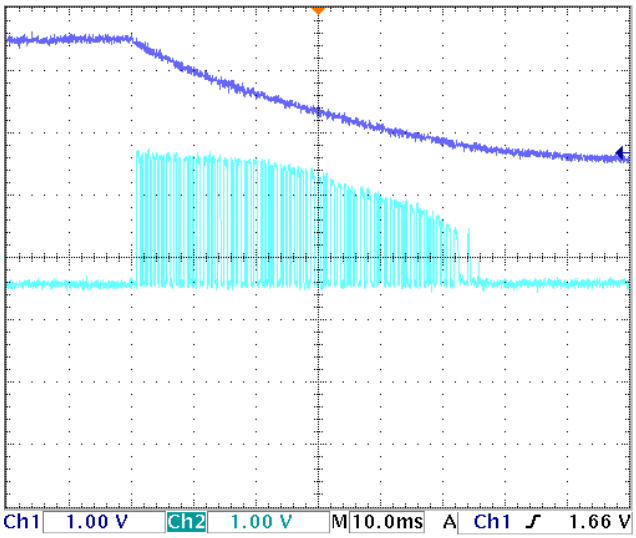


Figure 10. Storage capacitor voltage (dark blue trace) and LED5 voltage – data stream (light blue trace), $C1$ = 10 µF

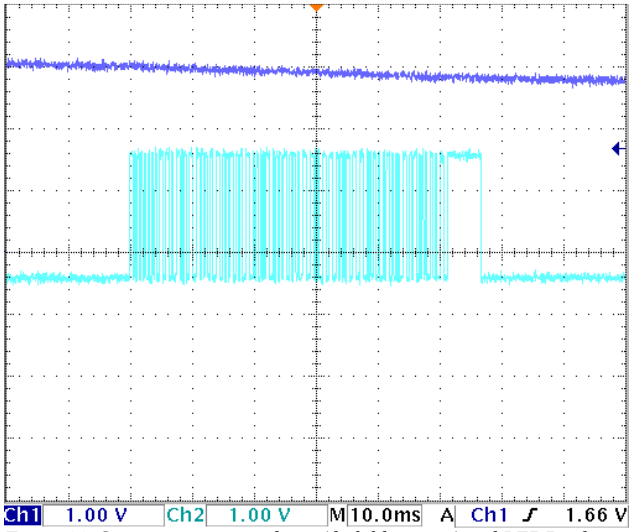


Figure 11. Storage capacitor voltage (dark blue trace) and LED5 voltage – data stream (light blue trace), $C1$ = 100 µF

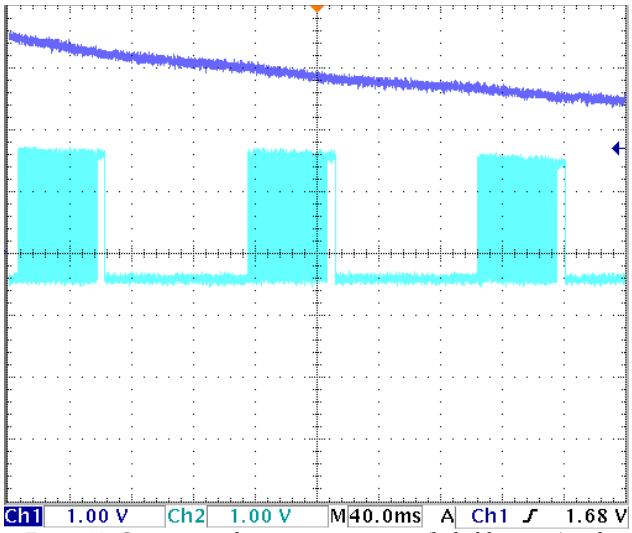


Figure 12. Consecutive data stream sequences (light blue trace) and storage capacitor voltage (dark blue trace), $C1$ = 100 µF

V. CONCLUSION

In this paper a concept of LED powered identification tag was presented. With characterization of common LED types as energy harvesters (using photovoltaic effect), the order of magnitude of available power for energy harvesting is determined. Accordingly, the power management circuit is developed and measurements on discrete realization of the circuit show maximal instantaneous current consumption less than 2 µA. Power management circuit is designed to enable continuous charge/discharge (i.e. energy harvesting/energy consumption) operation without settling at quiescent state.

In order to power a microcontroller from energy harvesting circuit as long as possible and to ensure the data integrity and high security of the identification process, an optimisation method was performed. The optimisation was performed using Response Surface Methodology and Box-Behnken Design where three parameters were chosen as inputs (data stream frequency, number of bits and inter frame delay) alongside with one response (Power consumption expressed in µA). The optimisation presented optimum values of input parameters for minimum power consumption, maximum data bits (security) and minimum inter frame delay (higher data integrity by means of retransmission).

Finally, measurements on optimized tag show that it is possible to use LED power harvesters to power a microcontroller and an output LED, alongside with the control of the power-on circuit. For larger capacitor values several consecutive data streams sequences can be transmitted but the time to charge the capacitor is also larger. This is a compromise that has to be taken into account when designing a circuit.

Overall, this paper presents a concept of LED powered identification tag that has the potential for development in the future work. One of the open issues for future work remains the choice of energy storage capacitor size and the number of LEDs for energy harvesting. The increased number of LEDs increases the available energy but it also increases the size of the tag. The size of the capacitor will be optimised for minimum charge time and maximum number of data stream sequences. Also, the communication protocol for the exchange of data between tag and light transmitter needs to be developed in the sense that the transmitter needs to have a duty cycled transmission in order for the tag to successfully transmit the data towards the light transmitter. Also, the problem of light beam direction needs to be analysed in order to estimate the required angle for best data reception and highest amount of harvested energy.

REFERENCES

[1] S. Kim, J.-H. Cho, S.-K. Hong: „A Full Wave Voltage Multiplier for RFID Transponders“, IEICE Trans. Commun., Vol. E91-B, No.1, January 2008 (388-391)

[2] H. Aziza, E. Bergeret, A. Perez, J.-M. Portal: „An Efficient Model to Evaluate the Impact of Design Parameters on Charge Pump Circuits' Performances: Application on RFID Circuits“, *Proc. of the International Conference on Design & Technology of Integrated Systems in Nanoscale Era*, DTIS 2007 (108-113)

[3] B. Arazi, “Open problems pertaining to RFID anti-cloning and some observations”, Dept. of Computer Engineering and Computer Science, University of Louisville, Louisville, KY 40292, March 24, 2007.

[4] T. T. Le, A. von Jouanne, K. Mayaram, T. S. Fiez: „Piezoelectric Micro-Power Generation Interface Circuit“, *IEEE Journal of Solid-State Circuts*, Vol. 41, No. 6, June 2006 (1411-1420)

[5] N. Soltani, F. Yuan: „A Step-Up Transformer Impedance Transformation Technique for Efficient Power Harvesting of Passive Transponders“, *Microelectronics Journal*, 41, 2010 (75-84)

[6] R. J. M. Vullers, R. van Schaijk, I. Doms, C. V. Hoof, R. Mertens: „Micropower Energy Harvesting“, *Solid-State Electronics*, 53, 2009 (684-693)

[7] B. C. Yen, J. H. Lang: „A Variable-Capacitance Vibration-to-Electric Energy Harvester“, *IEEE Transactions on Circuits and Systems-I Regular Papers*, Vol. 53, No. 2, February 2006 (288-295)

[8] F. Kocer, M. P. Flynn: „A New Transponder Architecture With On-Chip ADC for Long-Range Telemetry Applications“, *IEEE Journal of Solid-State Circuts*, Vol. 41, No. 5, May 2006 (1142-1148)

[9] Raymond H. Myers, Douglas C. Montgomery, Christine M. Anderson-Cook, Response Surface Methodology: Process and Product Optimization Using Designed Experiments, 3rd Edition, Wiley January 2009, ISBN: 978-0-470-17446-3.

[10] Vadde, K.K.; Syrotiuk, V.R.; Montgomery, D.C.; , "Optimizing protocol interaction using response surface methodology," Mobile Computing, IEEE Transactions on , vol.5, no.6, pp. 627- 639, June 2006, doi: 10.1109/TMC.2006.82

[11] Kulkarni, V.; Devetsikiotis, M.; , "Cross-Layer Response Surface Methodology Applied to Wireless Mesh Network VoIP Call Capacity," Simulation Symposium, 2008. ANSS 2008. 41st Annual , vol., no., pp.15-22, 13-16 April 2008, doi: 10.1109/ANSS-41.2008.23

[12] Kemper, P.; Muller, D.; Thummler, A.; , "Combining response surface methodology with numerical models for optimization of class-based queueing systems," Dependable Systems and Networks, 2005. DSN 2005. Proceedings. International Conference on , vol., no., pp. 550- 559, 28 June-1 July 2005, doi: 10.1109/DSN.2005.28

[13] Wang, L.; Kazmierski, T.J.; Al-Hashimi, B.M.; Aloufi, M.; Wenninger, J.; , "Response-surface-based design space exploration and optimisation of wireless sensor nodes with tunable energy harvesters," Design, Automation \& Test in Europe Conference \& Exhibition (DATE), 2012 , vol., no., pp.733-738, 12-16 March 2012

[14] Design-Expert® Data Analysis Software System, v. 8.0.7.1, Stat-Ease Inc., Minneapolis, MN, USA (2007).

[15] EM4100 RFID - Transponder Protocol Specifications

Self powered current acquisition system with wireless data transfer

V. C. Valchev*, A. St. Marinov * and E.N. Dimitrova*
* Technical University of Varna/Department of electronics and microelectronics, Varna, Bulgaria
*vencivalchev@hotmail.com, **a.marinov@tu-varna.bg, ***edimitrowa@yahoo.com

Abstract - **The current paper suggests a structure for smart electrical energy meter for (but not limited to) households. The advantage of the suggested structure is that it allows energy metering at different consumers, where the measuring units derive power from their measuring transformer, omitting the need to break the electrical line of the consumer at the measuring point. Simulation verification and experimental validation are presented. They show and describe the functionality of the suggested structure.**

I. INTRODUCTION

The term smart grid addresses electrical grids that benefit from modern information and communication technologies. Smart grids include a multitude of intelligent electronic devices that are implemented to aid and improve generation, transmission and consumption of electrical energy. Those electronic devices can augment the reliability, efficiency, sustainability and others major parameters of the electric grid. General functions of the smart grid electronic devices include monitoring and smart metering, advance control, decision support, integrated communications, etc. [1]

Smart grids can be implemented, as aforementioned throughout the whole chain of electric distribution - generation, transmission and consumption. This gives a wide variety of electronic devices that, based on the level that they operate, can work under high, medium and low voltages. The current paper addresses the last category, through a discussion on smart energy metering for household applications [2].

Households are a major consumer of electrical energy, with a total share for the European Union (EU) of about 28% [3]. Thus it is important to be able to accurately predict consumption of this particular branch of consumers in order to form adequate demand response and flexible load adjustment [4].

Electrical energy consumption is mostly measured for each household using a single central energy meter. This however provides information only on the monthly consumption of electrical energy. In-depth information on household energy consumption can be obtained through smart metering devices. Those devices allow monitoring of the energy consumption of each major electrical consumer in the household and in most cases provide data logging that can track the hourly behavior of the consumer throughout the month. This detailed information can be very useful to both the consumer and the distributer, and can lead to many significant benefits [5,6,7].

The current paper suggests a structure for smart electrical energy meter for (but not limited to) households. The advantage of the suggested structure is that it allows energy metering at different consumers, where the measuring units derive power from their measuring transformer, omitting the need to break the electrical line of the consumer at the measuring point.

II. SUGGESTED STRUCTURE

The suggested structure of the smart energy metering unit is presented at figure 1. The smart meter includes a multitude of remote measurement stations placed at the consumers that will be observed. Each remote station measures the average rectified current of its designated consumer. Information about the measurement is then sent wirelessly to a central station that measures the voltage and the total consumed current. The central station can then calculate the total power, consumed energy and the average power of each consumer. In this way the total share of consumed power for each of the monitored consumers can be deduced.

The novelty of the suggested structure is the design of the remote measurement station. Assuming further discussions are made for figure 1 - "remote station 1", the principle of operation of the designed topology can be explained as follows:

Each remote station has a current transformer CT1 that measures the current drawn by the observed consumer. The secondary winding of the current transformer is connected to center tapped rectifier. The rectified current then, assuming that initially switch S1 is in position "1", charges capacitor C1 trough the resistor R1. When the voltage on the capacitor C1 reaches the required power supply voltage of the wireless transmitter, S1 is switched into position "2". Thus the charge of C1 is used to power the wireless transmitter. The transmitter then sends its corresponding address. When the message is send, S2 is switched and the remaining charge in capacitor C1 is dissipated in resistor R3. S1 goes back into position "1" and the process is repeated. During the time when S1 is in position "2", the current transformer is loaded trough the

Bulgarian scientific fund

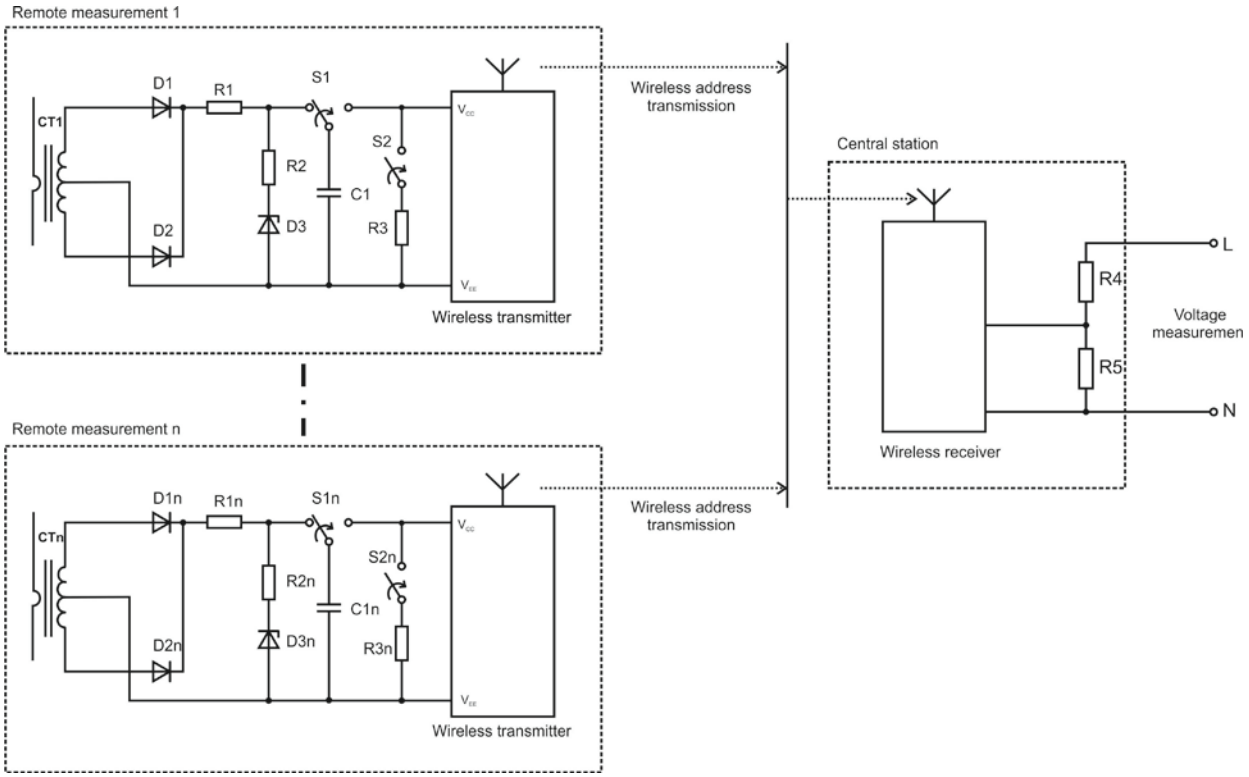


Figure 1. Structure of the suggested smart energy meter

zener diode D3 and resistor R2. The stabilizing voltage of the zener diode is selected slightly above the voltage at which capacitor C1 is being charged. Assuming this principle of operation the central station will receive the address of the corresponding remote station each time when the capacitor is charged to its designated voltage. Thus the time interval between receiving each address will correspond to the rate at which the capacitor is being charged. The rate of charging however will be directly proportional to the size of the capacitor, the voltage limit at which switch S1 is turned to position "2" and the average rectified current. In this way the only "unknown" is the average current, which can be calculated by the central station based on the measured time between received addresses and the known capacity and charge voltage of C1.

In this way the following benefits of the suggested structure can be noted:

- If switches S1 and S2 are realized using a set of static electronic switches it can be assumed that the total charge of the capacitor will be used to power the wireless transmitter, which could be low power device that would not require any additional power. In this way the remote station will be power only by the current transformer. This gives the benefit of total galvanic insulation and no disconnection in the measured line.
- The data rate could be very low - information is derived only from the received addresses and depends on the magnitude of the measured average current. This is an advantage when considering electromagnetic intensity in households, caused by radio transmission of wireless devices.
- The data will have a very high level of encryption, since it can only be decoded when the values of the capacitor and its charging voltage are known.

Further on the suggested structure is tested in simulation and partial experiment.

III. SIMULATION RESULTS

A dedicated simulation model, used to test the suggested structure is developed. The model includes two parts - the remote station and the central station - considering only one remote station is analyzed.

The remote station includes the following blocks:

- A model of the current transformer, where for the simulation primary to secondary transformer transfer ratio of 1:100 is used.
- The center tapped diode rectifier, where the diodes are modeled as real schottky diodes.
- Switch S1 from the structure from figure 1 build with two switches S1 and S2 in the developed model.
- Voltage measurement block of the capacitor that transfers data to a hysteresis relay. The relay turns "off" switch S1 and "on" S2 when the voltage on the capacitor reaches 3.3V. Inversely when the voltage at the capacitor falls at level close to zero S1 is turned back "on" and S2 is turned "off".

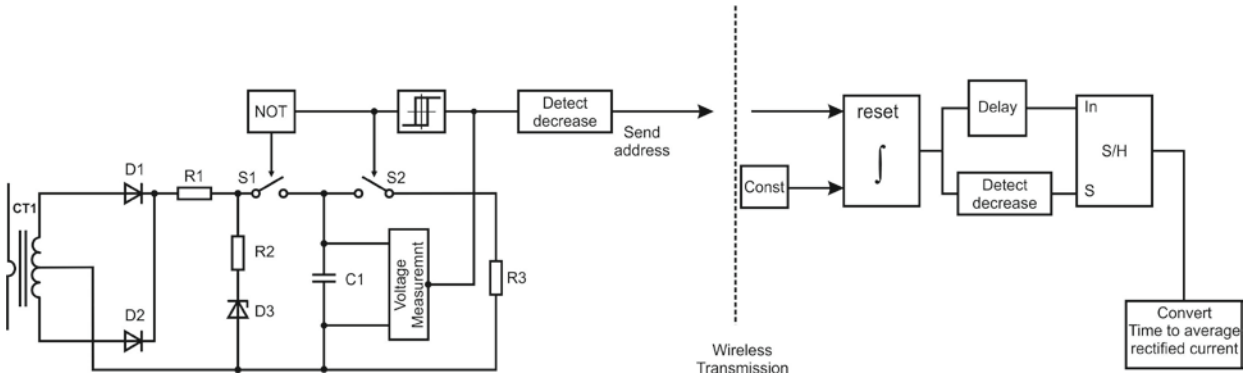


Figure 2. Structure of the suggested smart energy meter

- Detect decrease block that generates a single pulse when the capacitor is discharged.

The central station on the other hand includes:

- A resettable integrator that is controlled by the signal received by the remote station. Thus on its output the integrator provides the time between the received pulses.
- "Sample and Hold" block with control logic that holds the time value used to calculate the average rectified current.

For the initial simulation analysis the current transformer is fed with a sinusoidal current with frequency 50Hz and amplitude of 20A for duration of 3s and amplitude 10A for duration of 17s (Fig. 3). This analysis allows to test the functionality and responsiveness of the system and to evaluate its basic accuracy.

Sample waveforms from the outputs of the different blocks of the model are presented on Fig. 4 and represent the following depictions:

- The input current that is objected to the measurement.
- The voltage on the capacitor.
- The accumulated and held time between receiving the separate addresses.

The diagrams confirm the stated principle of operation, where with the last waveform shows the accuracy and responsiveness of the suggested structure can be viewed. It can be seen there is a minimal error between real value of the average rectified current and the one calculated using the suggested measurement structure. With the set parameters the responsiveness of the system is rather slow where big current changes occur.

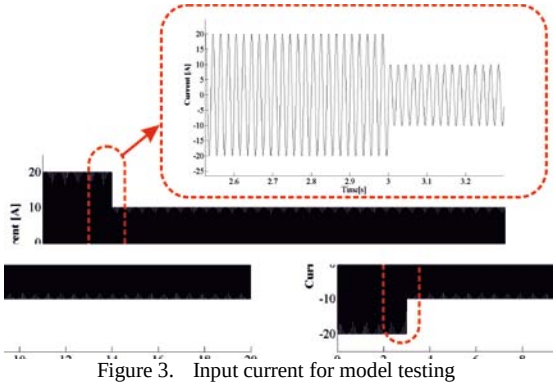


Figure 3. Input current for model testing

IV. Simulation with Experimental Data

Further studies of the suggested structure were implemented. Using the developed model, a real current is loaded instead of the sinusoidal input that was used in the previous section. This can provide information on how the suggested structure will operate under real live load. The stimulus used, is the data logged current consumed by a household refrigerator for duration of one hour - figure 5.

Results from the simulation using the experimental data, for a short duration of the repeating process that describes the operation of the refrigerator are shown on Fig. 6.

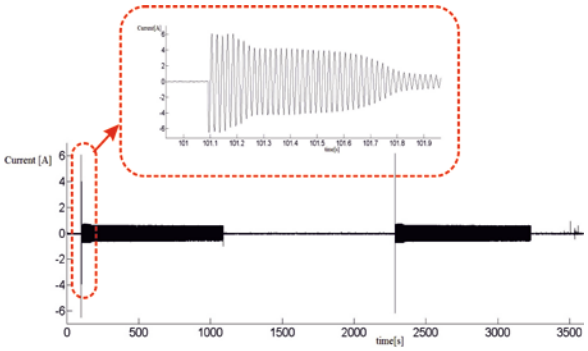


Figure 5. Experimental data for model simulation

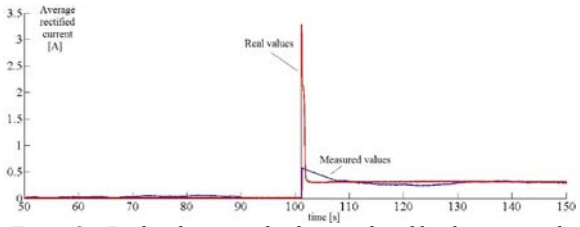


Figure 6. Real and measured values produced by the suggested structure using experimental input signal

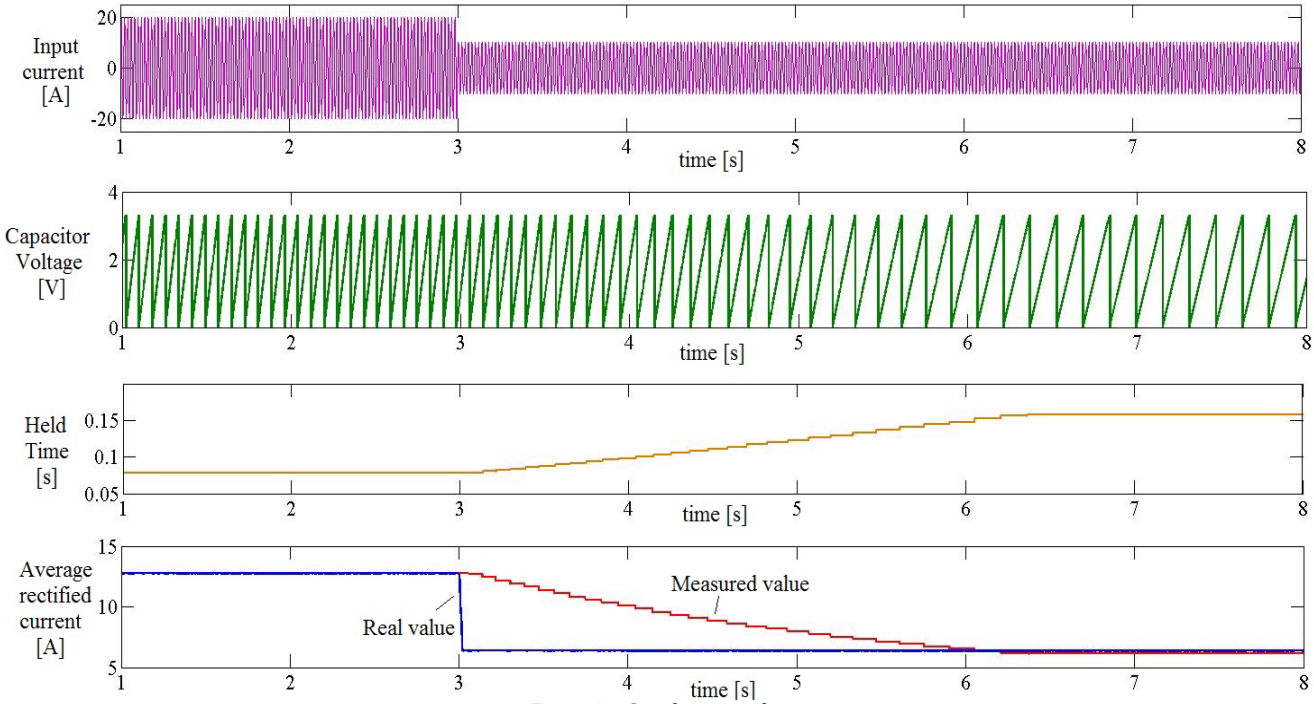


Figure 4. Simulation results

As it can be seen from the figure, the error between the measured and real current is very small.

In addition for all simulation a capacitor with capacity of 100µF is used. When charged to the voltage selected for the simulation 3.3V the charge of the capacitor will equal Q=330µC. This charge should be sufficient to drive the wireless transmitter, making the probe "selfpowered" - power is derived from the measurement itself.

V. DESIGN CONSIDERATIONS

A. *Calculating and logging the active power for each of the observed devices*

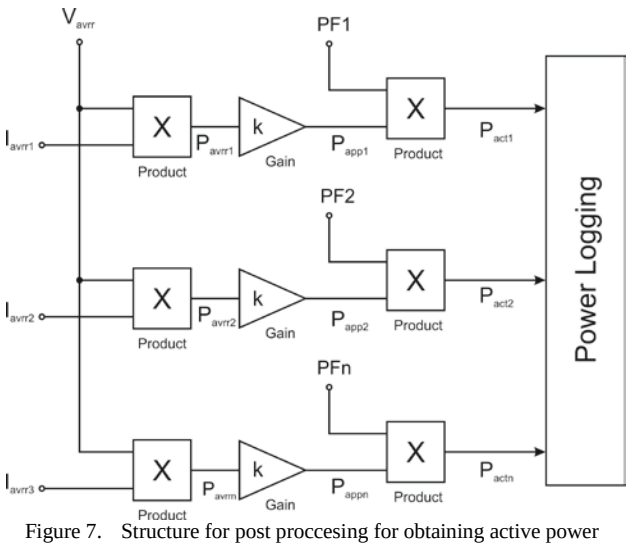


Figure 7. Structure for post proccesing for obtaining active power

Data recieved by each separate meter and processed by the central measurement unit provides the average rectified current, as described and presented above. In this way information about the phase of the current of each device, relative to the voltage is lost. This for power logging will result in being able to retrieve only the average rectified power of each of the observed devices. Consequently the apparent power of each device can be calculated. What presents interest for domestic users however, is the active power, since this is the power that is billed by electric companies.

Suggestion for post-processing for obtaining the active power using the measured average rectified current is depicted in figure 7. The figure presents a system with "n" members (remote stations) where:

- Currents I_{avrr1}, I_{avrr2},....I_{avrrn} are the average rectified currents acquired trough the measurement of the remote stations.
- Voltage V_{avrr} is the grid average rectified voltage measured by the central station.
- Powers P_{avrr1}, P_{avrr2},....P_{avrrn} are the average rectified powers obtained as a product of the voltage V_{avrr} and each corresponding current I_{avrr1}, I_{avrr2},....I_{avrrn}
- Coefficient k that acts as a gain to the average rectified powers in order to obtain the apparent powers of each device - P_{app1}, P_{app2},....P_{appn}. For sinusoidal waveform of the current and the voltage k as the relation between average rectified and apparent power is k≈1.232.
- Power factor specific to each device PF1, PF2,... PFn.
- Retrieved active power of each specific device - P_{act1}, P_{act2},....P_{actn}. Obtained as a product of the measured apparent power and the set power factors

In the suggested structure the nominal power factor of each device has to be set upon initializing the system. This means that for the central measurement station an initial data input for each consumer has to be set. The data input can be a short description of the observed device so a power logging record can be created and its nominal power factor provided in specification. In this way based on the measured average rectified current and the calculated apparent power, the active power contribution of each device can be retrieved.

Using the structure from figure 5 will mean that each remote station has to be fixed to a certain consumer, rather than a non-described group of consumers.

B. Establishment of a wireless data transfer with multiple metering deviced

Realization of the system using multiple measurement points will require the implementation of a specialized wireless communication protocol. This protocol will have to allow data exchange between the remote station and the central station without loss of information when two or more simultaneous send request occur. Furthermore the data transfer has to be completed within the power limit set by the measurement capacitor.

For an example a simple star network configuration can be used where each member is a radio frequency transceiver. Upon power-up the remote station could listen in and if no communication is detected a data transmit can begin - the device can wait until power is lost. After each transmission the device can prompt for acknowledge from the central station, which if not received will initiate a consequential data transmit, until power is lost or acknowledge is received. The communication process and switching within the device can be controlled trough a low power programmable logic such as PAL.

Due to the low data rate of the suggested structure of the remote station and relatively low number of members within the communication network it is expected that data losses due to lack of power when simultaneous communication occurs will be minimum.

C. Errors and uncertainties in the system

Several possible errors and uncertainties can be expected within the suggested system. They can be described as:

- Error due to the averaging of the system - the suggested structure averages the measured current which means that sudden changes and surges in the current may not be registered or significant measurement error could be present.
- Error due to lost data during wireless communication - as explained in the previous section some information loss can occur when a remote station is unable to transmit within the energy limits set by the charge accumulated in the capacitor.
- Error due to the usage of nominal power factor for the calculation of the active power. The suggested mean of retrieval of the average power involves the usage of the nominal power factor. Household devices that are the main object of observation however often operate outside their nominal power factor which can bring an uncertainty in the measurement.
- Error due to degradation of components - it could be expected that an uncertainty in the measurement can occur due to the degradation of the capacitor in the remote stations. The capacitor is loaded with a with a significant AC component and is expected to be electrolytic.

The expected errors and uncertainties in the measurement are not yet fully evaluated for the current system. Practical realization and experiments are planned as a future study that will examine them in detail and further evaluate the functionality and accuracy of the suggested powerless measurement structure.

VI. CONCLUSIONS

A structure of a specialized smart energy meter is presented. The structure is loosely based on conversion of current to frequency, where the charge accumulated during the conversion is used to power the remote measurement stations. The presented simulation and experimental based study confirm the stated principle of operation and show the potential functionality of the suggested structure. Further analysis is required to determine transmission error. This will be done in further study where a practical realization of the suggested structure will be developed.

ACKNOWLEDGMENT

This paper is prepared in the frames of Project MU03/164 - “Intelligent systems for energy management and control of consumer expenses”, Ministry of Education Youth and Science, Bulgarian National Science Fund.

REFERENCES

[1] Aman, S., Simmhan, Y., Prasanna, V., " Energy management systems: State of the art and emerging trends", IEEE Communications Magazine, Volume 51, Issue 1, 2013, Article number6400447, Pages 114-119

[2] Youjie, M. , Dalong, D. , Xuesong, Z., " Survey on technology of the smart grid", 2nd International Conference on Applied Mechanics, Materials and Manufacturing, ICAMMM 2012, Changsha, 17 November 2012

[3] “Country factsheets”, European commission, Energy, 2011

[4] Gungor, V.C., Sahin, D., Kocak, T., Ergüt, S., Buccella, C., Cecati, C., Hancke, G., "Smart grid and smart homes: Key players and pilot projects", IEEE Industrial Electronics Magazine, Volume 6, Issue 4, 2012, Article number6378528, Pages 18-34I.

[5] Lee, J., Park, G.-L., Han J., “Integrative in-home display development for smart places” 7th International Conference on Ubiquitous Information Technologies and Applications, CUTE 2012, Hong Kong, December 2012, Code94342

[6] Landi, C. , Del Prete, G., Gallo, D., Luiso, M., “The use of real-time instruments for smart power systems, 2012 IEEE International Energy Conference and Exhibition, ENERGYCON 2012, Florence, 9 September 2012

[7] D. Muñoz, D. Pérez, Design and experimental verification of a smart sensor to measure the energy and power consumption in a one-phase AC line, Measurement, Elsevier, vol. 42, 412–419.

Investigation of an intelligent power supply system for an electric heater furnace with silicon carbide heaters

Georgi Nikolov*, Borislav Dimitrov**, Vencislav Valchev*
* Technical University Varna, Department Electronics and microelectronics, Varna, Bulgaria
** Technical University Varna, Department Electrical Engineering and electro technologies, Varna, Bulgaria
Georgi.nikolov@tu-varna.bg, bdimitrov@processmodeling.org, vencivalchev@hotmail.com

Abstract - This paper is focused on the development and investigation of power supply system for an electric heater furnace with silicon carbide heaters. The goal is to present a new solution based on a microcontroller embedded system with power electronic converters. By using the proposed system are solved some of the problems, frequently occurring when low frequency power transformer is used - uneven distribution of the temperature inside the furnace and shorter life of the heating elements. Additional features like direct readout of the supplied voltage, current and power, set and actual temperature, computer interface and others can be easily realized.

I. INTRODUCTION

High temperature electric heater furnaces with silicon carbide heating elements are mainly used for heating, melting and hardening of different materials [1]. Their main disadvantage is the aging - the increase of the heater resistance with time (when used). Traditionally this leads to decrease of the temperature inside the furnace, as the power dissipated by the element is decreased also. Usually the elements are 2kW with supplied voltage in the range of 60-80V. All elements are connected in parallel to a single low frequency transformer. The temperature inside the furnace is controlled by simple On-OFF process by electromechanical contactors or solid state devices (mostly thyristors). However this leads to uneven distribution of the temperature, as the effects of aging are different and cannot be compensated.

The goal of this paper is the present an independent power supply for each heating element based on power electronic converters controlled by a small 8 bit microcontroller. This allows realizing a power stabilizer that can dissipate the same power in the heating element even with the change of the resistance. The investigations are based on previous research in the same field. Some authors use SCR to control the power to the heater [2, 3]. The same problem is investigated experimentally and by a mathematical model [4], a system with programmable logic controllers is built as well [5, 6].

II. PROPOSED SYSTEM

The proposed system for powering, monitoring and diagnostics of high temperature electric heaters furnaces has the following capabilities:

1. Individual control of each heating element.
2. Regulation and stabilization of the output power to compensate for the increase of the resistance of the element.
3. Regulation of the temperature according to P, PI or PID algorithm.
4. Displaying various information on a display.
5. Remote monitoring and diagnostics via RS232/USB connectivity to a PC.

The block diagram is presented in Fig. 2.
Main blocks are:

A. *Power stage*

Main purpose of this block is to convert the three phase 50/60Hz input voltage into high frequency (50kHz) voltage. A novel IGBT EasyPIM FP10R12W1T4_B3 [7] module is used. The internal structure of the module is presented in Fig. 1. The module can be divided in three parts:

- Input three phase bridge rectifier
- Three phase IGBT inverter with integrated reverse diodes (however four transistors are used as single phase inverter in the current system)
- Temperature sensor

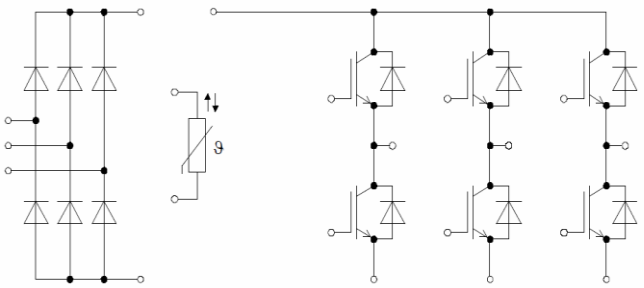


Figure 1. Internal structure of the IGBT module

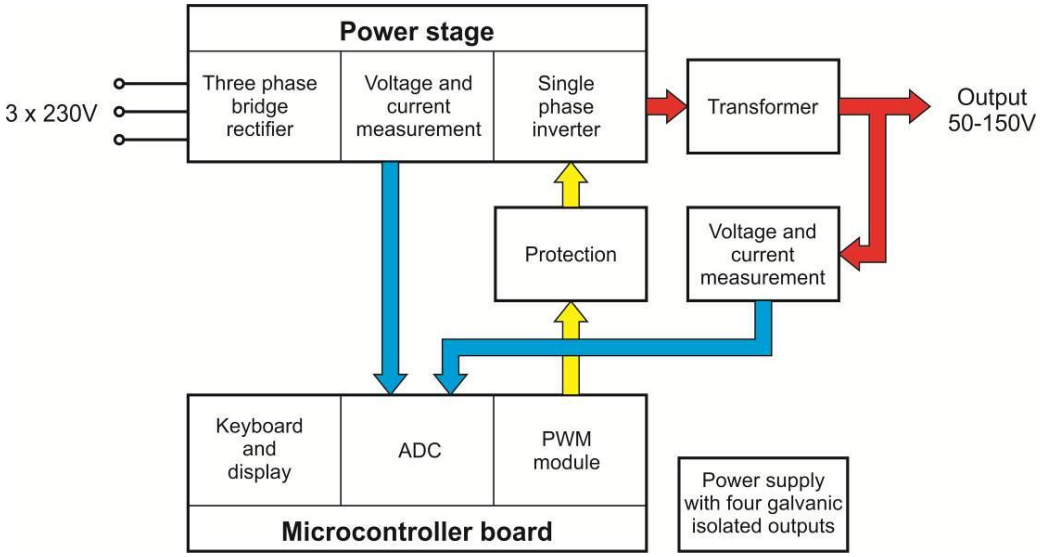


Figure 2. Block diagram of the power supply system

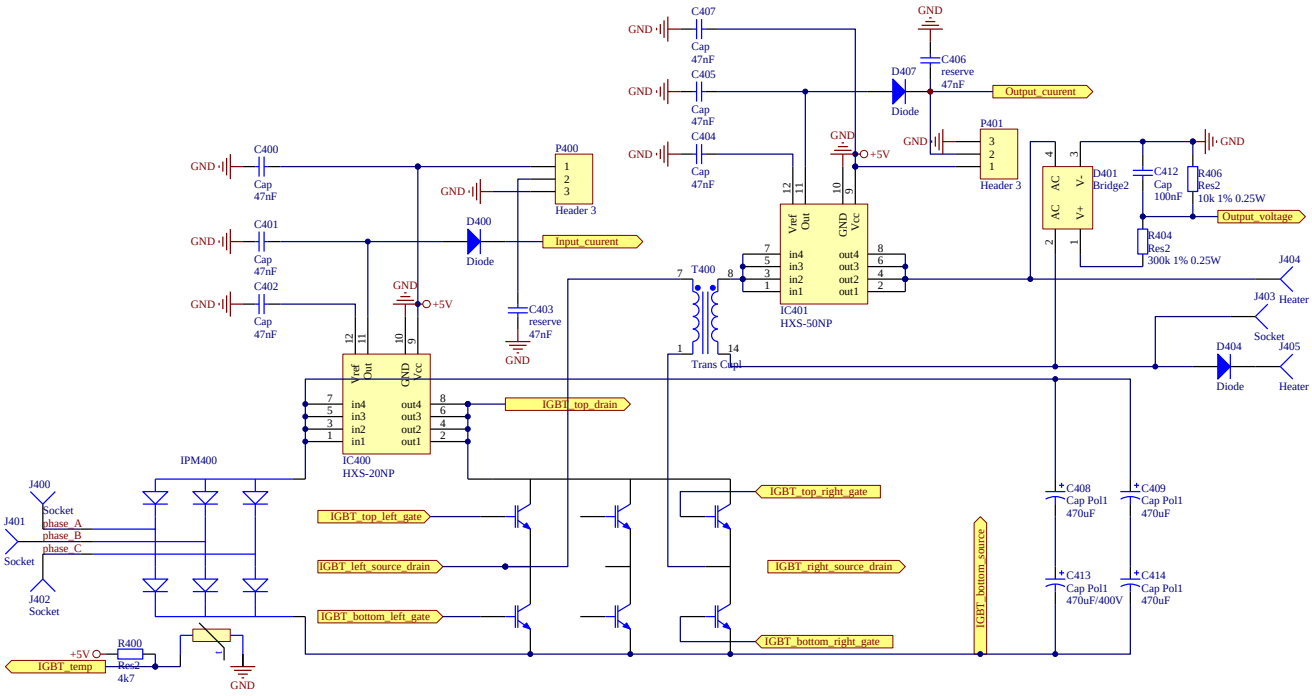


Figure 3. Power stage of the converter

For the current power supply system, only 4 transistors are used, as for the first prototype single phase transformer and inverter are used. Even that two of the six IGBTs are not used in the current configuration, it is still worthwhile using the module. There are three main reasons for this:

- The module save lots of space, by combining 6 rectifying diodes, 6 IGBTs, 6 freewheeling diodes and a temperature sensor. This also simplifies the evacuation of the generated heat with just a single heatsink.
- Once the system with the single phase output inverter is tested, it can be easily upgraded later to a three phase inverter.
- Finally, the price of the module is about 25euro (for more than 25pcs) which is better than 4 transistors for €3.80 (IGW15T120) and 12 diodes each for about €1.10 (DSI 30-12A). You also should consider the additional price for using several smaller heatsinks or the voltage isolation when using single one.

All the three reasons are beneficial to using the module even that it is not currently fully utilized.

B. Protection

Each of the four IGBT transistors is controlled by the PWM module of the microcontroller through an optically isolated driver (HCPL3120). Additionally a desaturation protection circuit is included in order to provide safe operation of the transistors in case of faults.

C. Microcontroller board

A general purpose development board (EasyPIC6 [8]) is used for controlling the power stage. This greatly reduces the required time for creating the required hardware for the microcontroller. If later required, this board can be replaced by smaller one designed specifically for the current application. The microcontroller that is used is an 8 bit PIC16F1937. It is 40 pin IC, member of member of Microchip's extreme low power microcontroller family featuring nanoWatt XLP technology. There are five PWM modules and 14 channel 10bit ADC. As there are 36 I/Os, it allows easy addition of LCD and keyboard for user interface.

D. Voltage and current measurements

Secondary voltage and current value are required in order to calculate the power supplied to the SiC heater. A separate module is designed for the voltage measurements in order to provide galvanic isolation, and still be simple and cheap enough. Measuring high voltages with galvanic isolation can be realized in different ways. Some of the options are presented here:

- Voltage transducer – like the AV100 from LEM [9]. Simple to use and effective solution, however the most expensive one.
- Isolation amplifier – like the AMC1200 from Texas instruments [10]. This is great option when high speed, accuracy and small PCB footprint is required. The price is lower compared to the voltage transducer but still not cheap enough (about 5-7 euro).
- Isolated analog to digital converter –Si8900/1/2 from Silicon Laboratories [11] is a nice option when the voltage is required in digital form. Such ADCs can convert several channels, which allow easy measurements of several parameters. However there are not some widely available, which makes them difficult to find.

In the current system is used alternative option to the third case. Instead of having a main microcontroller that communicates with the isolated ADC via optocouplers, the main microcontroller is communicating with a secondary one, which has an integrated ADC. The digital communication is again by optocoupler. The main advantage of this case is the ability to measure as much analog signals as you like, while having the lowest price of the former three cases. It also allows sending additional data for monitoring other parameters of the system. In the proposed system only one input channel is required, so a small microcontroller is used - PIC12F1822. This is an 8 bit, 8 pin device, with integrated PWM and ADC. The input voltage is divided by a factor of 151, and then supplied to the ADC input of the microcontroller. After this the voltage is converted to pulse width and this information is send by an optocoupler to the main microcontroller. An onboard galvanic isolated DC/DC converter is used to power the voltage measurement board. The total price for this voltage measurement system is less than €5.00, which can be several times lower (compared to the other cases) when having more than two input channels. However additional time is required to write the firmware for the secondary microcontroller.

The current measurement is realized by Hall Effect current transducer – HXS 50-NP [10], capable of continuous measurement of currents up to 50A. Great advantage of this sensor is that it provides galvanic isolated signal, which can be directly connected to the ADC to the microcontroller.

E. Transformer and output rectification

A single phase transformer is used. It steps down the voltage and provides galvanic isolation of the SiC heater. The material grade is K2008 and the size is E65/27. The primary winding is from copper wire, 37 turns with 2 x 1.00mm in parallel. It is winded in two layers, with the secondary winding between them (9turns, 4 x 1.90mm).

The output voltage from the transformer is single phase rectified (for the prototype) by schotky diode. However possible improvement is to use center type transformer and full wave rectification. The PCB is designed in such way, that it is also possible to easily bypass the rectifying diode and supply the SiC heater with AC voltage. However the high resistivity of the SiC (about 900 μΩ.m) and the high operating frequency of the inverter (50kHz) can lead to uneven distribution of the current through the element due to the skin effect. The small diameter of the currently used heating elements is beneficiary to reduce this effect, however the high power heating elements, with diameter of about 50mm, can have reduced performance at high frequency. The skin depth for the used SiC heating element is shown on the next figure.

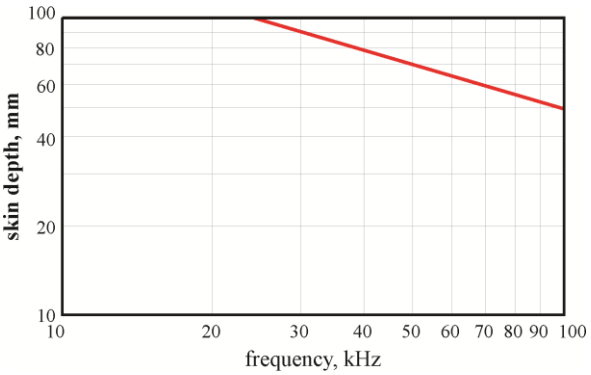


Figure 4. Skin depth for SiC heating element

F. Power supply

It is used to power the microcontroller board, the protection and measurement units. Four independent 15V/0.5A outputs are provided.

On this block diagram is not shown the optional RS232/USB interface for computer connectivity.

III. SYSTEM OPERATION

In the experimental system, the operation is realized in the simplest way. The block diagram for the two microcontrollers is presented on Fig. 6.

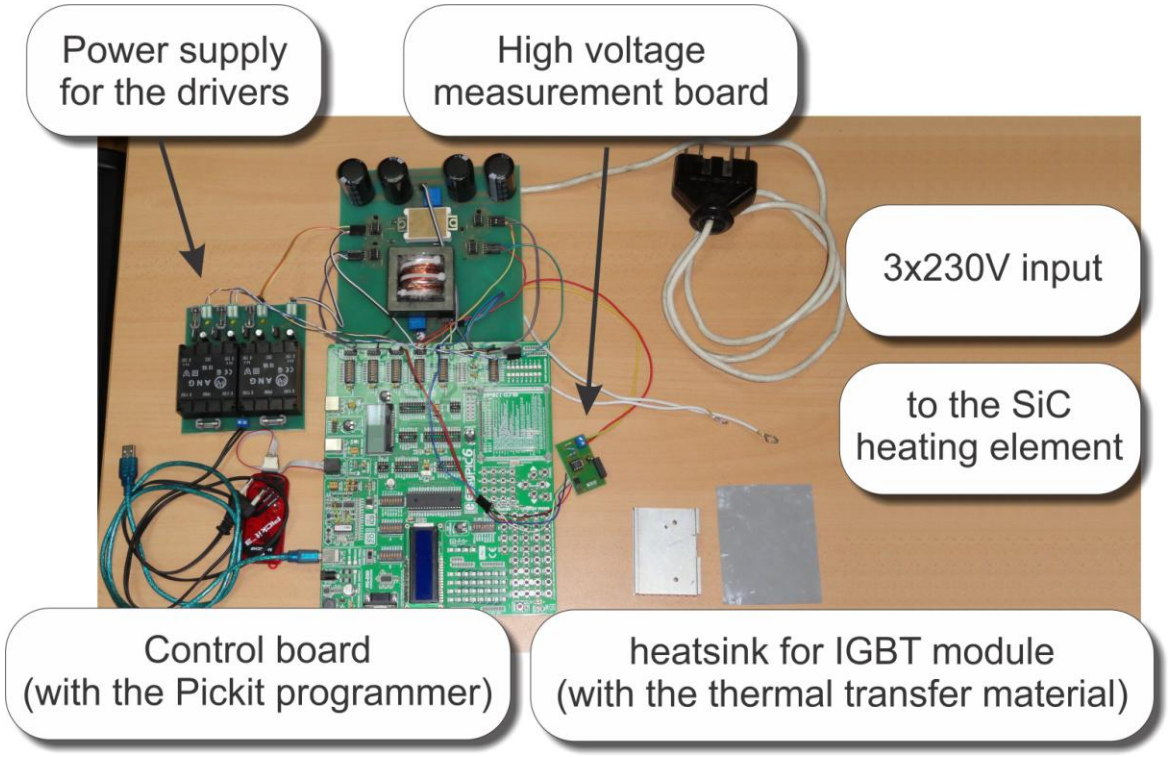


Figure 5. Photo of the experimental power supply system

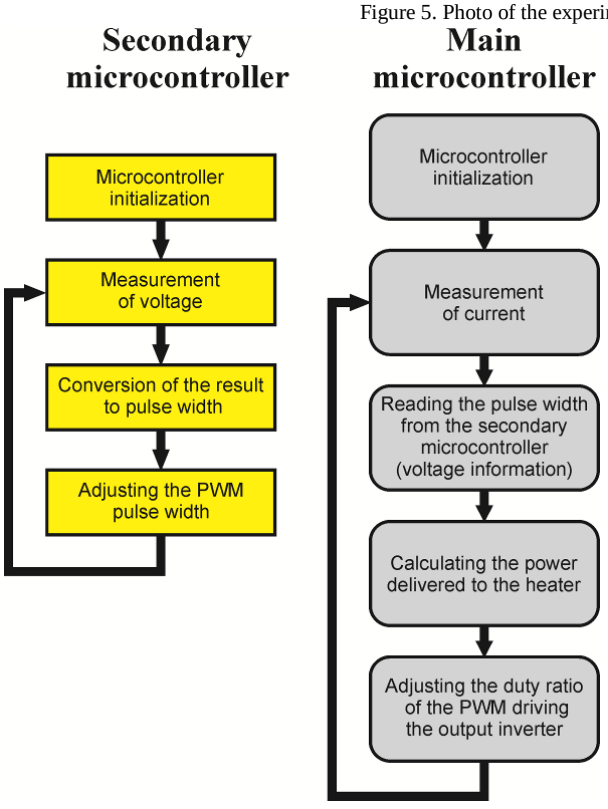


Figure 6. Block diagram for the firmware of the two microcontrollers

The secondary microcontroller is used to measure the voltage only. The integrated ADC is configured to do conversion about every 20µs, resulting in 50kHz sampling rate. The voltage information is coded as duty ratio of the PWM signal generated by the CCP module. The selected frequency is about 5kHz, with 8 effective bits of resolution. Every bit represents 3 volts, resulting in 750V maximum measured voltage. The input voltage divider also limits the voltage to 750V. After the calculation of the voltage and adjusting the duty ratio, new conversion is initiated and the process is repeated.

The main microcontroller measures the voltage from the current transducer and then calculates the current. After this the information for the voltage is read by the main microcontroller by measuring the time between 16 rising edges of the PWM signal. The power is obtained by multiplying the values for the current and voltage. Averaging of 8 consecutive values for the power is done to filter any small errors in the measurements due to different interferences. A simple proportional-integral regulator is realized for the test run. Every 2 seconds the value for the power is compared to the set value (2kW in this case) and the duty ratio of the output inverter is adjusted to compensate the error. For the feature improvement PID controller can be implemented.

IV. EXPERIMENTAL INVESTIGATION

On Fig. 5 is shown the experimental test board without the heating element.

The power supply system is shown with the additional 15V supply. There is a heat sink in the IGBT module which is not shown on the picture. The microcontroller board is not shown as well.

On Fig. 7 are presented the waveforms obtained from the measurements of the transformer. On channel 1 is shown the primary voltage, on channel 2 is the current through the winding, channel 3 represents the losses across the transformer and lastly on channel 4 are shown

the magnetic flux waveform. The measurements are done with no load on the secondary winding.

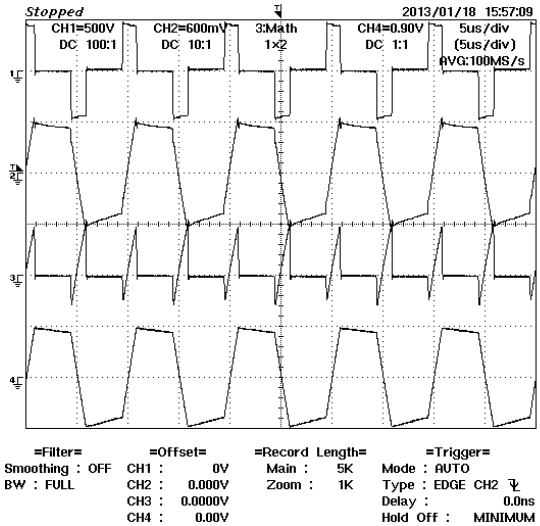


Figure 7. Waveforms at the primary side of the transformer

For the experiments with load we used two different setups. As the available laboratory SiC heater is still new (Fig. 8) it is difficult to observe the automatic compensation of the power supply system. Instead we used the SiC heater just to verify the correct operation of the system. Instead of it we used a water based 5kW load, that can easily change its resistance. This allows taking different measurements of the system. Some of the results are shown in the table below.

TABLE I. OBTAINED RESULTS FROM THE TESTS WITH THE WATER BASED LOAD

Equvalent resistance	Duty ratio	Current	Voltage	Power	Error
Ω	%	A	V	W	%
1.8	22.7	33.1	61.2	2025.72	1.286
3	30.0	25.4	77	1955.8	-2.21
6	42	18.5	111	2053.5	2.675
8	48	15	127.2	1958.88	-2.056
10	50	14	141	1974	-1.3

Figure 8. SiC heater for laboratory testing

Figure 9. Investigated high temperature chamber

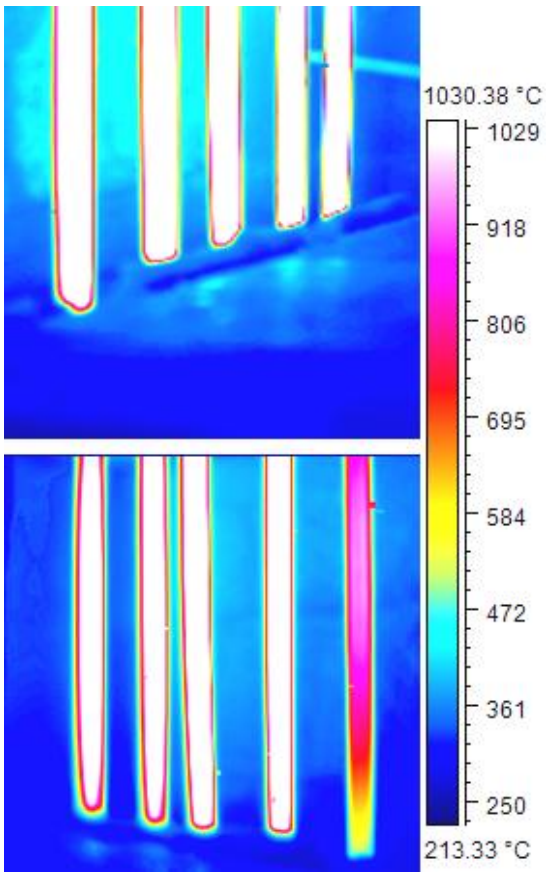


Figure 10. Thermography photograph of the SiC heaters, measured with thermovision camera

The results from TABLE I show, that the system is capable of maintaining the desired power across wide range of the load. As the transformer is winded in order to provide 150V at the secondary winding at full load it is obvious that load can change even further. Usually the SiC heaters can change their resistance in the order of 3 to 1 for both time and temperature variations [4]. However the current proposed system is able to regulate the load in the range of about 6 to 1. In this range the output power is regulated with error within ±3%.

After the initial laboratory tests were carried out. Some field tests were conducted. The investigated high temperature furnace is shown in Fig. 9. There are five SiC

heating elements on each side, resulting in total of ten heaters. Each heater is powered by a separate power supply system. After some time to settle down the heaters, the temperature distribution is observed by an infrared thermo vision camera Trotek [13]. As it can be seen on Fig. 10 All the heaters are evenly heated. The only exception is the right most heating element. Its resistance is beyond the capabilities of the power supply system to maintain the desired power.

V. CONLUSION

This paper presents a microcontroller based power supply system for controlling SiC heating elements. The proposed system is based on several boards that can be replaced/upgraded in the future. The system supplies 2000W to the heater, and is capable of maintaining it when the heater's resistance is changed up to 6 times. By changing only the transformer it is possible to increase that range. It is possible to create application specific control and measurement boards (and replace the general purpose ones here) to minimize the required space, and make a fully integrated solution. It should be mentioned that IGBT module used here requires additional forced cooling, as there is not enough space to mount bigger heatsink. All these remarks will be taken into account for future work. Anyhow, the system proves capable of controlling individual SiC heating elements.

ACKNOWLEDGMENT

This paper is prepared in the frames of Project MU03/163 "Increasing energy efficiency and optimization for electrotechnological processes and devices", Ministry of Education Youth and Science, Bulgarian National Science Fund.

REFERENCES

[1] Silicon Carbide electric heating elements, Kanthal Ltd., catalogue 10-B-1-3, 11.02 3000

[2] A. George, Controlling silicon carbide heating elements with SCR power controls, VP AMETEK HDR Power Systems, 2003

[3] Evalds P., SCR Power Controls Provide Reliable Industrial Electric Heating, Industrial Heating, October 2002,

[4] Howes G., Control of silicon carbide heating elements wsing SCRs, Eurotherm corporation. version 2, 1988

[5] Dimitrov B. A. Marinov, V. Valchev "Experimental based selection of a structure of a power electronic converter for control of silicon carbide electric heater" XVII-th Internacional Symposium on Electrical Apparatus and Technologies SIELA 2012, Burgas 2012, Volume 1. pages 82-89

[6] Dimitrov B., I. Yachev, "Study of high-temperature chamber-type electric resistance furnace and attainment of uniform temperature field" 11-th International Conference on Applied and Theoretical Electricity – University of Craiova ICATE

[7] Dimitrov B., G. Nikolov, P. Andreev, "PLC based control system for high temperature chamber resistive furnaces with silicon carbide heaters" Юбилеен конгрес с международно участие "50 години ТУ-Варна"

[8] FP10R12W1T4_B3 datasheet, http://www.infineon.com/

[9] EasyPIC6 datasheet, http://www.mikroe.com/easypic/v6/

[10] Current transducer HXS 50-NP datasheet, http://www.lem.com/

[11] Fully-Differential Isolation Amplifier AMC1200, http://www.ti.com

[12] Isolated Analog-to-Digital Converters Provide Flexible, Cost-Effective Solutions for AC Line Monitoring, Silicon Laboratories, Inc. Rev 1.0, 2012

[13] http://www.trotec.de

Končar Smart Metering – Innovative Solutions for Well Known Problems

Drago Cmuk, Davor Cihlar, Marijo Šimunović
Končar-Inem/PJ ICT sustavi, Zagreb, Hrvatska
{dcmuk, dchilar, msimunovic}@koncar-inem.hr

Abstract - EU climate and energy package and Energy Efficiency Plan 2011 has set very ambitious goals for all the Member States regarding Smart Metering (SM). In Croatia, different infrastructural problems and lack of independent regulating body allowed only partial solutions by each utility. On the technical level rollouts were slugged by lack of fully multifunctional solutions with minimal installation and maintenance costs, able to adapt to different meter producers with multiple access to the same data. Končar made strategic efforts to overcome these obstacles trough an integrated R&D approach. Realized solution Končar AMI described in this paper relies on the optimized network architecture with innovative long range and short range wireless communication solutions and modular communication equipment. Such a novel approach allows great flexibility, excellent network coverage, good signal propagation and easy installation process with reliable data transfer with minimum operating expenditures. Combined with a cloud based MDM solution they make the milestones for a powerful and highly adaptable, integrated Smart metering solution that could bring numerous advantages to final users and short ROI. Končar AMI was tested trough a number of pilot projects and proved a concept particularly appropriate for bigger SM rollouts.

I. INTRODUCTION

Multidisciplinary field that deals with the collection, management and use of real data on energy consumption and possible savings is called Smart metering (SM). The existence of a large number of different approaches to this problem so far has actually been an aggravating factor in the creation of a universal solution that can be applicable in most situations. Requirements for low capital (CAPEX) and operational (OPEX) costs of such systems only further exacerbated the situation [1]. This paper describes the solution developed in Končar INEM, which has the potential to overcome several important technical limitations that impeded the massive installation of such systems in recent years.

Benefits arising from the use of SM systems are multiple, from increased energy efficiency and energy savings to decrease in environmental pollution, which makes them very attractive to public and private investors. Positive effects of the use of SM are the result of energy saving from the users itself, on the basis of timely and frequent information that the system provides [1].

SM system is actually at its core a complex information and communication system that allows for the collection, processing and storage of data on energy consumption from end users. The basic components of such a system are shown in Fig. 1. However, the SM can be so much more.

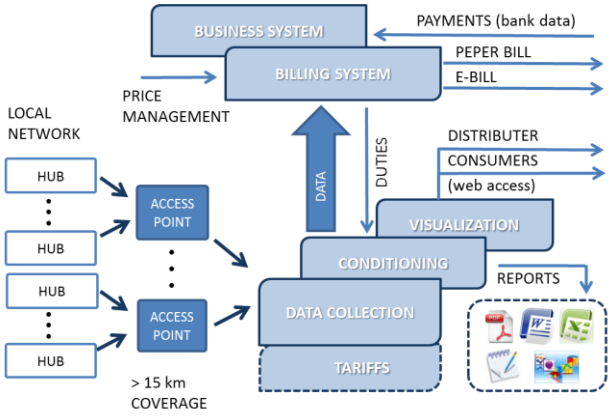


Fig. 1. Smart metering system

Connecting all information services of a company in a single central system, AMR solution becomes advanced Advanced Metering Infrastructure (AMI), which beside being bidirectional, includes the implementation of advanced billing system, customer web portal, asset management and many other features. At the end of the chain, SM includes the installation of smart meters in the user object and allows data collection, data processing and informs the users about the current consumption of electricity and other energy sources. Just based on the increased amount of timely information user can plan and optimize their consumption. In special circumstances, the advanced metering system can control the delivery of energy, and sometimes even enable regulation of the consumption and production of energy.

Končar AMI is based on two key components: a two-way radio frequency (RF) communications network (long-range and low-consumption), which enables low installation and maintenance costs of the system and a centralized, network-based software that ensures maximum performance and maximum service availability. RF component of the system is the solution suitable for remote, rural and industrial environments with low population density, where the feasibility of the realization of wired systems is questionable, but also in urban, densely populated areas where a single device can

communicate with a large number of other devices. The application is available via the Web and is best suited for implementation in cases where the applicant access to multiple users on different domains. The modular architecture of hardware and software components of the system guarantees the flexibility and adaptability at the highest level.

II. HIGHER PERSPECTIVE: ICT IN ENERGY SECTOR

In recent years, ICT technologies become firmly positioned in the energy market. Conservative and robust energy market came to a point where its long-term stability is threatened - if short-term investments do not increase to improve performance and achieve savings. Climate change and the reduction of energy reserves make this picture even more serious. At the same time, the chance of ICT to enter this market is growing on a daily basis, especially taking into account a number of EU projects in the field of advanced metering.

A. Business framework

European Council in December 2008 adopted The EU climate and energy package which requires the EU Member States to achieve the following by 2020:

- reducing greenhouse gas emissions by 20%
- introducing a 20% renewable energy, and reducing the consumption of primary energy in the EU by 20% [1].

These requirements oblige member states to radical changes and to start launching projects that will enable the achievement of the goals set in this short time. The European Commission believes that the ICT sector can help reduce the impact of greenhouse gases through three framework plan:

- provide energy savings in the global economy, especially in most profitable sectors,
- turn the intangible economy where teleworking, e-sales and cloud computing provide significant energy savings,
- introduce a measurement of energy consumption in homes, power grids and industry to increase awareness of the importance of reducing energy consumption.

At first glance, it does not sound surprising that the European Commission has such high expectations of ICT because ICT improved way of life and business in wast range of aspects. Such a climate is the long-awaited release lever for projects in the field of SM, which have the unique capability of providing accurate data on energy consumption to all involved parties, from individual users to the government and the European Commission. However, the SM technologies have not yet reached the required level of maturity necessary to enable making of informed and accurate decisions at the highest level.

On the technical level, the growth in SM was slugged by a variety of factors. Most important of them in Croatia include:

- lack of universal solutions with minimal installation costs and maintenance (different meters and incompatible systems)
- a wide range of technologies, communication channels and protocols used in existing solutions
- different utility companies and government bodies require access to the same data [2, 3].

Although a number of EU projects became catalysts of the standardization processes and facilitate the definition of appropriate technologies in the field of SM, much work is still undone. Although a number of standards such as DLMS/COSEM etc. provide a common foundation for interoperability in the future, currently implemented systems include a range of proprietary solutions that prevent real competition among different manufacturers, revealing a number of stairs to pass towards truly interoperable solutions.

B. Technical framework

SM systems are usually organized in communication networks in which great number of nodes collect measurement data from sensors and send them to central node. Since these are often wireless networks, it is necessary to consider the proper communication protocols, frequency channels and modulations which will be used in system development. From a technical point of view, the choice is made as a best fitting answer to number of questions. The key elements of this selection process are:

- signal propagation distance
- environment configuration (physical obstacles, reflections, etc.)
- antenna properties (dimensions, sensitivity)
- output power
- interference and channel selectivity
- generated data size and network throughput [4].

Systems in the domain of SM generally use frequency bands around 868 MHz. However, the use of the 169 MHz radio spectrum band for ERMES systems has dramatically decreased with the result of the spectrum not being efficiently used. Sequentially, this frequency band is addressed for use in six types of preferred applications, including meter readings [4] [5].

The devices that use above-mentioned frequency band must fulfill further demands:

- radiofrequency transcievers have high selectivity, sensitivity and good blocking properties
- system core is low-power MCU
- network topology is suitable for efficient powering system implementation (if network nodes are not battery powered)
- battery powered devices have long autonomy (more than five years).

WMBus, ZigBee, 6LoWPAN, 802.15.4. are standardized solutions already used in the domain of advanced measurements, with WMBus being only

European Norm for wireless metering. WMBus standard can be used for e-meters as well [6].

Receiver sensitivity, interference filtering from other devices in the same frequency band (selectivity) and devices in other, neighboring and far bands (blocking) define device quality performance.

Regarding the immense number of protocols and devices that work in the same frequency band (meters, alarms, wireless audio, LTE base stations, etc.), the interference problem rises as a most restrictive element in system design. Some solutions become much more widely adopted in USA, Australia and similar countries than in EU due to the higher power and bigger number of channels available (eg. DigiMesh dynamic routing depends heavily on several channels available in 900 MHz US spectrum, while in EU only one in 868 MHz is available).

Besides the device quality performance in radio-frequency devices, the system development is mostly leveraged by energy consumption problem. In state-of-the-art sensor networks, the RF communication makes the biggest part of device power consumption. In order to increase longer lifetime, sensor nodes must keep their communication components off most of the time. To achieve this, a great number of synchronization mechanisms between transmitters and receivers is developed. ZigBee standard based its functionality on a presumption that devices has constant power source. Accordingly, ZigBee standard doesn't provide any power saving mechanisms for routing nodes, which decreases battery powered devices lifetime significantly. Consequently, this narrowed down the possible application range for ZigBee devices and caused emersion of numerous alternative solutions [7].

C. Existing solutions

Energy efficiency is a fundamental subject that pervades the design of communication protocols for wireless sensor networks (WSN) including routing protocols and MAC layer protocols (when spectrum limitations are kept firm). One of the primary mechanisms used for achieving low power consumptions in energy restrained networks is the use of power cycles, which consist of activity periods and sleeping periods for power saving. Because of the sleeping periods, the greatest challenge is reducing power delays for big data throughput with maintained low power consumption. Existing approaches to this problem can be roughly divided in two categories: techniques that use some synchronization method (S-MAC, T-MAC, DigiMesh) in order to achieve overlapping awake periods of network devices and techniques that don't use synchronization mechanisms but are based on communication with elongated preamble and low power listening – LPL (B-MAC, WiseMAC) or frequent transmission of short data packages (Wireless M-Bus) [8].

These approaches are motivated by the requirement to decrease awake period durations in which node listens medium even though there is no data being send to that node. In these intervals, significant amounts of energy are consumed so the use of these techniques should be avoided in WSNs. DigiMesh [9] successfully solved a number of the "sleeping router" issues by using synchronization algorithms, however some important drawbacks remained. Besides being a proprietary protocol, important drawback are cyclic periods in which whole network is awake (and dynamic routes are generated by using broadcast messages). This way awake time of average node rises significantly - when dealing with disturbances of neighbor networks, and no additional channels (like in EU 868 band) are available. Advantages in one sence (eg. sleeping routers) always brings about drawbacks in other fields (effective data rates, minimum response time, etc.). All the mentioned solutions have numerous advantages in different utilization scenarios, but still an optimized solution for typical 15 min read intervals of energy consumption and long battery life in bigger rollouts (with disturbances from neighboring networks, and only one channel available) is required. Although direct performance comparison of WSNs is out of the scope of this paper, Končar AMI brings several improvements in this area also, and they, although described briefly, clearly show advantages of the system as a whole and the integrated approach which is presented.

III. KONČAR AMI

A. Končar AMI system design

Končar AMI system is designed to meet the following requirements:

- low installation and implementation expenses (CAPEX)
- minimal operating expenses (OPEX)
- interoperability with broad range of different measurement and communication systems
- multifunctional, modular system applicable for water, gas, heat and electricity consumption measurement
- multifunctional, modular, network based software for data management, advanced user segmentation and advanced data sharing capabilities, appropriate for *off the shelf* implementation in different utilities

Končar solution, showed in Fig. 2, consists of optimized network architecture based on a short-rage wireless network at 868 MHz (HAN, IEEE 802.15.4) [9], long range wireless network at 2.4 GHz (WAN) with advanced modulation technique of ultra-link processing (ULP) and an unique device that acts like a bridge between them – concentrator.

This approach made possible great signal coverage of wide urban and suburban, as well as remote, rural and industrial areas, eased the process of installation and provided means of reliable data transfer without additional monthly costs and paying any licence fees. These system components, combined with Končar MARS software, are foundations for creating highly adaptive and universal solution for smart metering that

could make it preferable solution in comparison with existing systems [3].

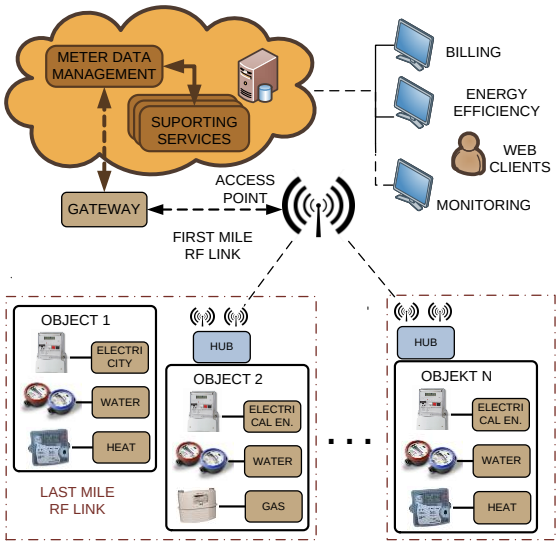


Fig. 2. Končar AMI

B. *Končar AMI system communication model*

Node is a physical device in Končar AMI network. Assigned to every node is a 24-bit number MID, which is unique to each device in network. Node possesses two distinct functions: communication and measurement. Communication is related to sending and routing messages through the network to the central node. Measurement function on end devices involves different interfaces for accessing measurement devices (RS485/ IEC62051-21, RS232, pulse inputs, WMBus). Measuring node can use multiple interfaces, and each interface holds multiple channels for identification of devices being read. Each participant in the network implements communication architecture showed in Fig. 3.

Architecture can be roughly described through three layers: application layer, network layer and communication (COM) layer.

Application layer generates and consumes messages. MID number is used for message destination addressing. Application layer assigns destination MID to the message and additional data generated or used in each of the application modules. Those modules are *MeterApp* which holds application logic for measurement channel, *ServiceApp* which takes care of network maintenance and service and *DiagnosticsApp*, which serves for continued network monitoring by sending periodical diagnostic messages and asynchronous alarms in case of errors.

Network layer – Konect, reroutes incoming messages to the application layer and generated messages to the appropriate communication interface. This layer implements message redirecting over the network topology, and selecting appropriate network layer (ORW, K-MAC, GPRS, RS232, ETH).

Network layer retransmits the message to its final destination in the network. There are two types of redirecting. The first case is when data destination is network root node. This redirecting is trivial since with each hop message travels upwards to the parent, and finally to the root of the network tree. Second case comprises multiple hops both in direction of node's parent and in direction of node's children until it gets to the final destination in the path. For this purpose, dynamical routing tables that are part of the network layer are being used. Tables contain entry for each descendant node which connects offspring's node MID with means of access for that particular node. Node access description contains COM interface (ETH/ORW/RS232/GPRS/K-MAC) and address which depends on the particular interface (IP/nodeID/HAN). This information enables determining next hop in path to the message destination. Routing table is refreshed each time when descendant node message is passing to the network root node. Inactive members (e.g. the ones reconnected to the other network node) of the network are being removed according to the predefined network rules.

COM layer is modular and its implementation depends on device hardware. Each device possesses one or more interfaces that are being used through COM layer. This layer is used for data transmission between two neighbouring devices in the network. Therefore, this layer has no knowledge of entire network topology, but only about „local network tree“ which includes connection to parent node and connections to children nodes.

COM layer implements reliable communication with transfer acknowledge and retransmissions (in case of errors). For a single connection to the neighbour node network link quality is calculated and stored. If necessary, connections with poor quality are closed and new ones are created. Also, COM layer has a synchronization task – child node adopts parent's current time. Time uncertainty increases with network depth with acceptable offsets (in practice, network topologies are only few levels deep).

Incoming messages propagate from the lower layer of the diagram (Fig. 3) to the upper layers. Outgoing messages go in opposite direction. In particular, when the data is not addressed to the current device, it is directed up towards the Konect layer. After determining the appropriate path, using available information on network topology (routing tables), message is transmitted to its actual destination using the appropriate COM module.

a) *Local wireless network*

K-MAC is a MAC protocol for wireless sensor networks, and it is developed as a part of Končar Advanced Metering Infrastructure. K-MAC is an adaptation of MAC communication layer described in 802.15.4-2006 standard. Unlike the asynchronous protocols that use LPL techniques [10], K-MAC implements communication protocol with synchronization between all network nodes

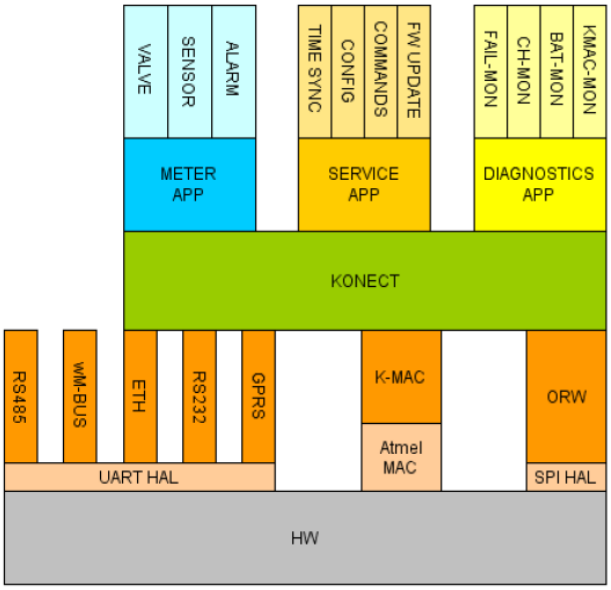


Fig. 3. Končar AMI system communication architecture

Network nodes are battery powered modules and they can be divided in different groups: registrators – nodes that collect measured data and routers – if they retransmit messages to the central network node – concentrator. In each device, duty cycle mechanism is used to extend battery lifetime:

- nodes are in sleep mode for most of the time
- nodes are awake only couple of times a day for network synchronization and diagnostic messages sending
- nodes awake synchronously to send measurement by predetermined schedule.

Since only one channel is being used at 868 MHz separation of two networks in same RF signal range can be done through use of different network identification numbers (PAN IDs). For address assignment in a single network C-skip algorithm is used [7].

Duty cycling decreases device awake time significantly. Scheduler conducts timely task completion. Advanced system for assigning tasks enables low power consumption in overall network communication. Additionally, network parameters are optimized to network data traffic which increases system efficiency significantly.

Synchronization is conducted from network root to children nodes on deeper network levels. During the scanning process, node gets information from all the nodes in RF range (duty cycle parameters, link quality) and based on that information, chooses best parent candidate. After synchronization, each device is aware when its parent will be awake again. Equally, since the synchronization is performed with acknowledgment, parent node is also aware when the child node is awake. When there is data that needs to be sent, a transmitting node will calculate time to the next awake period of a destination node. Until that moment, node can sleep, saving the battery power and awake just in time to send the message successfully.

Additionally, as a result, unsuccessful transmission number is decreased so the network configuration is more stable which means fewer number of repeated scans and node re-connections. K-MAC solves problem of increased power consumption caused by listening in inactive state that can be found in asynchronous protocols like B-MAC and WiseMAC. These protocols are based on preamble sampling where transmitting node must send the preamble as long as the receiving node is sleeping [8]. K-MAC also solves the problem of node saturation during increased data traffic intervals found in X-MAC protocol. This is caused due to constraint of sending at most one data package per awake cycle [10].

b) Long range wireless network

Long range coverage is implemented through multiple communication paths – GPRS, Ethernet or OnRamp Wireles ULP. Path selection comes down to use of the appropriate communication module in modular concentrator. In addition to standard long range communication solutions, OnRamp Wireless ULP module conducts the communication of distributed network nodes and central system in which transmitted data is being collected, stored and processed. Parent side uses *Acces Point* (AP) connected to *Gateway computer* (GW) over Ethernet. Children side uses eNode or dNode module.

OnRamp Ultra-Link Processing system has star topology with half-duplex communication. Access point communicates with nodes directly, in one hop. ULP enabled signal reception even under thermal noise level (sensitivity up to -142dBm) which makes possible distribution of the devices which communicate in unlicensed frequency range over long ranges, independently of possible interference or challenging end locations. Network of multiple access points simply creates redundant signal coverage over large area (preliminary test shows over 15 kilometers range in licence-free spectrum in Zagreb).

IV. KONČAR MARS

Končar MARS (Managament and Automated Reading Software) is central part of the Končar AMI systems and is a programming backbone of the system. MARS is a software for centralized measurement data collection, manipulation and visualisation. Additionaly, software can be used for report generation and connection to the other systems (e.g. billing systems).

Končar MARS consists of client and server application. Server architecture is a proprietary solution and enables data collection from broad range of devices using different communication technologies. System implements numerous communication protocols like Ethernet, OnRamp ULP, GPRS, RS232 and RS485. Each protocol is implemented as an independent service with unique MARS interface.

Modular architecture enables fast upgrades and expanding to other technologies. MARS server acts as a bridge between different data sources and user

application for processing and visualisation of measured data. This process is hidden from end users. User only accesses an interactive client application available through any Internet browser. This is one of the most important MARS features – data availability from any computer that has Internet access. MARS features and capabilities are grouped over different modules and are divided in three main group according to different user groups.

First group comprises physical assets of the system: meters, counters, sensors, substations, etc. and link quality to that devices. Fundamental modules of this group are Data Collection Modul and Data Integrity Module. Data Collection Module integrates real, physical data sources (meters, sensors) into system. Data Integrity Module informs users if problem arises during data measuring or device accessing.

Second group is related to logical grouping tasks and data visualisation. This group includes Logic Management Module which allows user to structure system elements according to the preferred logic hierarchy and criteria (e.g. device location, equipment owner, etc.).

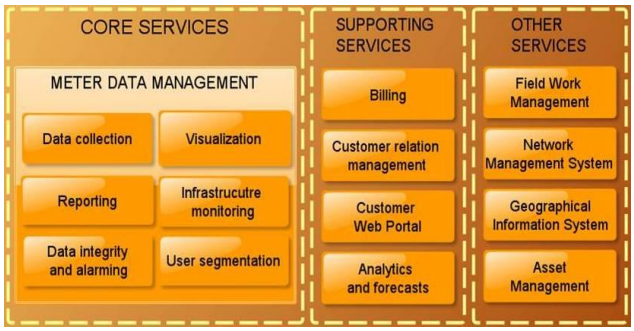


Fig. 4. Končar MARS application modules

Other modules are Data Visualisation Module, Data Report Module and LiveView Module which enables real-time display in cases of increased data traffic.

Third group are administrative modules: User Management Module, Template Production Module, and System Event Monitoring Module. Through use of modern tools and technologies, Končar MARS introduces device location display, with geographical maps, which gives fast and intuitive environment for system monitoring. These tools represent the framework for AMWR system maintenance and ease this process significantly.

V. CONCLUSION

This paper presents a large market potential of SM systems that, when confronted with specific implementation constraints, became significally decelerated. More precisely, solutions not optimized for given situation often cannot provide a sufficient ROI to enable the mass installation. Solutions to some of mentioned constraints are the result of several years of work of the AMI team in Končar - Electronics and Informatics Inc. on solving them. Proposed long-range wireless network with unusually high coverage, bidirectional local network with advanced sleeping repeater system and modular hardware solutions are most important of them. Additional advantages rise from the software package for the collection and processing of measurement data - MARS. Implemented as a Web app, MARS puts the full functionality available to distributed users without the need for installation, it provides advanced network diagnostic (which greatly facilitates maintenance and lowers its cost) and allows easy connection with a billing system. All these solutions provide ease of installation, low maintenance costs, and a remarkable system performance which are prerequisites for ensuring the critical ROI for mass installations. System performance was shown on several important pilots (Zagreb Holding , Končar - Infrastructure and Services, City of Zagreb, etc.) that clearly proved (beside technical maturity) a remarkable market potential of Končar AMI. Further development of new functionalities and the supported protocols is underway unleashing new capabilities and further expanding the areas of its applicability.

REFERENCES

[1] EU Commission, »Official Journal of the European Union,« 11 2012. [Mrežno]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:315:0001:0056:EN:PDF. [Pokušaj pristupa 3 2013].

[2] D. Cmuk and M. Puškarić, *Brak iz interesa,* Zagreb: InfoTrend, 2012.

[3] D. Cmuk, M. Puškarić and Ž. Delač, *Što to mudri rade sa pametnim mrežama?,* Zagreb: InfoTrend, 2012.

[4] CEPT, »Paging frequencies get new applications,« Europe's Information Society Thematic Portal, 2012.

[5] *EN13757-4: Communication systems for meters and remote, Part 4: Wireless meter readout,* Wien: Austrian Standards Institute, 2011.

[6] M. Stefanov, "Embedded World," Lipanj 2011. [Online]. Available: http://www.embedded-world.eu/fileadmin/user_upload/pdf/smart_home_metering_summit/Session_2/06_Milen_Stefanov_-_Hochperformante_WMBUS_subsystem_26Oct11_Release.pdf. [Accessed Veljača 2013].

[7] D. Gislason, ZigBee Wireless Networking, Washington, USA: Elsevier, 2008.

[8] M. Buettner, G. Yee, E. Andersen i R. Han, »X-MAC: A Short Preamble MAC Protocol for Duty-Cycled Wireless Sensor Networks,« u *Proceedings of the 4th international conference on Embedded networked sensor systems*, Colorado, USA, 2006.

[9] *IEEE Std 802.15.4™-2006 Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks,* IEEE Standard for Information Technology, 2006.

[10] O.Yang and W. Heinzelman, "Modeling and Performance Analysis for Duty-cycled MAC Protocols with Applications to S-MAC and X-MAC," *IEEE Transactions on Mobile Computing*, vol. 11, no. 6, pp. 905-921, 2012.

Micropositioning mechatronics system based on FPGA architecture

Ervin Kamenar and Saša Zelenika
University of Rijeka, Faculty of Engineering and Centre for Micro and Nano Sciences and Technologies, Rijeka, Croatia
ekamenar@riteh.hr, sasa.zelenika@riteh.hr

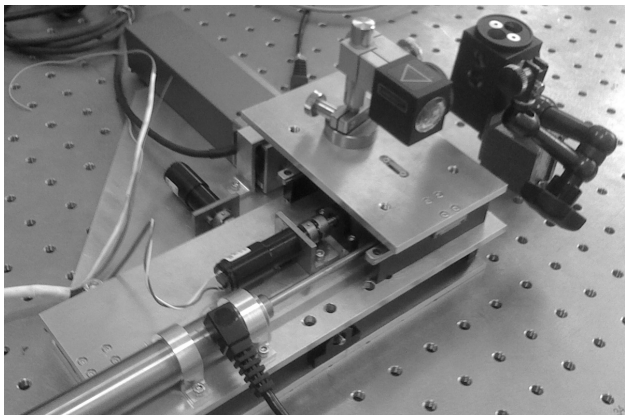
Abstract – In this paper a micropositioning mechatronics system is conceived and described. Its mechanical design is optimized in order to achieve high precision displacements. High speed closed-loop feedback control is obtained by using both PID and ramp control algorithms programmed as virtual instruments (VI) on an FPGA (Field Programmable Gate Array) module. User controls are programmed in an independent Host VI. A Linear Variable Differential Transformer (LVDT) is employed as a position feedback sensor, while positioning accuracy and repeatability are experimentally assessed by using a Michelson-type laser Doppler interferometric system. Different experiments with point-to-point positioning are conducted. In the first set of experiments, positioning with 100 µm steps is performed. In a second step, 10 mm positioning experiments are done. The system shows a marked nonlinearity when longer positioning steps are used; this nonlinearity influences significantly the output error. It is hence found that the main contribution to system's nonlinearity is caused by the LVDT. The error is compensated via system linearization by an experimentally obtained analytical function which is programmed in the LabVIEW Host VI. Positioning accuracy and repeatability are finally assessed experimentally again and true micrometric positioning is achieved.

I. INTRODUCTION

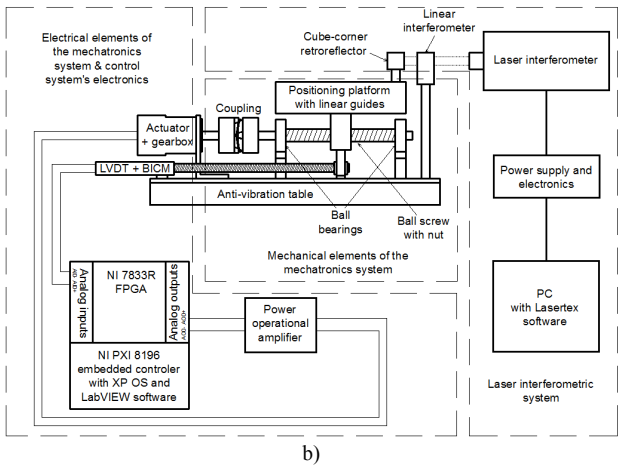
High-precision positioning over stroke lengths from several tens of millimetres up to several metres is often needed in manufacturing systems where this aim is frequently achieved by using positioning mechanisms driven by rotary motors [1]. In this frame, an accurate tracking performance coupled with specified requirements for positioning accuracy and repeatability is essential in realizing ultra-precision devices. Moreover, nowadays there is a clear tendency towards reducing the dimensions of the products, i.e. towards miniaturization [2]. In dealing with these challenges, improving the mechanical elements that make up the considered system is generally rather expensive [1-3]. On the other hand, reliable control systems offer a good alternative [2-5].

Field Programmable Gate Arrays (FPGA [6]) represent a possible technological approach in the implementation of suitable control typologies. FPGA technology has been recently widely used in many commercial, as well as in industrial applications. FPGAs are characterized by flexibility, high performance and low-level reconfigurability. Also, a very important differentiating feature with respect to conventional Central Processor Units (CPUs) is that a single CPU can process only a single instruction per time cycle. On the other hand, an FPGA can be configured as several virtual processors capable of functioning in parallel. Moreover, FPGAs allow digital solutions for the controllers used in closed-loop systems. In fact, an important prerequisite for good control is a feedback-based closed-loop control system. A good feedback implies, in turn, the usage of sensors with high resolution and minimal nonlinearity [7-9].

In this work, a micropositioning mechatronics system is conceived. Digital high speed closed-loop feedback control is implemented. Two types of control algorithms are used: a digital proportional-integral-derivative (PID) controller and a ramp control algorithm. The PID controller is custom designed since the available LabVIEW software does not include the LabVIEW PID Control Toolkit license. On the other hand, the ramp



a)



b)

Figure 1. Experimental set-up (a) and its scheme (b)

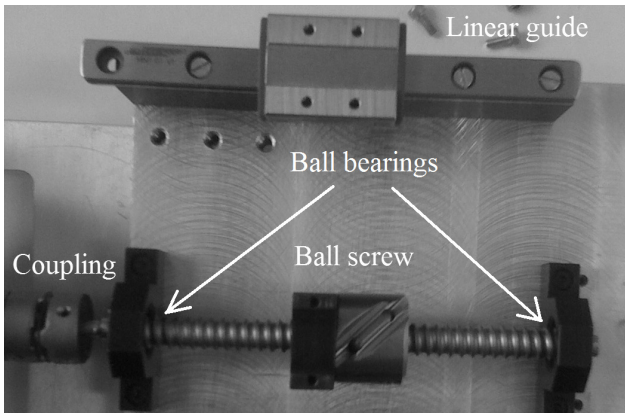


Figure 2. Main mechanical parts of the system

controller is based on a linearly decreasing actuator input voltage proportional to the distance of the moving part to the defined reference position.

II. EXPERIMENTAL SET-UP

The development and assessment of the performances of the conceived micropositioning mechatronics system is performed on a suitable experimental set-up depicted in Figure 1. The single axis high precision system is constituted by a DC actuator, a feedback sensor, mechanical elements, the control system and the laser interferometric system used for the nanometric assessment of the positioning accuracy and repeatability.

A. Actuator and Feedback Sensor

The mechatronics system is driven by a Faulhaber M 1724 006 SR DC actuator with a nominal voltage U_N = 6 V, a no load speed n_0 = 8600 rpm and dimensions ϕ = 17 x L = 24 mm. The motor is integrated with a two-stage L_g = 17.7 mm Faulhaber series 15A plastic planetary gearhead having a reduction ratio i = 19:1 [10].

The used feedback sensor is an Omega series LD610-50 Linear Variable Differential Transformer (LVDT) with a measurement range of l = ± 50 mm (100 mm). An LVDT is a linear displacement/position sensor which consists of a static and a moving element. The static element bears a central primary winding excited with an AC excitation voltage V_e = 2 V RMS, which is located between two symmetrical secondary windings. The moving element consists of a separated cylindrical core made of a Ni-Fe alloy, mechanically connected to the moving part of the system. When the system moves, a sinusoidal voltage with an amplitude proportional to the movement is generated on the secondary windings [11]. AC voltage is conditioned by the Boxed Inline Conditioning Module (BICM [12]) which is embedded with the LVDT and its output is set to cover the full scale of the induced voltage $V_{BICM\ max}$ = ± 10 Vdc. The BICM module also generates the AC excitation voltage of the primary winding.

B. Mechanical Elements of the Mechatronics System

Although, in general, various types of sliding and rolling leadscrews are used in high precision applications, ball screws are the most common solution in practical applications [13]. In the developed experimental set-up,

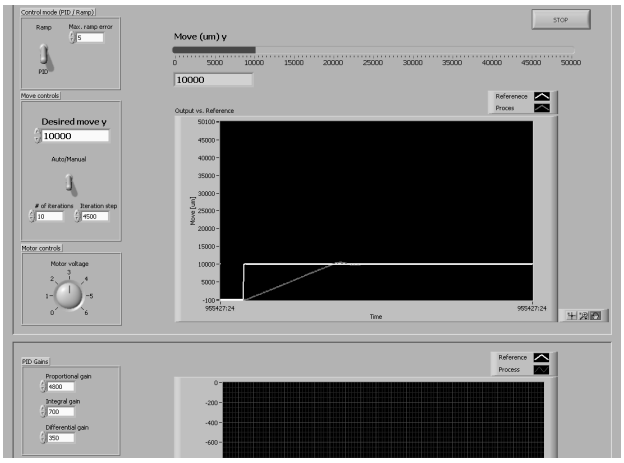


Figure 3. Host VI front panel

the rotation of the DC motor is therefore converted to linear displacement by using an SKF type SH miniature ball screw with integrated tube recirculation having a nominal diameter d_0 = 6 mm, a p = 2 mm lead, a nominal maximum backlash B = 50 μm and a practical efficiency of η_p = 94% [14]. The screw is supported by employing Miniature SKF 618/4 d = 4 mm, D = 9 mm, b = 2.5 mm ball bearings [15]. The motor is connected to the ball screw by using a miniature aluminium Misumi MCGS13-3-3 coupling with a compliant polyimide disc enabling the compensation of lateral and angular misalignments and having a nominal outer diameter D_C = 16 mm, inner diameters $d_1 = d_2$ = 3 mm and the length l_C = 13 mm [16].

Schneeberger MINIRAIL profiled miniature guideways MN 7 with fixed part dimensions: $l_f/w_f/h_f$ = 85 /7/4.5 mm and moving part dimensions: $l_m/w_m/h_m$ = 24.6/ 17/6.5 mm are used as linear guides allowing the sliding of the movable part of the system [17].

The described ball screw, ball bearings and linear guideways are shown in Figure 2.

C. Control System

The control system of the conceived mechatronics device consists of a National Instruments (NI) PXI-1050 chassis, including a PXI-8196 embedded controller and a reconfigurable PXI-7833R FPGA module.

The NI PXI-1050 chassis has eight PXI (PCI eXtensions for Instrumentation) slots with four integrated Signal Conditioning Extension for Instrumentation (SCXI) subsystem slots. The SCXI system is connected via a bus to the PXI slots enabling the control of the

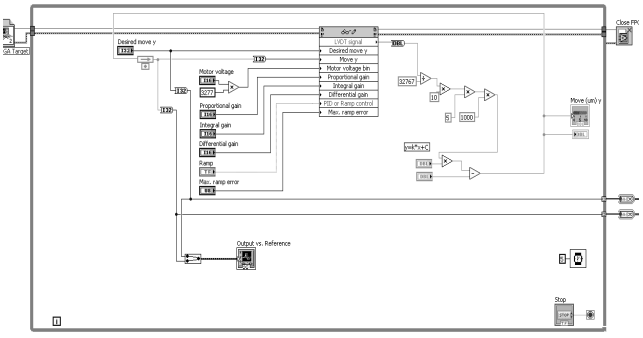


Figure 4. Host VI block diagram

SCXI system in a multiplex mode [18]. The NI PXI-8196 embedded controller is a modular PC in a PXI running on Windows XP OS [19].

The NI PXI-7833R FPGA module with a Virtex-II 3M gate FPGA chip is equipped with a reconfigurable input/output (RIO) hardware which allows processing of data on the card. It has eight analog 16-bit inputs, eight analog 16-bit outputs and 96 digital lines configurable as inputs, outputs, counters, or custom logic at rates of up to 40 MHz. Analog inputs and outputs have a full scale range of $V_{AO/I} = \pm$ 10 V [20]. The analog input 0 (AI0) of the FPGA module is used to connect the LVDT position feedback sensor of the developed mechatronics system. The obtainable theoretical resolution r of the LVDT can thus be calculated as:

$$r = \frac{l}{2^{16}} = \frac{100\,mm}{65535} \approx 1.5\,\mu m \qquad (1)$$

where l is the measurement range of the used LVDT.

On the other hand, the analog output AO0 is used to drive the DC actuator. Since each analog output of the NI PXI-7833R module can deliver a maximum output current of 25 mA, it cannot directly power the DC actuator which has a no-load current I_0 = 20 mA and a stall current of approximately I_H = 400 mA. An inverting Texas Instruments LM675 power operational amplifier with a maximum output current I_{max} = 3 A and a voltage gain set to A = 1 is thus used [21]. The value of the voltage gain is set to one since the maximum actuator voltage is 6 V and the output of the FPGA module has a full scale range of 10 V. Only current amplifying is thus needed in order to drive the selected DC actuator.

The control algorithms are programmed in the LabVIEW program environment [22]. The latter consists of two main virtual instruments (VI [23]): the Host VI and the FPGA VI. The Host VI is executed on the host computer (the PXI-8196 embedded controller) and includes user controls and indicators. The FPGA VI is designed and compiled on the host PC and exported to the FPGA module. Each VI has its front panel that consists of user controls and indicators, and its block diagram, which consists of the coded parts of the VI. The Host VI front panel is shown in Figure 3 while its block diagram is given in Figure 4.

The FPGA VI is used to implement the control algorithms, i.e., in the herein considered case, the PID and the ramp control algorithms described in section III.

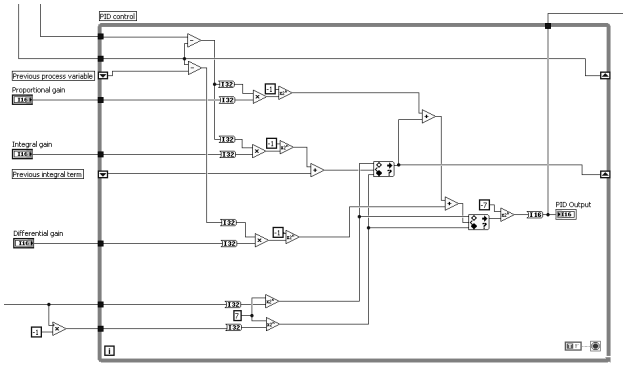


Figure 5. PID control FPGA VI block diagram

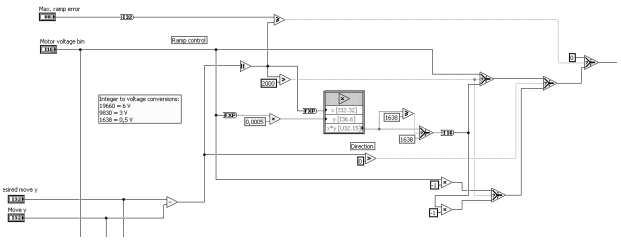


Figure 6. Ramp control FPGA VI block diagram

In Figure 5 the PID version of the FPGA block diagram is depicted, while in Figure 6 the ramp algorithm, implemented as the FPGA block diagram, is shown.

The independence of the host computer (where the Host VI is running) and the FPGA VI is one of the main advantages of systems based on FPGA architecture. In fact, if the host computer stops to operate, the user will not be able to read the indicator values or interactively control the system, but the FPGA VI will still be running and executing its program.

D. Laser Interferometric System

In this work the mobile Lasertex LSP 30-3D [24] Michelson-type laser Doppler interferometric system is used to assess the positioning accuracy and repeatability achieved by the used mechatronics system. This is a two-frequency laser interferometer which can be used in optical set-ups for linear and angular positioning or velocity measurements, as well as for straightness, flatness or squareness measurements. The available sensors allow compensating in real-time the environmental influences due to pressure, temperature and humidity variations during the measurements. Mounting and alignment errors such as the dead path, cosine and Abbe errors [25] have been minimized by performing a careful iterative mounting procedure.

To perform the needed nanometric measurements, a linear optical set-up is used. As shown in Figure 7, it is made of a laser head with an integrated receiver, the linear interferometer (reference cube-corner) and the linear cube-corner retroreflector. In the specific application, the laser head and the interferometer are mounted on the static base, while the retroreflector is mounted on the moving part (see Figure 1a).

III. CONTROL METHODS

High speed closed-loop feedback control is obtained by using a PID and a ramp control algorithm.

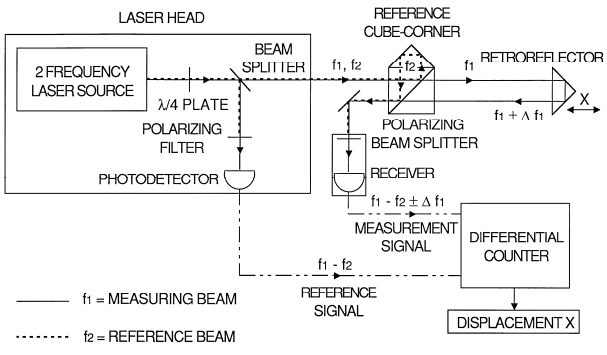


Figure 7. Michelson-type laser Doppler interferometric system

A. PID Controller

PID controllers have been widely used for many years in various industrial applications and control systems; in fact, PID is the most commonly used feedback controller. It is a closed-loop feedback control typology based on three terms – proportional (K_P), integral (K_I) and derivative (K_D). These parameters, known as PID gains, multiply in the feedback loop the error determined by the feedback sensor of the considered system and have different numeric values depending on the system architecture and its operating conditions [26].

The error value $e(t)$, represented by the difference between the set reference position (set point) $y_0(t)$ and the reached position (process value – PV) $y(t)$, measured via the feedback sensor, can be expressed as:

$$e(t) = y_0(t) - y(t) \tag{2}$$

In the time domain, the output of the PID control algorithm $u(t)$, can then be defined as [26]:

$$u(t) = K_P \cdot \left(e(t) + \frac{1}{T_I} \cdot \int_0^t e(\tau) \cdot d\tau + T_D \frac{d\,e(t)}{dt} \right) \tag{3}$$

where T_I is the integral time constant, T_D is the derivative time constant and τ is the integrating variable that can vary in the time interval from 0 to the actual time t.

To get the discrete form of the controller, the integral and the derivative terms are approximated as [27]:

$$\int_0^t e(\tau) \cdot d\tau \approx T \cdot \sum_{k=0}^{n} e(k)$$

$$\frac{d\,e(t)}{dt} \approx \frac{e(n) - e(n-1)}{T} \tag{4}$$

$$t = n \cdot T$$

where $e(n)$ is the discrete proportional and derivative error term, $e(k)$ is the integral error term, t is time, k is the summation variable for the discrete integral term, n is the discrete time step, and T is the signal sampling period.

The integral and derivative PID gains can hence be calculated from the integral and derivate time constant as [27]:

$$K_I = \frac{K_P \cdot T}{T_I}, \quad K_D = \frac{K_P \cdot T_D}{T} \tag{5}$$

Equations (3-5) make it possible to express the discrete PID controller as [27]:

$$u(n) = K_P \cdot e(n) + K_I \sum_{k=0}^{n} e(k) + K_D \left[e(n) - e(n-1) \right] \tag{6}$$

where $u(n)$ is the output from the discrete controller.

To avoid that a change in the defined set point (in this case the reference position) induces unwanted rapid changes (i.e. sharp spikes or derivative kicks) in the control signal, the controller is improved by multiplying the process value (defined as the reached position in two consecutive time steps – i.e. not the error term) by the derivative gain. These rapid changes could otherwise create serious problems for electronic circuitry [28]. Finally, the discrete PID algorithm, which can be implemented on the FPGA module, can be expressed as [28]:

$$u(n) = K_P e(n) + K_I \sum_{k=0}^{n} e(k) + K_D \left[y(n) - y(n-1) \right] \tag{7}$$

where $y(n)$ and $y(n–1)$ are the reached positions measured via the feedback sensor in two subsequent discrete time steps. The block diagram of the digital PID controller implemented on the FPGA module is shown in Figure 5.

It is worth noting here that various methods are commonly used for tuning the PID gains, some of them being the Ziegler Nichols, the Cohen-Coon or the Tyreus-Luyben tuning method [29]. In this work the tuning of the PID controller parameters is conducted in two steps. First, the Ziegler-Nichols method is used to achieve a rough estimate of the gains; in a second step, an experimental method of fine-tuning of the PID parameters is performed.

B. Ramp Control

The ramp control algorithm is based on a linearly decreasing driving voltage proportional to the distance of the moving part to the reference position. Maximum voltage $V_{R\max}$ is therefore delivered to the actuator up to the point distant two millimetres from the reference point. The input voltage is then linearly decreased from the maximum available value to a user-defined minimal value (in the considered case, this value is $V_{R\min} = 0.5$ V so as to enable the overcoming of mechanical nonlinearities such as sticktion). When the moving part finally reaches the reference position – within the limits of the defined maximal acceptable positioning error, the driving voltage is set to zero. The whole cycle is adapted to the operation of the used mechatronics system so as to avoid overshoots when reaching the reference position. The used ramp control function is shown in Figure 8.

IV. EXPERIMENTS

To assess the performances of the described mechatronics system, whose control is based on FPGA architecture, in achieving micropositioning, a set of point-to-point experiments using both the PID and the ramp control typologies is performed.

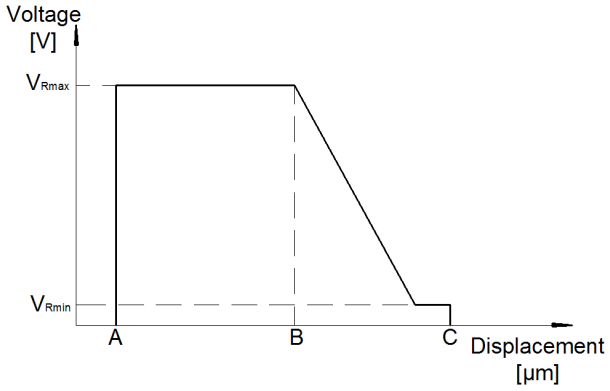


Figure 8. Ramp control function

TABLE I. MEASUREMENTS FOR 100 μm STEPS – ERRORS IN μm

	PID	Ramp		PID	Ramp
Point no.	Error	Error	Point no.	Error	Error
1	-2.8	-1.2	6	-6.2	-1.8
2	-7.6	-0.8	7	3.9	-0.8
3	-2.5	-1.9	8	-1.1	-1.1
4	-1.1	2.5	9	-6.5	0.4
5	-6.5	3.0	10	-4.6	-1.5

In a first instance, point-to-point positioning in a micrometric range of displacements, i.e. with 100 μm steps is implemented. Long range, ten millimetre steps are executed next.

A. Hundred Micrometre Displacements

In this set of experiments, each measurement is repeated 10 times. The mechatronics system performs in each move a displacement from the initial position to the reference position in 100 μm steps. The first set of measurements is performed by employing PID control, the second one using the described ramp control algorithm. To obtain optimal system response, the PID gains are set to the following values: $K_P = 4700$, $K_I = 600$, $K_D = 190$.

The thus obtained measurement results in terms of the errors in μm are given in Table I. It can hence be calculated that, in the case when PID control is used, positioning accuracy and repeatability are, respectively, 2.68 μm and 3.07 μm. When the ramp control algorithm is used instead, these values are 1.75 μm and 0.32 μm.

B. Ten Millimetre Displacements

When longer travel ranges are implemented, the system output results in a marked nonlinearity which significantly influences the resulting positioning error. In fact, from the results listed in Table II it is evident that, when PID control with the same gains as in the previous case is used, positioning accuracy and repeatability are, respectively, 44.34 μm and 3.53 μm. With ramp control, accuracy is 48.62 μm, while repeatability is 3.37 μm.

C. Ten Millimetre Displacements with Linearization

A careful analysis of the system architecture allowed establishing that the LVDT sensor has the main influence

TABLE II. MEASUREMENTS FOR 10 mm STEPS W/O LINEARIZATION – ERRORS IN μm

	PID	Ramp		PID	Ramp
Point no.	Error	Error	Point no.	Error	Error
1	37.3	47.4	6	49.4	55.5
2	48.8	50.3	7	46.4	52.7
3	44.5	49.8	8	42.5	48.5
4	40.3	46.6	9	45.1	44.1
5	42.9	46.7	10	46.2	44.6

TABLE III. MEASUREMENTS FOR 10 mm STEPS WITH LINEARIZATION – ERRORS IN μm

	PID	Ramp		PID	Ramp
Point no.	Error	Error	Point no.	Error	Error
1	1.1	0.6	6	3.6	3.6
2	0.7	2.4	7	3.2	-0.5
3	2.6	-2.7	8	2.2	0.2
4	-0.6	-2.6	9	-1.7	-1.2
5	2.4	-3.6	10	-1.7	-3.5

on the observed system's nonlinearity in the case of longer travel ranges. By using the laser interferometric system, repetitive measurements with 1 mm steps in the whole 0 – 10 mm range, coupled with an interpolating procedure implemented in the Matlab software [30], allowed determining that the linearization function to be applied to the values of displacements x measured via the LVDT has the form:

$$f(x) = 1{,}006 \cdot x + 7 \tag{8}$$

This linearization function is hence programmed in the LabVIEW Host VI and, considering that the linearization function causes a different behaviour of the system, the PID gains are slightly modified ($K_P = 4800$, $K_I = 700$, $K_D = 350$). Long travel range experiments are hence performed again and the obtained results are given in Table III. It is clear that a marked improvement is obtained and a positioning accuracy of 1.18 μm with a repeatability of 1.86 μm can be calculated. In the case of ramp control, positioning accuracy and repeatability are, respectively, 0.73 μm and 2.34 μm.

V. CONCLUSIONS AND OUTLOOK

In this work a single-axis micropositioning mechatronics system is developed. Its mechanical design is optimized in order to achieve high precision positioning. High speed closed-loop feedback control is obtained by using PID and ramp control algorithms programmed as a VI on an FPGA module. User controls and indicators are programmed in an independent Host VI. An LVDT is employed as a position feedback sensor, while positioning accuracy and repeatability are experimentally assessed by using a laser interferometric system.

Numerous experiments with point-to-point positioning are conducted. In the first set of experiments short range positioning with 100 μm steps is validated. The calculated positioning accuracies and repeatabilities are always within 3 μm. Ramp control gives roughly two times better results than PID control.

In the case of long (10 mm) travel ranges, a marked nonlinearity is observed. It induces errors in the range of 40 – 50 μm. The nonlinear effect, caused mainly by the LVDT feedback sensor, is characterised via interferometric measurements and compensated via system linearization. The resulting linearizing analytical function is thus programmed in the LabVIEW Host VI. Positioning accuracy and repeatability are finally assessed

experimentally again, and their values are limited to roughly 2 µm, both in the case of the PID and of the ramp control algorithms.

In a future work, the authors plan to investigate the usage of more complex control typologies, e.g. by implementing on the FPGA module a pulse-width modulation (PWM) based control. Means of optimising the control algorithms so as to maximise positioning accuracy and repeatability will also be investigated. Other types of feedback transducers, e.g. optical encoders, will be tested as well. Finally, multi-axes micropositioning mechatronics systems based on the described FPGA architecture will be developed. Their foreseen applications are in machine tools, devices for handling and assembly of microsystems, robotics, metrology, high-end optical system, scientific instrumentation, ICT, medicine and the growing micro-electro-mechanical (MEMS) devices field.

LIST OF SYMBOLS

A	power operational amplifier voltage gain [V/V]
b	ball bearing width [mm]
B	ball screw nominal maximum backlash [µm]
d	ball bearing inner diameter [mm]
d_0	ball screw nominal diameter [mm]
d_1, d_2	coupling inner diameters [mm]
D	ball bearing outer diameter [mm]
D_C	coupling nominal outer diameter [mm]
$e(k)$	value of integral error term in the discrete domain [mm]
$e(n)$	value of proportional and derivative error terms in the discrete domain [mm]
$e(t)$	error value [mm]
$f(x)$	LVDT linearization function [mm]
i	planetary gearhead reduction ratio
I_H	DC actuator stall current [mA]
I_{max}	power operational amplifier maximum output current [A]
I_0	DC actuator no-load current [mA]
k	summation variable for the discrete integral term
K_P, K_I, K_D	proportional, integral and derivative controller gains of the PID controller
l	LVDT measurement range [mm]
l_C	coupling length [mm]
$l_f/w_f/h_f$	guideways' fixed part dimensions (length/width/height) [mm]
$l_m/w_m/h_m$	guideways' moving part dimensions (length/width/height) [mm]
L	DC motor length [mm]
L_g	planetary gearhead length [mm]
n	discrete time step
n_0	DC motor no load speed [rpm]
p	ball screw lead [mm]
r	theoretical LVDT resolution [µm]
t	time [s]
T	sampling period [s]
T_I, T_D	integral and derivative time constants [s]
$u(t)$, $u(n)$	output of the controller [V]
U_N	DC motor nominal voltage [V]
$V_{AO/I}$	FPGA module analog input/output full scale voltage [V]
$V_{BICM\,max}$	BICM full scale range output DC voltage [V]
V_e	LVDT excitation voltage [V]
V_{Rmax}	maximal voltage of ramp control algorithm [V]
V_{Rmin}	minimal voltage of ramp control algorithm [V]
x	current LVDT position (displacement) [mm]
$y(t)$, $y(n)$	reached position (process value) [mm]
$y_0(t)$	reference position (set point) [mm]
η_p	ball screw practical efficiency [%]
τ	integration variable
ϕ	DC motor diameter [mm]

REFERENCES

[1] S. Fukada, B. Fang and A. Shigeno, "Experimental analysis and simulation of nonlinear microscopic behaviour of ball screw mechanism for ultra precision positioning," Precision Engineering, vol. 35/3, pp. 650-668, 2011.

[2] K. H. Kim, P. J. Kim and K. H. You, "Ultra-precision positioning system using robust sliding mode observer and control," Precision Engineering, vol. 37/1, pp. 235-240, 2013.

[3] S. Zelenika and F. De Bona, "Nano-positioning using an adaptive pulse width approach," J. Mechanical Engineering Science – Proc. Institution Mech. Engineers, vol. 223/C8, pp. 1955-1963, 2009.

[4] B. A. Bucci, J. S. Vipperman, D. G. Cole and S. J. Ludwick, "Evaluation of a servo settling algorithm," Precision Engineering, vol. 37/1, pp. 10-22, 2013.

[5] E. Bin Linggin, "DC motor position control with PID controller," Fakulti Kejuruteraan Elektrik, Kolej Universiti Teknikal Kebangsaan Malaysia, 2005.

[6] URL: http://www.ni.com/fpga

[7] Y. P. Kondratenko and E. Gordienko, "Implementation of the neural networks for adaptive control system on FPGA," Annals of DAAAM & Proc. 23rd Int. DAAAM Symp., vol. 23, pp. 389-392, 2012.

[8] B. Mac Cleery and Z. M. Kassas, "New mechatronics development techniques for FPGA-based control and simulation of electromechanical systems," Proc. 17th World Congress – Int. Federation Automatic Control, Seoul, pp. 4434-4439, 2008.

[9] F. H. Ali, M. M. Hussein and S. M. B. Ismael, "LabVIEW FPGA implementation of a PID controller for DC motor speed control," Iraq J. Electrical and Electronic Engineering, vol. 2, pp. 139-144, 2010.

[10] URL: http://www.faulhaber.com/uploadpk/EN_1724_SR_DFF.pdf

[11] URL: http://www.omega.com/pptst/LD610.html

[12] URL: http://www.omega.com/manuals/manualpdf/M3342.pdf

[13] S. Fukada, M. Matsuyama and J. Hirayama, "Ultraprecision positioning with sub-nanometer resolution by using ball screw and aerostatic guide way," Proc. 12th Euspen Int. Conf., Stockholm, pp. 306-310, 2012.

[14] URL: http://www.skf.com/files/264744.pdf

[15] URL: http://www.skf.com/skf/productcatalogue/Forwarder?newlink=&action=search&lang=en&prodid=1010026184

[16] URL: http://in.misumi-ec.com/asia/ItemDetail/10300119820.html

[17] URL: http://www.schneeberger.com/products/linear-and-profiled-guideways/profiled-miniature-guideway/minirail-mini-x/#c2106

[18] URL: http://sine.ni.com/nips/cds/print/p/lang/hr/nid/201565

[19] URL: http://www.ni.com/pdf/products/us/20055052501101iLR.pdf

[20] URL: http://sine.ni.com/ds/app/doc/p/id/ds-98/lang/hr

[21] URL: http://www.ti.com/lit/ds/symlink/lm675.pdf

[22] URL: http://sine.ni.com/np/app/main/p/docid/nav-104/lang/hr

[23] URL: http://www.ni.com/pdf/manuals/373427c.pdf

[24] URL: http://lasertex.eu/pages,download.htm

[25] C. R. Steinmetz, "Sub-micron position measurement and control on precision machine tools with laser interferometry," Precision Engineering, vol. 12/1, pp. 12-24, 1990.

[26] G. F. Franklin, J. D. Powel and A. Emami-Naeini, "Feedback control of dynamic systems," Addison-Wesley series in electrical and computer engineering: control engineering, 2nd ed., 1991.

[27] URL: http://lorien.ncl.ac.uk/ming/digicont/digimath/dpid1.htm

[28] URL:http://www.palgrave.com/science/engineering/wilkie/sample/0333_77129Xcha18sample.pdf

[29] URL:http://www.ee.up.ac.za/main/_media/en/undergrad/subjects/ebt410/skogestad_simple_pid_tuning_rules.pdf

[30] URL: http://www.mathworks.com/help/signal/ref/interp.html

On-water measurement and analysis system for estimating kayak paddlers' performance

G. Vadai*, G. Makan*, Z. Gingl*, R. Mingesz*, J. Mellár*, T. Szépe*, and A. Csamangó**
* Department of Technical Informatics, University of Szeged, Szeged, Hungary
** EDF DÉMÁSZ Szeged Water Sport Association, Szeged, Hungary
E-mail: vadaig@inf.u-szeged.hu

Abstract - In recent years, the use of computerized measurement systems to help coaching professional kayak paddlers has become rather common. Note that it is not possible to evaluate the quality of the paddling exactly, however the measured inertial signals can provide useful quantitative information. Therefore, it is a challenge to find really good indicators that could help the trainers. Keeping these in mind we have developed a microcontroller-based expandable system that can record the kayak's motion accurately using the built-in 3-axis accelerometers, gyroscopes and various optional sensors. Several hours of paddling can be recorded and the off-line analysis can be done using a PC. After examining many athletes with different age and technical skills at training and races, we have evaluated the commonly used indicators and defined useful additional indicators based on the statistics and fluctuations of the parameters obtained with peak search algorithms and frequency–domain analysis. .

I. INTRODUCTION

Modern training and professional coaching are supported by various devices that can measure and evaluate inertial and biomedical signals, forces and even more [1-9]. There are common analysis techniques that can be easily applied and understood like higher velocity, higher force, elevated heart rate, and so on. However, in a race less than a second can make the difference even though the technique of the athletes can be rather different. During a race the integral of the quantities matter in many cases, a simple example is the distance as the integrated instantaneous velocity. All sensors have a certain accuracy, non-linearity and noise that affect the overall accuracy of the measurement. On should not forget many additional circumstances like weather, mental condition, support of the fans. The above mentioned facts mean that it is really challenging to extract quantitative information that can be used to distinguish between racing techniques at a professional level.

It is very unlikely that one can make so accurate measurements and analysis that can express clearly the differences of different techniques, it seems to be a better approach to find indicators based on the averaging and other statistical processing of the sensor signals. Of course, trend of these averaged quantities with a certain time resolution can be very informative as well. Engineers and scientists of different fields work together with coaches and athletes to find the most useful indicators that are simple, informative enough and easily readable – a few numbers, some curves.

This work was supported by DEAK Plc., EDF DEMÁSZ Zrt., TÁMOP-4.2.2/B-10/1-2010-0012.

We have followed the above mentioned approach to develop a special experimental system to evaluate and analysis the technique of kayak paddlers. Our device is universal, allows connection of many different sensors and the analysis is done off-line after the data acquisition. There are devices on the market that can do real-time measurements and analysis, can provide a feedback for the athlete, allow the coach to monitor the signals [7, 8], however according to the opinion of the coaches and athletes the information they can get is still far from ideal, only partial usage is common. Our aim is to provide a hardware and software platform that supports extensive research. In collaboration with coaches we have carried out several measurements and proposed indicators that can be used to evaluate the efficiency of the paddlers.

II. MEASUREMENT SYSTEM

Our device is based on a mixed-signal microcontroller, the C8051F581 [10]. This chip integrates precision analogue peripherals including a multichannel 12-bit analogue-to-digital converter, voltage reference and precision internal oscillator. High performance digital communication peripherals allow glueless connection of memory cards and provide easy interfacing to a host computer. Wide variety of analogue and digital output sensors can be used without additional signal conditioning electronics.

The block diagram and the photo of the device are shown on Fig. 1 and Fig. 2, respectively.

A three-axis accelerometer and a three-axis gyroscope are integrated on the printed circuit board as a basic set of inertial sensors. Since the sample rate is kept much higher – 1000 Hz per channel – than in commercial devices [7-9], the aliasing and noise reduction could be performed by simple one pole passive filters with cutoff frequencies of 50 Hz, on the other hand, further digital signal analysis can be applied to determine the optimum sample rate later.

We have chosen inertial sensors (LIS352AX, LPR530AL, LPR503AL, LY503ALH [11]) considering low noise, proper signal range and availability. We have selected sensors with the following dynamic ranges: ±2 g for the three-axis acceleration, ±300 °/s for the roll axis angular velocity, ±30 °/s for the pitch axis and yaw axis angular velocities, respectively.

Typical calibrated accuracy of the sensors was below 3%, which was sufficient for our estimation, especially

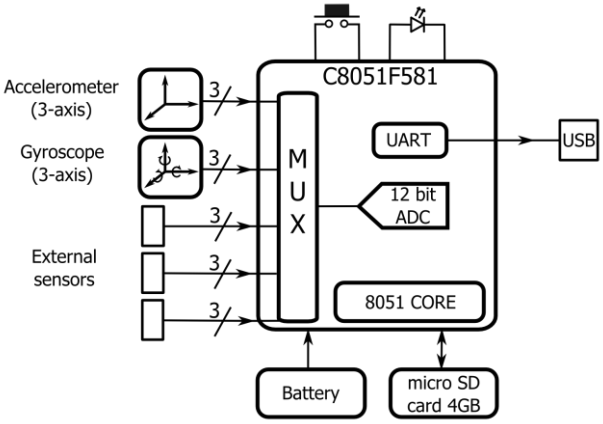


Figure 1. Block diagram of the developed device

because our analysis is rather based on the relative changes and the timing of the signals. Examples include the stroke rate, the symmetry factor and pull time.

The device incorporates a rechargeable battery and charger circuitry that is enabled upon connection to the host computer via the external USB 2.0 interface dongle. The applied Li-ion 3.6 V, 1.2 Ah battery ensured 8 hours of active operation. We have used a 4GB micro SD card for data storage that allowed data acquisition with high sample rate and many channels for hours. The microcontroller code is written in C and compiled with the open source SDCC compiler. The embedded software measures the raw signal data based on the parameters (sample rate, included channels, configuration) received from the host computer. It monitors a simple button to start and stop the acquisition and provides audible and light signals to inform the user about the status.

The device is housed in a water-proof aluminum enclosure and can be mounted easily under the seat. Besides the six degrees of freedom inertial sensors there are three available universal sensors ports. These five-pin ports provide power for external sensors, can accept both analogue and digital sensor output signals and can be flexibly set up to accommodate the needs. We have used one of the ports for monitoring the instantaneous velocity, but paddle force sensors or biomedical sensors can also be connected.

The host computer software is written in JAVA and

Figure 2. The developed device

can communicate with the device via the USB 2.0 port. The application allows setting up the sample rate, selecting the sensor signals to be recorded, setting the configuration for external sensors. The measured data can be uploaded form the device and can be used for extensive signal processing. The simple and informative graphical user interface helps the user to visualize and analyze data quickly.

In the following we show the main implemented analysis functions and evaluated indicators briefly. Our results we obtained using a large set of experiments with many paddlers.

III. PROCESSING THE MOTION SIGNALS

As it was mentioned already the sampling frequency of the motion signals was set to 1000 Hz in our measurements. This value provides sufficient accuracy for motion analysis and allows to modify the method of evaluation and do further processing or filtering procedures in the digital domain.

The built-in six degrees of freedom inertial sensors (three-axis acceleration, three-axis angular velocity) were used as follows. Naturally, the forward (x-axis) acceleration signal plays the most important role in the analysis of the motion, in addition, information about other directions displacement is needed for accurate evaluation of paddling technique. However, measuring acceleration in the other two directions is rather sensitive to mechanical effects and the rotations of the kayak, so the measured signals were quite noisy. On the other hand, the three-axis angular velocity that is measured by gyroscope, was much less noisy, so it suited better for detecting the main characteristics of the movements, helping also in the interpretation of the forward acceleration signal and determining the estimated parameters with sufficient accuracy, as will be shown later.

Video analysis was used to interpret the measured x-axis acceleration, whereby the signal of pitch axis angular velocity (kayaks "nodding") was very useful to compare video frames with the signals (see Fig. 3). A stroke is divided for two phases: underwater phase (which have "catch", "power" and "exit" stages), and air-work recovery phase [4]. As it can be seen, the placing of the

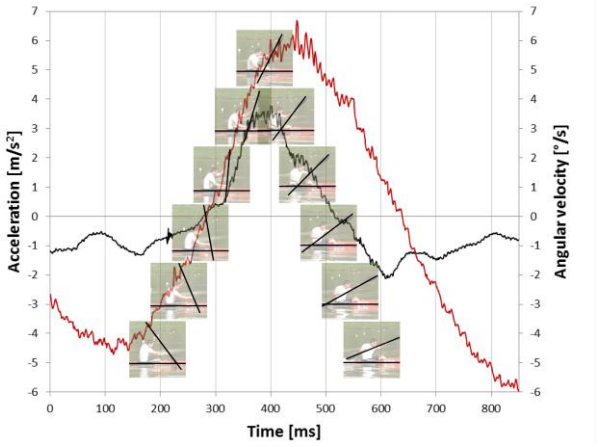


Figure 3. Comparison of video frames with measured signals using pitch axis angular velocity (red curve), to interpret x-axis accerelation (black curve)

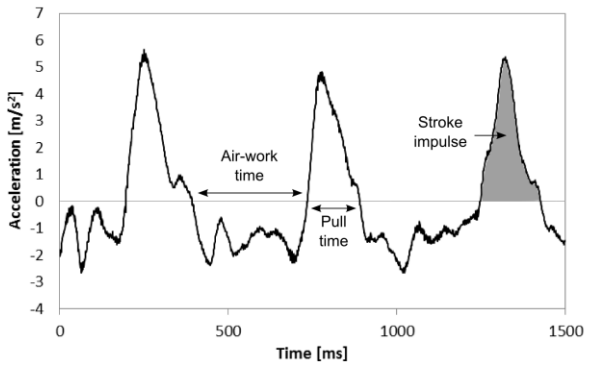


Figure 4. Parameters of a stroke cycle on the x-axis acceleration

paddle into the water happened in the negative acceleration range, however, defining the pull phase as the positive parts of x-axis acceleration signal could be a good approximation, that supposes the power stage starts at zero-intersection of x-axis acceleration.

Therefore, a period of the x-axis acceleration signal, i.e. a stroke cycle is divided to two parts: the time of *pull* and the time of *air-work*, as it demonstrated on Fig. 4. Using the time of a stroke cycle we can calculate the *stroke rate*, which is very important parameter for trainers in the kayak-canoe sport. Integrating the positive (pull) part of the signal, we can calculate the *stroke impulse* (as in [3]) (see Fig. 4), which is a speed-dimension quantity that could be interpreted as the speed-growth caused by the actual pulling. Multiplying it with the weight of the paddler and the boat, we can calculate the mechanical impulse, however, as we can see hereinafter, this is not necessary for estimating the paddlers performance.

These four physical quantities provide the basis of paddlings' analysis. Using yaw axis gyroscope signal (rotation around vertical axis) we can detect easily, the actual stroke was produced by left or right hand, so we can examine the symmetry of the mentioned quantities. Furthermore, examination of the fluctuations and trends of a longer training section or a race provides further parameters and indicators, as will be discussed later. Therefore, sufficient accuracy at the calculations of these quantities is substantial.

Fig. 4 and Fig. 5 (a) show an x-axis acceleration signal of a professional athletes (Paddler 1) paddling, who has a clear, good technique and the rowing performed at training, therefore the basic parameters are rather easily definable. Unfortunately, in many cases the signals are more complex, for example at a 200 m sprint race, the detection of the strokes is a more difficult task. Furthermore, in some paddlers' signals, a negative acceleration period shows up on the middle of the pull phase of a stroke, and one or more positive period on the air-work phase, as depicted on Fig. 5 (b), which shows Paddler 2's paddling at training. This effect, which was discussed by another work [3], indicates a technical fault, therefore its presence is useful information for trainers.

On the other hand, with this effect it is easy to come across detecting errors, such as detecting two separate strokes instead of one real pull phase. To avoid these errors, our algorithm filters out the wrong peaks by using time limits and the gyroscope signals, which are much less sensitive to the mentioned effects. Of course, this method makes the calculation of parameters defined above less accurate.

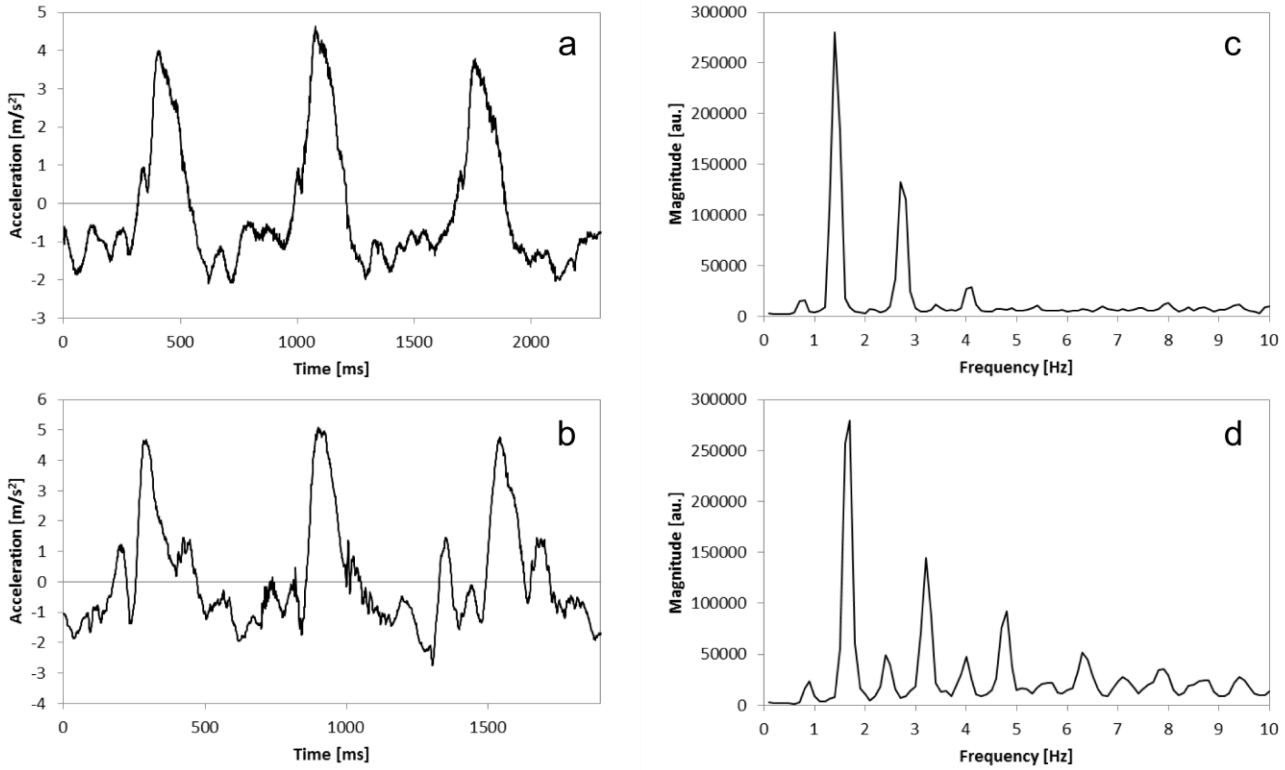


Figure 5. Forward acceleration signals of technically good (a) and faulty (b) paddlings (produced by Paddler 1 and Paddler 2, respectively) and a plots of them in frequency domain (c, d), which were calculated on an 80 seconds range with FFT algorithm using Hanning window after averaging over 10 seconds long intervals.

Besides detecting technical faults, its characterization is useful for athletes and trainers. On Fig. 5, we can see both paddlers respectively with technically good and faulty paddling, and a plots of them in frequency domain, which were calculated on an 80 seconds range with Fast Fourier Transform (FFT) algorithm using Hanning window after averaging over 10 seconds long intervals. As depicted on Figure 5, in case of the technically faulty paddling, we can observe much more intensive harmonics, therefore the factor representing the harmonic distortion can be used to characterize the technical faults.

Frequency domain analysis could help at calculating other useful parameters also, for example to calculate the paddling symmetry factor. As shown on Fig. 5 (c-d), the highest peak is the first harmonic, which represents the frequency of one stroke cycle. The fundamental frequency (which has much smaller magnitude) represents a period, where a left and a right hand stroke occurs. Therefore, the ratio of these two frequencies provides information about symmetry of the paddling. Examining the other measured signals at frequency domain provides further possibilities, such as analysis of transient periods of paddling (start, sprint), which will be discussed in detail in our subsequent work [12].

IV. PERFORMANCE ESTIMATION

The interpretation of the measured paddling signals and the analysis of a stroke cycle are very complex and have been discussed in several publications [13-19]. Nevertheless, at everyday work of coaches and athletes, fast and easily useable but still accurate indicators are needed. Therefore, we were looking for informative and revealing trend curves and factors, which will help the involved coaches, these results are detailed below.

Plotting the above mentioned quantities' evolution in time is highly informative for the coaches. The task at training is producing a particular type of paddling and performance steadily, conversely at simulated or real competition the purpose is rawing a distance as fast as possible using a specific strategy.

Evidently enough, examination of stroke rates evolution is substantial. Because the detection of stroke cycles is needed for stroke rate calculation, the measured gyroscope signals are suitable for this task. Moreover, the angular velocity signals are not influenced by the mentioned technical faults, so these signals can produce more accurate results than x-axis acceleration in most cases. The trend curve of stroke impulse provides important information about the quality of pulls and paddlers' performance. This indicator allows to examine the stability of the performance at training and the strategy and fatigue at race. For raising velocity in a race situation, paddlers are often increasing the stroke rate, nevertheless it could make the strokes quality worse. Therefore it is interesting to examine the two curves together, as well as combining the analysis of the two quantities with introducing a total impulse/min metric, which denotes sum of impulses of strokes occurred in one minute. On Fig. 6 one can see two trend curves of a paddling of Paddler 2 at a 500 m race, without the transient phases at start and finish. The trends of stroke impulses shows apparently the fatigue of the athlete.

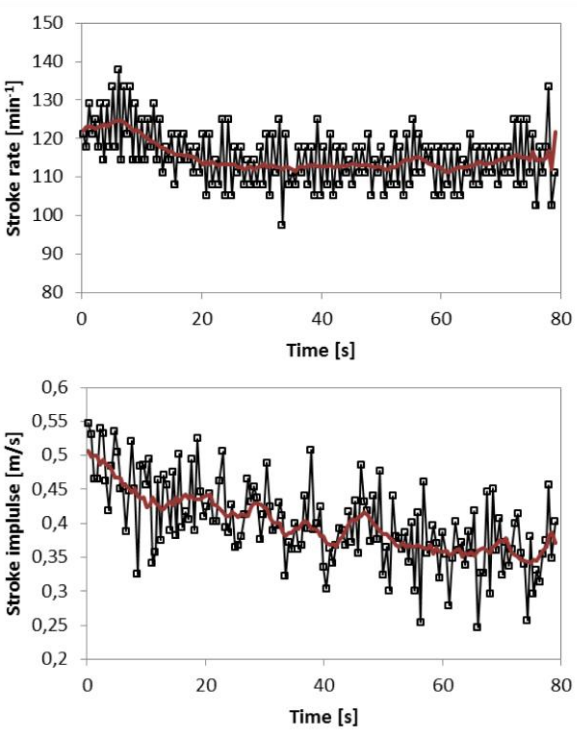


Figure 6. Trend curves calculated for paddling of Paddler 2 at a 500 m race, without the transient phases at start and finsih. While the stroke rate (upper plot) does not change significantly in time, the stroke impulse (lower plot) decrease indicates the fatigue of the athlete. Smoothed curves (red curves) was calculated using 10 samples long moving avarage.

Visualization of pull time, air-work time and their rate indicates the stroke rythm, and its stability. Comparison of paddlings produced at training and race situations showed that when the stroke rate increases, the air-work time decreases only, the time of pull phase does not change significantly. Plotting the stroke impulses of the two hands with two separate curves provides information about the symmetry of produced paddling. As depicted on Fig. 7, for professional athlete Paddler 1, the average impulse of left hand strokes more than 10% greater than impulse right hand strokes. The mentioned rates of stroke parts and left-right hand impulses are different for different athletes, therefore it characterizes the personal paddling technique.

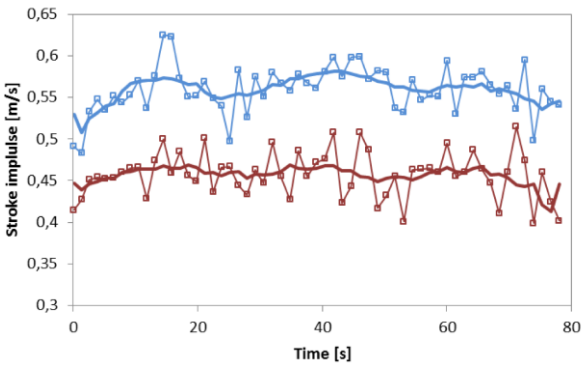


Figure 7. Trend curves of stroke impulses of left and right hands (blue and red curves, respectively) calculated for paddling of Paddler 1 at training. Smoothed curves (thicker curves) was calculated using 10 samples long moving avarage.

The developed PC software plots a smoothed version trend curves too, similar that were shown on Fig. 6 and 7. The smoothed curve was calculated using 10 samples long moving average. In addition, the mean and standard deviation of parameters for the examined stage of signals are also shown.

The classical parameters presented above allows the analysis of paddlings for strategy, technical and other aspects [4]. Nevertheless, the coaches, who supervise the work of many athletes every day, have no possibilities to do a detailed analysis in many cases. Easily understandable indicators and factors could help at quick evaluation of performance and at detecting of the technical faults. Searching helpful parameters and defining new indicators, statistical analysis of classical parameters was carried out on data of 26 athletes with different age, gender and technical skills. In all cases, a typical 80 second paddling phases have been studied, the transient stages were skipped.

Although technical skills are not characterized by the age of athletes perfectly, as well as the speed of technical improving different in different ages of paddlers, it allows demonstration of the evolution of the parameters and the detection of certain effects.

As it was mentioned above, the stroke rate, stoke rythm and paddling symmetry characterize personal paddling techniques and can be irrespective for technical skills. Naturally, the stroke impulse (and mechanical impulse) increases with the age of athletes, however the effect of stroke impulses' standard deviaton is more interesting. As Fig. 8 shows apparently, the standard deviaton of impulses decresases for older athletes. This effect is shown at examining the impulse trend curves of paddlers: while values of a world champion athletes stroke impulse are changing slightly, values of a young, less-practiced paddlers the stroke impulses vary significantly. Naturally, this effect influences the kayak's speed and position in the water and much more kinetic energy is dissipated by unoptimal movements. Consequently, a factor calculated from stroke impulses standard deviaton could be useful marking paddlers technical skills.

The standard deviaton of stroke impulses and relative standard devitaton of mechanical impulses show the mentioned effect too, therefore both impulse quantites are suitable. As we can see, the temporal fluctuations of the calculated physical parameters provides useful additional information about paddlers' performance. Detailed analysis of this fluctuations and more detailed parameter evaluation will be shown elsewhere [12].

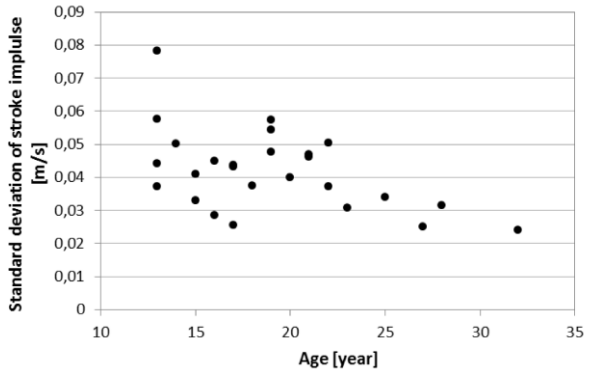


Figure 8. Standard deviaton of stroke impulse as a function of paddlers age.

The technical fault detection and characterization, as well as standard deviation of stroke impulse is clearly proved to be a useful indicator of paddling technique, while the trend curve and the total impulse/min proved to be useful in current performance assessment.

V. CONLUSION

We have shown a universal and flexible hardware based on six degrees of freedom inertial sensors and software platform especially developed for the evaluation of the performance of kayak paddlers. We have carried out several experiments, reviewed analysis methods and proposed additional evaluation methods to aid professional training. Our hardware solution provides universal ports for several additional external sensors and can be used in other sports including canoe movement analysis.

We have found that taking the raw six degrees of freedom inertial data can be efficiently used with statistical and spectral analysis to optimally describe most of the characteristics of paddling accurately enough, therefore not paddle force and other signal measurements are needed. The statistical and spectral parameters are discussed and proved to be rather straightforward and simple to use for the trainers and athletes as well. Further fluctuation analysis methods are planned to find more useful performance indicators.

ACKNOWLEDGMENT

The authors thank P. Kocsis, Prof. G. Szabó, I. Gyémánt, A. Dervarics, K. Petrovics and athletes, trainers for their help and valuable discussions.

The publication is supported by the European Union and co-funded by the European Social Fund. Project title: “Broadening the knowledge base and supporting the long term professional sustainability of the Research University Centre of Excellence at the University of Szeged by ensuring the rising generation of excellent scientists.” Project number: TÁMOP-4.2.2/B-10/1-2010-0012.

REFERENCES

[1] J. P. Stothart, F. D. Reardon, and J. S. Thoden, “A system for the evaluation of on-water stroke force development during canoe and kayak events,” 4th Int. Symp. Biomechanics in Sports, pp. 146-152, 1986.

[2] D. A. Aitken, and R. J. Neal, “An on-water analysis system for quantifying stroke force characteristics during kayak events,” Int. J. Sport Biomech., vol. 8, pp. 165-173, 1992.

[3] M. G. Robinson, L. E, Holt, T. W. Pelham, and K. Furneaux, “Accelerometry Measurements of Sprint Kayaks: The Coaches' New Tool,” Int. J. Coaching Sci., vol. 5, pp. 45-56, 2011.

[4] Z. Ma, J. Zhang, Y. Sun, and T. Mei, “Sports Biomechanical Information Acquisition and Evaluation for Kayaking Events,” Int. J. Inf. Acquisition, vol. 6, pp. 213-223, 2009.

[5] I. Janssen, and A. Sachlikidis, “Validity and reliability of intra-stroke kayak velocity and acceleration using a GPS-based accelerometer,” Sport Biomech., vol. 9, pp. 47-56, 2010.

[6] B. Gomes, N. Viriato, R. Sanders, F. Conceição, J. P. Vilas-Boas, and M. Vaz, “Analysis of the on-water paddling force profile of an

elite kayaker," 29th Int. Symp. Biomechanics in Sports, pp. 259-262, 2011.

[7] DigiTrainer measurement system, Polaritás-GM Ltd. http://www.polaritas.com/DigiTrainer_main

[8] Catapult minimaxB4 measurement system, Catapult Sports Pty Ltd. http://catapultsports.com/sports/rowing-kayak-canoe

[9] Excalibur Data Aquisiton Paddle, Merlingear http://www.merlingear.com

[10] http://www.silabs.com

[11] http://www.st.com

[12] G. Vadai, Z. Gingl, and R. Mingesz, "Performance estimation of kayak paddlers based on fluctuation analysis of movement signals," unpublished.

[13] A. G. W. Carter, J. P. Peach, T. W. Pelham, and L. E. Holt, "Discrete measures of C1 craft acceleration using various paddle designs," 12th Int. Symp. Biomechanics in Sports, pp. 190-193, 1994.

[14] C. López López, and J. Ribas Serna, "Quantitative analysis of kayak paddling technique: definition of an optimal stroke profile," Rev. Andal. Med. Deporte, vol. 4, pp. 91-95, 2011.

[15] J. Baker, D. Rath, R. Sanders, and B. Kelly, "A three-dimensional analysis of male and female elite sprint kayak paddlers," 17th Int. Symp. Biomechanics in Sports, pp. 53-56, 1999.

[16] R. H. Sanders, and J. S. Kendal, "A description of Olympic Flatwater Kayak Stroke Technique," Aust. J. Sci. Med. Sport, vol. 24, pp. 25-30, 1992.

[17] R. V. Mann, an J. T. Kearny, "A biomechanical analysis of the Olympic-style flatwater kayak stroke," Med. Sci. Sports Exerc., vol. 12, pp. 183-188, 1980.

[18] J. S. Michael, R. Smith, and K. B. Rooney, "Determinants of kayak paddling performance," Sports Biomech., vol. 8, pp. 167-179, 2009.

[19] L. McDonnell, P. Hume, and V. Nolte, "Sprint kayaking stroke rate reliability, variability and validity of the digitrainer accelerometer compared to GoPro video measurement," 30th Int. Symp. Biomechanics in Sports, pp. 316-319, 2012.

Random Interleaver Design in Turbo Coding for Digital Magnetic Recording Channel

St. Stancescu, C. Sandu
University POLITEHNICA Bucharest,
Iuliu Maniu 1-3 Blvd , Bucharest, ROMANIA
E-mail: stst@elia.pub.ro, catal1981@yahoo.com

Abstract - Turbo codes interleaver logic is analyzed in Digital Magnetic Recording (DMR) Partial Response (PR) channels environment. The performances of turbo encoding/decoding process are estimated, in the presence of appropriate DMR noise, generally a particulate noise, manifested as burst errors of various intensities and dimensions. Depending of interleaver, turbo code performance (Bit Error Rate BER) is estimated, in presence of noise, over variations with its parameters. Optimal code length and structures design recommendations are obtained, related to required recording densities and Signal Noise Ratio (SNR) ranges.

Index terms: turbo coding, interleaver, magnetic recording

I. INTRODUCTION

Partial Response PR channel was first introduced in digital communication systems, in view to combat inter-symbol interference ISI. The same ISI problem occurs in the digital magnetic recording systems, accentuated at high record density. The partial response channel do not eliminate the ISI, it is kept under control. In digital magnetic recording DMR a combination between PR equalization and maximum likelihood ML detection produce the well known Partial Response Maximum Likelihood (PRML) channel. In turbo coding, the classical code is replaced with a parallel concatenation coding/decoding scheme. The change is analyzed with a PRML DMR channel simulator [3] and effects on the BER at various values of SNR and record density are evidenced.

The general scheme of a typical PRML channel is given in Figure 1. We replace classic Run Length Limited encoders - RRL with TURBO encoder (RSC Reed Solomon Coding encoder) combined with a MAP - Maximum A Posteriori decoding algorithm.

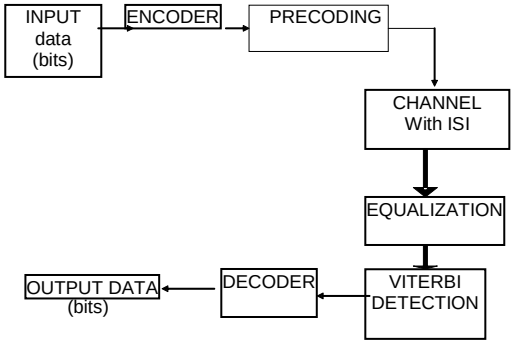


Figure 1: Block scheme of a typical PRML channel.

Section II provides a brief overview of parallel concatenation Turbo schema. The simulator is presented in Section III. Section IV will give the optimal interleaver choices. Simulation results and final remarks are given in section V.

II. INTERLEVER IN TURBO CODING

The purpose of DMR encoding is to protect stored information, in the event of material permanent or random defects, caused by electromechanical or electromagnetic phenomena as noise sources. The TURBO schemas are obtained by interleaving outputs of 2 Recursive Systematic Coders, RSC as in [1], working in parallel, as shown in Figure 2, where x, input data (binary), is permuted as interleaved data (input bits rearranged), output bit symbols v0 are same as x, output bit symbols v1 are generated from RSC Encoder 1 and output bit symbols v2 are generated from RSC Encoder 2.

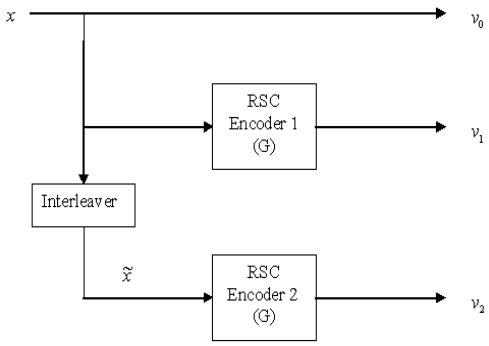


Figure 2: Turbo coding schema[1].

The interleaver important role in turbo coding is to induce a sufficient and uniform distance between the bit positions possibly affected by noise. Large burst errors are difficult do decode and correct, due to major error distances. The inteleaver before the second encoder spreads uniformly errors and avoid these difficulties by decongestioning decoding and correcting efforts. This leads to the possibility that the remaining errors from one encoder output may be corrected from the other encoder output, in an iterative way. Inteleaver structure should be well known at the reception. A random structure of inteleaver was tested against classic structure designs[4] (*shift interleaver, helical and odd-even*) and theirs performances measured.

Designed with the rand() function from C++ libraries the structure implements interleaver structure with 128, 256 and 1024 bits. The rand function in initialized with the current time value, to ensure a random seed. The generated array is then shuffled to ensure a proper randomization of the bits. For deinterleaving, we have to read this rule from a file every time. For the hardware implementation this procedure produces a delay time. The results were obtained with a structure with best results selected from a list of 10 structures that have been tested.

III. THE CAMAG SIMULATOR

We analyzed some interleaver structures on software platform CAMAG[3], dedicated to magnetic recording channel analysis (PR4, EPR4 and E2PR4 partial response channels). The in house built platform is dedicated to the read/write process, for a suite of coding and decoding procedures, R/W parameters and interfering noises. In our case, CAMAG was patched with MAP decoding of a stream of data input, randomly affected by controlled noise over a controlled length. The block diagram of the simulator is presented in Figure 3.

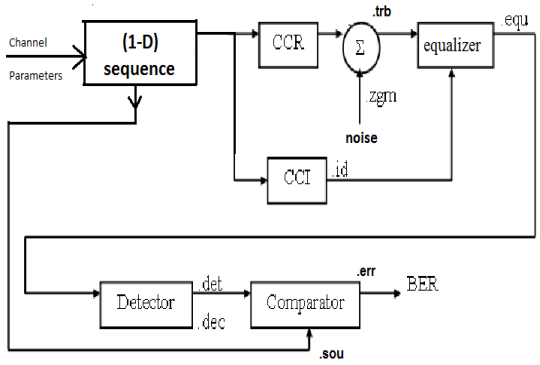


Figure 3: Block diagram of CAMAG

The first block performs a precoding of data to reduce propagation of errors that could occur in detection. The formula for precoding is: (1+D2) mod 2 for PR4 channel; (1-D+D2+D3) mod2 for EPR4 channel and (1-2D+2D3+D4) mod2 for E2PR4 channel. For read/write head we used 1-D operator to simulate the characteristic differential. The CAMAG approximate magnetic channel behavior using Lorentzian function [2]. The detection is Maximum A Posteriori, presuming that at each of some iterations set for the channel, an exchange of the extrinsic information is made between the two codes (the one with interleaved and the one without, both affected by noise). This way, after each iteration, the first decoder receives more information regarding the bit that is to be detected for each position, minimizing the chance of errors with each step taken [2]. BER (Bit Error Rate) values are calculated.

The Figure 4 provide a view of the MAP decoding that takes place in the simulation and how the interleaver affects it:

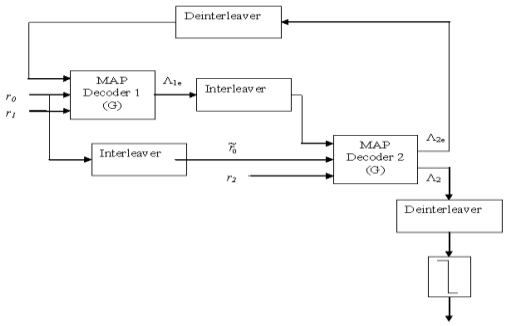


Figure 4: Turbo MAP decoding schema[4]

IV. SIMULATION REZULTS AND OPTIMAL INTERLEAVER

The interleaver permit turbo codes to act like a block code and express the interleaver length influence on code performance. A long length improves the performance but increases the complexity of the decoding algorithm. We studied the influence of fixed length interleaver with variable structures by measuring the BER at various SNR values and record densities. Turbo parameters used in simulations are: the encoders are described by 35O and 23O polynomials forward and backward generators, the size of memory was 5 for each component RSC encoders, and the decoding algorithm was MAP with 15 iterations for all tests. The noise was generated with different lengths and randomly distributed over data strings.

Below in Figure 5 are test results on a 2Kb data message, over which we applied first a set of three 1/2Kb bit interleavers, massage affected by a gaussian white additive noise (aleator1/2/3) over a length of 1/2Kb.

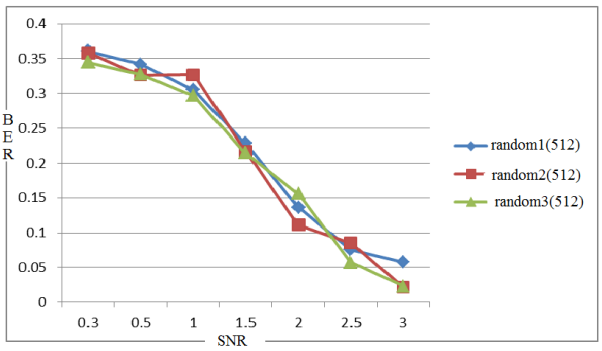


Figure 5: BER for 1/2Kb random interleavers

In the Figure 6, the 1/2Kb and 1Kb interleavers are compared and the slight advantage of the 1Kb interleaver is visible.

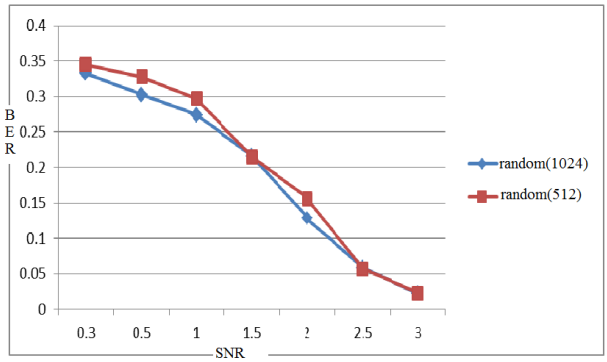


Figure 6: BER for 1/2Kb and 1Kb random interleavers

Examining random interleavers against 3 other well known types of interleavers, odd-even, helical and shift[x], it is visible that BER behaviour is practically no influenced over SNR simulated values .

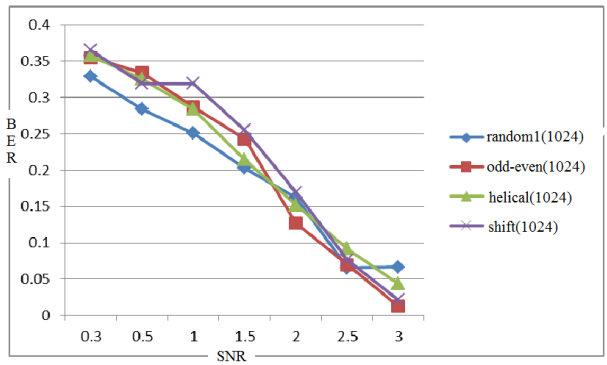


Figure 7: Comparative BER for interleaver structures

V. CONCLUSIONS

Over many possible scenario and parameters, we observed that at low SNR (<1dB), the 1Kb random interleaver design performed best, at the cost of greater computing resources , due to the fact that the structure of interleaver must be well known at the reception and we must save the structure every time.

Extending the interleaver length even further would only add to the complexity without bringing a considerable improvement to the performance of the turbo code.

The CAMAG simulation platform shows that the turbo codes used in digital magnetic recording systems express good performances, but with the price of hardware implementation complexity and redundancy. No increase in complexity is justifiable over 1/2 Kb interleaver length. This work and results shows that the Turbo Codes technique implemented in PR channels performs exceptional BER values, near Shannon limit, with no excess complexity in hardware.

REFERENCES

[1] C. Berrou, A. Glavieux, and P. Thitimajshima, “Near Shannon limit error- correctting coding and decoding: Turbo codes” *Proc. 1993 Int. Conf. Commun., pp. 1064-1070*.

[2] Ryan, W.E., "Performance of high rate turbo codes on a PR4-equalized magnetic recording channel," *Communications, 1998. ICC 98. Conference Record. 1998 IEEE International Conference on* , vol.2, no., pp.947,951 vol.2, 7-11 Jun 1998.

[3] Victor Soare, Ştefan Stăncescu, “Modular Partial Response Channel Software” *MeleCon 2000, May 2000, pp. 786-789.*

[4] A.S. Barbulescu, Iterative Decoding of Turbo Codes and Other Concatenated Codes, *Ph.d.thesis, University of South Australia, February 1996.*

[5] H Kobayashi, D.T.Tang, “Application of partial-response channel coding to magnetic recording systems”, IBM Journal of research and development, iulie 1970, pp. 368-375.

[6] Shan X. Wang, A.M.Taratorin: ”Magnetic Information Storage Technology”, *Academic Press, 1999, pp. 361-421*

Interpretation of pulmonary function test results in relation to asthma classification using integrated software suite

Almir Badnjević*, Dragan Koruga **, Mario Cifrek***, Hans. J. Smith, Tamer Bego
* New Technology d.o.o., Sarajevo, Bosnia and Herzegovina
** Army Medical Center Novi Sad, Petrovaradin, Serbia
*** University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
badnjevic.almir@gmail.com

Abstract - Asthma is a disorder characterized by chronic airway inflammation, airway hypersensitivity to a variety of stimuli, and airway obstruction that is at least partially or even completely reversible. To make correct diagnosis of asthma, diagnosis must be based on typical symptoms of asthma and identification of airway hyper responsiveness (AHR) or variable airway obstruction. In some cases, additionally we need to use bronchial provocation test using nonspecific stimuli, such as methacholine or histamine, what is useful for the determination of AHR. However, it is somewhat invasive in nature and it is not available in primary-care clinics or even in many general hospitals. Bronchodilator response (BDR) to short-acting β2-agonists is a valuable test to evaluate variable airway obstruction, which is only useful in patients with reduced lung function at the time of visit. Developed integrated software suite is user friendly tool which considers all this parameters to assist clinicians in the analysis and interpretation of pulmonary function tests data to better detect, diagnose, and treat asthma conditions. A total sum of 72 patient reports with previously diagnosed asthma by clinicians was tested with this tool. The software has performed the classification of asthma in the same way as doctors in 65 (90.27%) cases.

I. INTRODUCTION

Asthma is an inflammatory condition of the airways resulting in airway function becoming hyper-reactive, and generating increased mucus, mucosal swelling and airway smooth muscle contraction all of which contribute to airway obstruction. The symptoms include chest tightness, coughing and wheezing, and in severe cases shortness of breath and low blood oxygen [1]. In addition, the prevalence of asthma is increasing in recent decades, accompanied by the rise of the cost of asthma care as well [2-4]. One of many problems in asthma managements is that a lot of patients with asthma are not properly diagnosed or misdiagnosed as other respiratory diseases such as common cold, acute bronchitis, or chronic obstructive pulmonary disease [5,6].

Making a correct diagnosis of asthma is very important since that reduces costs of inappropriate treatment and leads to effective treatment of asthma.

The pulmonary function tests most commonly used to detect asthma is Impulse Oscillometry System (IOS) and spirometry. Spirometry as a test of lung function is most commonly used in the diagnosis of asthma and chronic obstructive pulmonary disease. Spirometry is is running by measuring forcing maneuver [7].

In contrast to forced spirometry, the forced oscillation technique (FOT) superimposes small air pressure perturbations on the natural breathing of a subject to measure lung mechanical parameters. IOS measures respiratory impedance using short pulses of air pressure [8-11].

Based on presented facts one can conclude that the most successful diagnosis is achieved by combination of IOS and spirometry. In this way we obtain a static assessment of the patient. In order to get dynamic assessment of pulmonary function patients, it is necessary that besides reports obtained from testing at diagnosis, from lung function testing to take in to account the patient's symptoms and allergies, perform auscultation of patient and do bronchial dilation (BDT) and bronchial provocation tests (BPT), if necessary.

The actual benefit of the integrated software suite is to help the clinical utilization of impulse oscillometry, and at the same time to assist in the diagnosis of asthma and differentiation of Chronic Obstructive Pulmonary Disease (COPD). Data which are first entered to software are related to a directory, symptoms and risk factors drawn from two major consensus for asthma and COPD, Global Initiative for Asthma (GINA) and Global Initiative for chronic Obstructive Lung Disease (GOLD). The parameters of IOS, spirometry, information about the symptoms, allergies and auscultation of the patient, are included in the neuro - fuzzy system, in order to help the software to suggest proper classification of asthma, COPD or normal lung condition. In the cases where it is not possible to determine the diagnosis the software will indicates to do BDT and/or BPT, after which new tests are required for IOS and spirometry, in order to get a complete diagnosis. This software increases the accuracy of the interpretation of data obtained by measuring the impulse oscillometry, and friendly guides the user to quickly obtain data to be used in lung function.

II. INTEGRATED SOFTWARE SUITE

Integrated software suite described in this chapter is developed as an integrated, interactive and user-friendly software. Software is using data from IOS and spirometry

measurements as inputs, and after processing all the related programs with information's such symptoms and allergies, classify the severities of small airway conditions in asthmatic and normal subjects to user, which actually represents the clinical assessment of lung function with software. This was achievable by utilizing the object oriented design methodologies, model based feature extraction and an inference system that incorporates the strengths of artificial neural networks and fuzzy decision making. The programming language to develop this software was Microsoft Visual C# Express. The software uses .NET technology to deploy applications, while in the background calls MATLAB functions for the main applications through MATLAB's .NET interface. The software has a feature of automatically reading and parsing standard format for IOS, Spirometry (SP) and Body Plethysmography (BP) recorded in the .pdf files.

Architecture for disease recognition is presented in Figure 1. After the software reads the data from .pdf files each of the reports ceptualise in fuzzy system and performed a preliminary classification of diseases on the basis of individual report.

Fuzzy system is designed according to the instructions and experience of experts in the field of medicine and equipment manufacturers for diagnostics of respiratory disease. Example of the implemented fuzzy reasoning for the case of IOS is shown in Figure 2.

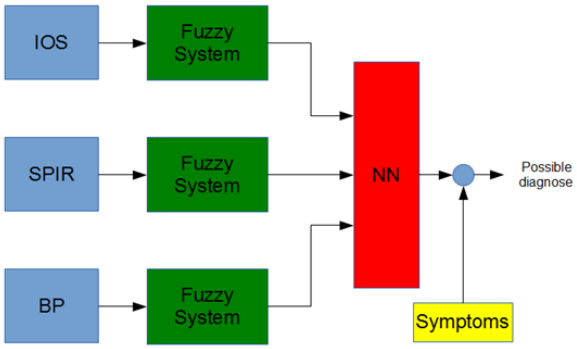


Figure 1. Architecture for disease detection

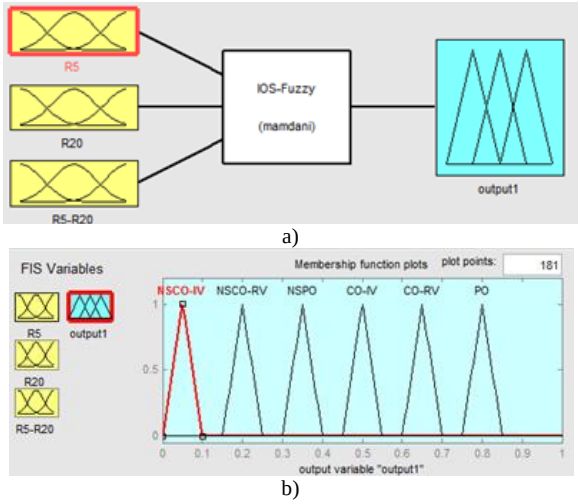


Figure 2. a) Implemented fuzzy reasoning for IOS, b) Output variables

The results of the analysis and classification based on the reports of IOS, spirometry and BP are the input vector of the neural network. Basic test does not take into account the report of BP, because it is not required in the initial test. In case of BDT or BPT, the result of the report is one of the inputs into the neural network for the next test. For BP test added is another input which is the result of BDT's (positive or negative), which is involved in the classification of disease after a period of BDT's.

The basic architecture of the neural network is shown in Figure 3. It is a linear feed forward network that is due to the application expert enough to properly perform the separation, and do not need complex architecture of network due to the presence of additional fuzzy classification system.

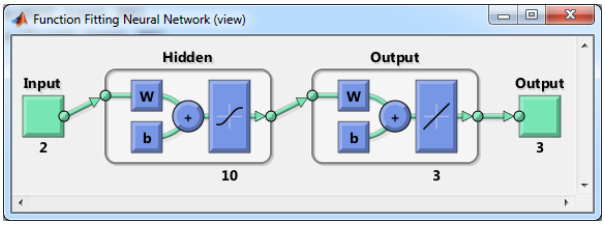


Figure 3. Basis architecture of implemented neural network

Complete neural network architectures in a system of classification of disease is shown in Figure 4.

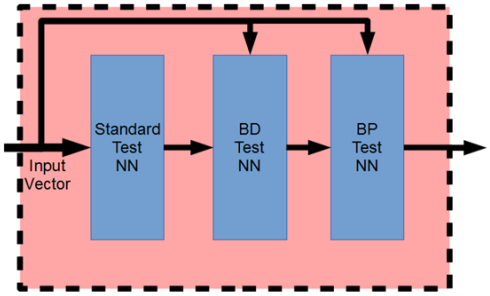


Figure 4. Complete architecture of implemented neural network

Schematic diagram of developed integrated software suite is presented on the Figure 5.

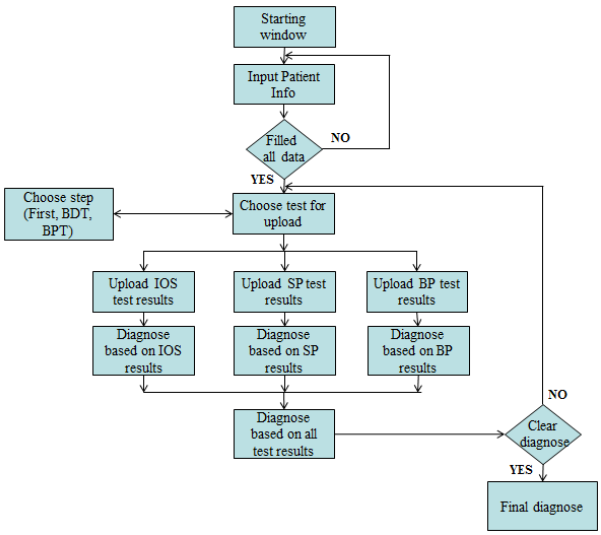


Figure 5. Schematic diagram of integrated software suite

Starting window of integrated software suite is presented on Figure 6.

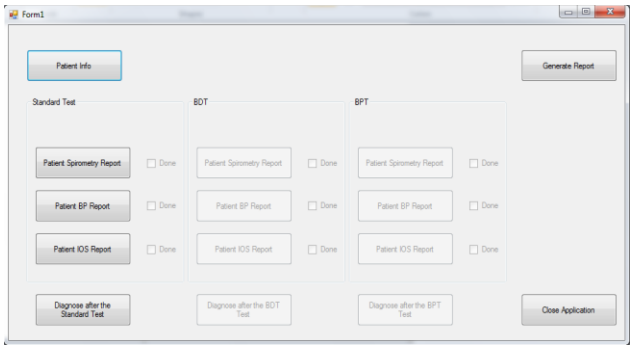


Figure 6. Starting window of integrated software suite

The software has been realized by the actual needs of physicians. Therefore, first you need to enter the Patient Info such as symptoms, allergies and perform auscultation of patient. Example of a Patient Info form is shown in Figure 7.

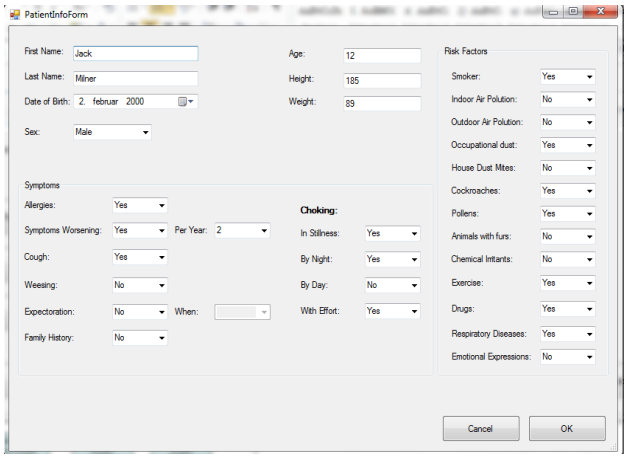


Figure 7. Exampe of Patient Info form

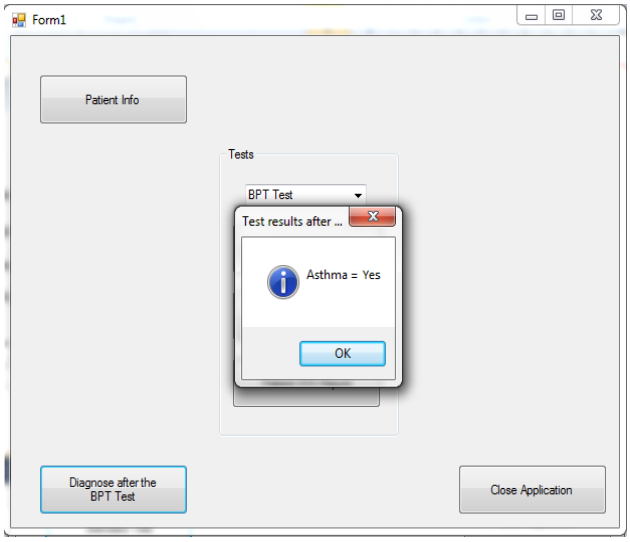


Figure 8. Classification of asthma by integrated software suite

Once you have entered information from the Patient Info form, it is possible to access the uploading measurement results of IOS and spirometry. If after the first measuring results software can't classify the disease, it suggests to approach BDT and/or BPT, and then we receive a final classification of diseases. Examples of classification of diseases asthma and healthy controls of fictitious patient data by integrated software suite are shown in Figures 8. and 9.

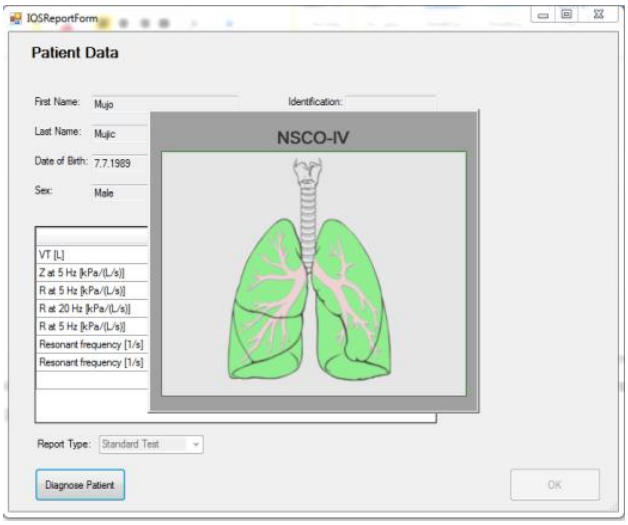


Figure 9. Classification of healthy controls by integrated software suite

III. MATERIALS AND METHODS

We enrolled 156 patients who previously visited departments for lung diseases in Clinical Centre University of Sarajevo, Bosnia and Herzegovina and Army Medical Center Novi Sad, Serbia. In this study, first all patients were asked by physician to respond to questions regarding symptoms, allergies and risk factors of asthma. Second step was measuring lung function with spirometry and IOS. Using Impulse oscillatory system (IOS), total respiratory resistance R5, proximal resistance R20, distal capacitive reactance X5 and resonant frequency Fres, were measured. Using spirometry, forced vital capacity (FVC) and forced expiratory volume in one second FEV1 were measured, while ratio FEV1/FVC was calculated. Patients with positive symptoms and R5>150%, R20>150%, X5>0.15, FEV1/FVC<0.8 were diagnosed with asthma. Patients with different test results were evaluated for bronchial hyper reactivity (BDR), first with BDT, and then with BPT. The bronchial dilation test was performed according to international guidelines using 400 µg of salbutamol. If R5 increases by more than 25% (ΔR5), X5 increases by more than 20% (ΔX5) and FEV1 decreases by more than 12% of baseline value and 200 mL, those patients are diagnosed with asthma. In cases where BDR didn't achieve any improvement for final diagnose, patients have done methacholine BPT. Serially increasing doses of methacholine were administered, and FEV1 was measured after each methacholine dose until a 20% decrease in FEV1. After that physicians were able to give final diagnosis. Patients were divided according to definite diagnosis of asthma into two groups, 'asthmatics' and 'healthy controls'. Of 156 patients who were enrolled in this study, 72 (46,15%) were grouped as asthmatics, while 84 (53,85%) were grouped as healthy controls. Integrated software suite was used for both groups of subjects.

IV. RESULTS AND DISCUSSION

Results of this study showed a high efficiency of integrated software suite in the classification of asthma disease (90.27%) and healthy patients (94.04%). In Table I are expressed percentage of hits and misses of the integrated software suite in the classification of asthmatics and healthy patients who were involved in this study.

TABLE I. EFFECTIVENESS OF THE INTEGRATED SOFTWARE SUITE IN THE CLASSIFICATION OF ASTHMA PATIENTS AND HEALTHY PATIENTS

Patients with asthma		Healthy patients	
72		84	
Hits by software	Misses by software	Hits by software	Misses by software
65 (90,27%)	7 (9,73%)	79 (94,04%)	5 (5,96%)

From table I we can see that in the case of patients affected with asthma the software showed a small percentage of false positive results (9.73%) as compared to physician diagnosis. In these 7 cases, the software classified those patients with other forms of respiratory diseases, not as asthma. In the case of healthy patients the software also showed a small percentage of false positive results (5,96%) as compared to physicians diagnosis. In these 5 cases, the software classified those patients as not healthy.

Testing the efficiency of integrated software suite at each single step of classification of asthma patients and healthy controls were obtained from the results of this study. These results are presented in Table II, while graphical representation is given on Figure 7.

TABLE II. HITS AND MISSES OF INTEGRATED SOFTWARE SUITE AT EACH SINGLE STEP OF CLASSIFICATION IN ASTHMATICS AND HEALTHY CONTROLS

Steps for classification of asthma	Criteria for classification of asthma	Patients with asthma (Hits by software) Σ = 65	Patients with asthma (Misses by software) Σ = 7	Healthy controls (Hits by software) Σ = 79	Healthy controls (Misses by software) Σ = 5
IOS and spirometry test results	R5 > 150% R20 > 150% X5 > 0.20 FEV1/FVC < 0.8	7 (10,77%)	0 (0%)	74 (93,67%)	1 (20,00%)
Bronchial dilation test results	R5 > 25% Fres > 20% FEV1 < 12	41 (63,08%)	1 (14,28%)	2 (2,53%)	3 (60,00%)
Bronchial provocation test results	R5 > 35% Fres > 30%	17 (26,15%)	6 (85,72)	3 (3,80%)	1 (20,00%)

R5 - Total respiratory resistance
R20 - Proximal resistance
X5 - Distal capacitive reactance
FEV1 - Forced expiratory volume in one second
Fres - Resonant frequency

Based on the study results shown in Table II and Figure 7. it's clear that the software classify the most patients with asthma in a second step, i.e. after bronchial dilation test (63,08%). In these cases of patients R5 increased by more than 25%, X5 increased by more than 20% and FEV1 decreased by more than 12%.

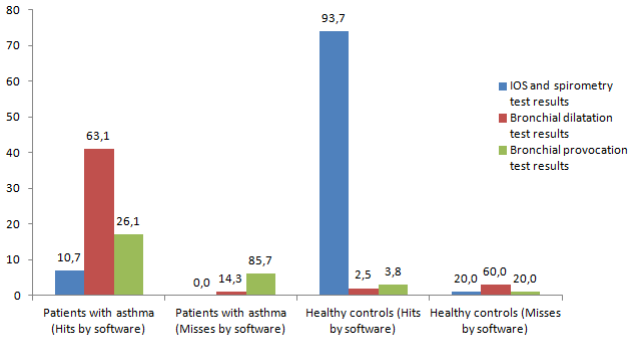


Figure 1. Graphical representation of integrated software suite in classification of asthmatics and healty patients (Results are expressed in percentages)

In the first step, i.e. after the results of IOS and spirometry, the software has classified 10,77% of patients with asthma, while in the third step, i.e. after bronchial provocative test, classified 26,15% of patients with asthma. These patients have done methacholine bronchial provocation test and R5 increased by more than 35% and Fres increased by more than 30%.

In the case of healthy controls, after the first step software classified 93.67% of patients as healthy controls. These patients had normal IOS and spirometry test results (also and symptoms). In the second and third step, where the test results of the patients contained minor deviations from normal predictive values, the percentage were 2.53% and 3.80%, respectively.

V. CONCLUSION

In this work we present results of classification of asthma using integrated software suite. Based on the total number of 156 patients involved in this study, the software correctly classified asthma in 90.27% of cases, whereas in the case of healthy subjects correctly performed the classification in 94.04% of cases. Also, it is shown that for classification of diseases integrated software suite takes into account the complete dynamic picture of patients. Was also presented that the clinical evaluation of lung function is a more comprehensive with developed software.

REFERENCES

[1] ALA. American Lung Association. [Online] 2011. http://www.lungusa.org/lungdisease/asthma/about-asthma/symptoms-diagnosis.html.

[2] Singh M. The burden of asthma in children: an Asian perspective. Paediatr Respir Rev 2005; 6: 14-9.

[3] Sennhauser FH, Braun-Fahrlander C, Wildhaber JH. The burden of asthma in children: a european perspective. Paediatr Respir Rev 2005; 6: 2-7.

[4] Beasley R. The burden of asthma with specific reference to the united states. J Allergy Clin Immunol 2002; 109 (Supple 5): 482-9.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] Helms PJ. Issues and unmet needs in pediatric asthma. Pediatr Pulmonol 2000; 30: 159-65.

[6] Enright PL, McClelland RL, Newman AB, Gottlieb DJ, Lebowitz MD. Underdiagnosis and undertreatment of asthma in the elderly. Cardiovascular health study research group. Chest 1999; 116: 603-13.

[7] Nielsen, KG and Bisgaard, H. “The Effect of Inhaled Budesonide on Symptoms, Lung Function, and Cold Air and Methacholine Responsiveness in 2- to 5-year-old Asthmatic Children”. Am J Respiratory Critical Care Medicine 2000, Vol. 162, pp. 1500-1506.

[8] Song, T W, et al. “Utlility of impulse oscillometry in young children with asthma”. Pediatr Allergy Immunol. 2008, Vol. 19, pp. 763-768.

[9] Song, T W, et al. “Correlation between spirometry and impulse oscillometry in children with asthma”. Acta Paediatrica, vol. , pp.,. 2008, Vol. 97, pp. 51-54.

[10] Hur, H Y, et al. “A comparison between impulse oscillometry system and spirometry for detecting airway obstruction in children”. Korean Journal of Pediatrics. 2008, Vol. 51 , 8, pp. 842-847.

[11] Larsen, G L, et al. “Impulse oscillometry versus spirometry in a long-term study of controller therapy for pediatric asthma. J Allergy Clin Immunol. 2009, Vol. 123, 4, pp. 861-867.

DC-VIS

International Conference on

DISTRIBUTED COMPUTING AND VISUALIZATION

Steering Committee

Chairs: Karolj Skala, Ruđer Bošković Institute, Zagreb, Croatia
Roman Trobec, Jožef Stefan Institute, Ljubljana, Slovenia
Uroš Stanič, Biomedical Research and Innovative Society, Ljubljana, Slovenia

Members: Piotr Bala, N. Copernicus University, Toruń, Poland
Leo Budin, University of Zagreb, Croatia
Yike Guo, Imperial College London, UK
Gordan Gulan, Klinika za ortopediju, Lovran, Croatia
Ladislav Hluchý, Slovak Academy of Sciences, Bratislava, Slovakia
Peter Kacsuk, MTA SZTAKI, Budapest, Hungary
Aneta Karaivanova, Bulgarian Academy of Sciences, Sofia, Bulgaria
Charles Loomis, Paris-Sud University, Orsay Cedex, France
Luděk Matyska, Masaryk University, Brno, Czech Republic
Laszlo Szirmay-Kalos, Technical University of Budapest, Hungary
Tibor Vámos, Hungarian Academy of Sciences, Budapest, Hungary

PAPERS

DISTRIBUTED COMPUTING

High Productivity Processing – Engaging in Big Data around Distributed Computing

Morris Riedel*, M. Memon, A. Memon, G. Fiameni **, C. Cacciari **, and Thomas Lippert*
* Juelich Supercomputing Centre, Juelich, Germany
** CINECA, Bologna, Italy
m.riedel@fz-juelich.de

Abstract - The steadily increasing amounts of scientific data and the analysis of 'big data' is a fundamental characteristic in the context of computational simulations that are based on numerical methods or known physical laws. This represents both an opportunity and challenge on different levels for traditional distributed computing approaches, architectures, and infrastructures. On the lowest level data-intensive computing is a challenge since CPU speed has surpassed IO capabilities of HPC resources and on the higher levels complex cross-disciplinary data sharing is envisioned via data infrastructures in order to engage in the fragmented answers to societal challenges. This paper highlights how these levels share the demand for 'high productivity processing' of 'big data' including the sharing and analysis of 'large-scale science data-sets'. The paper will describe approaches such as the high-level European data infrastructure EUDAT as well as low-level requirements arising from HPC simulations used in distributed computing. The paper aims to address the fact that big data analysis methods such as computational steering and visualization, map-reduce, R, and others are around, but a lot of research and evaluations still need to be done to achieve scientific insights with them in the context of traditional distributed computing infrastructures.

I. INTRODUCTION

More and more scientific user communities share the demand for new solutions for the steadily increasing amount of scientific data that arise from scientific instruments (i.e. higher resolutions), an increased amount of sensors in scientific field studies (i.e. constantly created real-time datasets), or better computing power (i.e. more granular and realistic simulations). As a consequence, the term *'big data'* emerged as a kind of new research field that aim to address the aforementioned challenges and to provide reasonable solutions in scientific ecosystems (i.e. research labs, universities, shared scientific instrument collaborations, etc.).

The term *'big data'* is very vaguely defined and directly leads to a wide variety of data-driven approaches in commercial environments and industry rather than to specific methods and scientific approaches. For instance, an often quoted definition is as follows: *'big data is data that becomes large enough that it cannot be processed using conventional methods'* [1]. The question arises what are these *'conventional methods'* in the particular context of the distributed computing community?

This paper aims to survey some of these conventional methods that have been used for decades in scientific computing in general and distributed computing infrastructures (DCIs) in particular with large scientific datasets. We argue that the aforementioned definition is wrong and do believe that several of the scientific methods and approaches have their role to play in the more and more emerging *'big data wave'*, especially around the requirement of *'big data processing and analysis'* what we refer to as *'high productivity processing'* in this contribution. This refers to known *'conventional methods'* such as High Performance Computing (HPC) and High Throughput Computing (HTC) aligned with large-scale federated infrastructures including a wide variety of storages and databases as well as available data analytics techniques (e.g. computational steering and visualization).

The reasons for the aforementioned statement will be underpinned by existing solutions and by comparing the scientific ecosystem with the current industry realm that tends to state that *'big data'* solutions are available not only for commercial applications, but also for scientific user communities. We thus argue in this paper that the scientific analysis processes and its methods are different from analysing large commercial datasets that often include the analysis of log files or connection details in order to understand customer behaviour or to optimize hardware for better utilization.

While we do agree that several methods in industry are interesting for scientific investigation (e.g. the map-reduce method [2]) there is still a lot of potential to apply the conventional methods, perhaps in a tuned and improved form, to the challenges of the emerging *'big data waves'* in science and engineering. Some scientific *'big data'* applications that we work with are given as examples for the processing with conventional methods also in the *'big data'* field.

The remainder of this paper is structured as follows. Section I introduces the problem space while Section II outlines the meaning and challenges of scientific data and *'big data waves'* in our DCI community. Section III then presents the different rather conventional *'big data processing'* methods and approaches including several scientific application domains in context of *'big data'* applications. Section VI surveys related work in the field and this paper ends with some concluding remarks.

II. SCIENTIFIC DATA AS INFRASTRUCTURE

The DCI community with many of their institutional and scientific community members already analyse large-scale scientific data in the context of data-intensive computing since decades. The view in this community was traditionally very much oriented towards computing (cf. to the DCI term) instead of putting the data in the center of the view as stated by the high-level expert group document on scientific data: *'Our vision is a scientific e-infrastructure that supports seamlesss access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure'* [4]. As a consequence, also other more DCI-oriented high-level bodies have written or address more data related topics such as e-Infrastructure Reflection Group (e-IRG) report on data management [5] and the European e-Infrastructure Forum (EEF) report of data requirements from European user communities [6].

As the previous sources suggest there are a wide variety of requirements and demands for data infrastructures and there are many more high-level reports, press releases, or rather political statements that all share in common that the technical realisation of them are often rarely described. In order to derive concrete architectures for data infrastructures that are able to satisfy the high-level requirements, we have started in realizing a reference model-based approach on how such infrastructures look like in Riedel et al. [3], illustrated in Figure 1. The aim of the reference model is not only to increase the understanding of different stakeholders from various scientific domains with desribed entities and their relationships, but also to provide *'a point of reference to compare required functionality with the functionality already existing in established DCIs or scientific discipline specific data infrastructures'*.

Before the overall paper and later sections will go into more details about these comparisons and overlaps in desired and existing functionality, this section briefly outlines the core building blocks of the reference model described in more detail in Riedel et al. [3]. This includes some hints on how we realize one out of many possible derived concrete reference architectures within the European Data Infrastructure (EUDAT) [7].

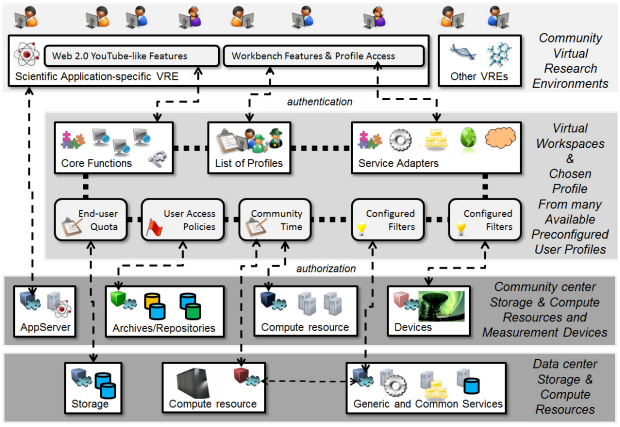


Figure 1. Conceptual view of the data infrastructure reference model including computing resources for 'big data' processing [3].

The foundation of the reference model are entities in the two layers that form the *'collaborative data infrastructure (CDI)'* among multi-disciplinary large data centers with massive computational power and large storage systems as well as scientific discipline-specific community centers with data archives and moderate computational power. The aim is to federate the available resources among these centers and to provide a stable and enhanced access to unique scientific devices and large data archives as shown in Figure 1 for the scientific end-users. As the illustrations reveal, the reference model also models the required resources for performing *'high productivity processing'* on scientific data with either HTC or HPC resources possibly available in both the community centers and data centers. The required methods that are realized as a wide variety of *'common and generic services'* are illustrated in Figure 1 as well.

In this context, one could argue that the architectural design of the *'European Grid Infrastructure (EGI)'* [8] is similar while we argue that the concrete roles of the EGI resource providers are rather generally defined as equal while in the CDI the roles are more clear and distinguished from each other. The realization of the EUDAT CDI however points to several overlaps in architectural approaches, technology requirements, and required analysis methods. There are also several overlaps in the architectural design with the *'Partnership for Advanced Computing in Europe (PRACE)'* [9] especially when we consider the large-scale HPC resources and their ecosystems (e.g. parallel file systems, etc.) available at the large data centers. When we consider the data archives and repositories, there are also architectural and functionality overlaps with a wide variety of existing scientific discipline-specific data infrastructures such as the *'Common Language Resources and Technology Infrastructure (CLARIN)'* [10] or the *'Digital Research Infrastructure for the Arts and Humanities (DARIAH)'* [11] just to list a few.

In order to realize the *'data as an infrastructure'* without *'reinventing the wheel'* and with building on established technologies and services emerging from existing aforementioned infrastructures, we outlined the vision of a ScienceTube [3] that federates all the added values from the aforementioned existing infrastructures with an increased interlinking. The goal is that users are able to 'dive into the data' instead of just using data having at the same time a powerful *'virtual workspace'* that they can use for scientific data analysis. As illustrated in Figure 1, these virtual workspaces need to have *'service adapters'* to the wide variety of established services available at the community and data centers. The view on data itself, metadata as well as additional information such as clear quotas of storage, lists of community times on rare computing resources are core functions of the virtual workspaces. The *'list of profiles'* enable the view on data with different roles and lightweight Web 2.0 features in a wide variety of scientific application-specific potentially browser-based *'Virtual Research Environments (VREs)'*.

The EUDAT project is currently realizing parts of this vision and the first service *'Safe Data Replication'* [3] is ready for production being also the basis for 'high productivity processing' on large scientific *'big data sets'*.

III. BIG DATA PROCESSING METHODS & APPROACHES

The previous section outlines the general reference model approach of a CDI while this section provides details on the different types of available conventional analysis methods in the context of processing large scientific datasets. The fundamental goal here is to change the traditional view on *'complicated technology (i.e. one specific middleware)'* towards *'general methods and approaches'* that can be, in principle, realized with several technologies (i.e. multiple middlewares) today. In context of the reference model illustrated in Figure 1, this section aims to provide a technology-agnostic specialization on *'common and generic services'* that are required on both the community and data centers as *'big data processing and analytics services'* illustrated in Figure 2.

A. (Iterative) Map-Reduce

The *'Map-Reduce'* method [2] introduces the idea of processing (scientific) data in parallel instead of serial and represents a current trend in computer science, especially in those commercial enterprises that analyze 'big data' of consumer behavior (e.g. phone connections). Applications include information retrieval, inverted indexing, graph algorithms, and PageRank just to list a few. The method is simple to program with existing Java-based frameworks like Apache Hadoop [11,12] by just specifying a MAP and REDUCE function while the framework sorts results of the MAP execution (by using a hash function) before it is given to the REDUCE execution. The key idea is that the MAP and REDUCE functions can be executed in parallel using HTC resources to perform the specified data analysis. That is one of the reasons why this method can be applied in community and data centers as shown in Figure 2. In some cases it makes sense to use MPI implementations of Map-Reduce [34] on HPC resources.

Combining research in the field of scientific computing with the *'Map-Reduce method'* has been performed, among others, by Geoffrey Fox et al. in [13]. The result is a new method named as *'Iterative Map-Reduce'* that enhances the *'classic Map-Reduce'* algorithm with multiple ITERATIONS of the MAP-REDUCE steps. This enables more applications such as parallel linear algebra-based solutions and can be easily used by existing frameworks such as Twister [14].

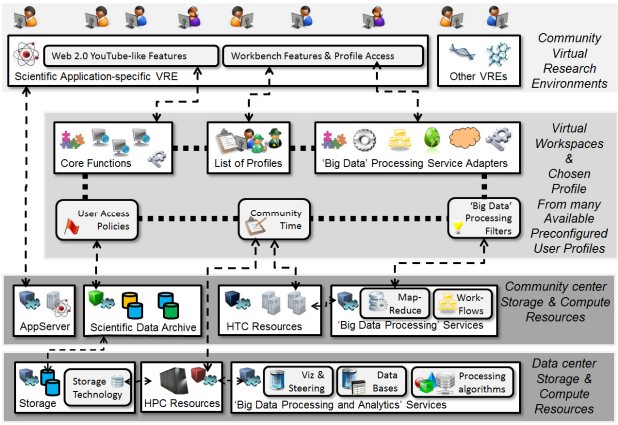


Figure 2. Conventional 'big data' processing methods and approaches in context of a service-based federated collaborative data infrastructure.

B. Scientific Workflows

The methods of *'scientific workflows'* represent an approach of chaining different tasks for data processing or the preparation of it (e.g. transformation of data formats into one another). This method need to specify two important *'information flows'* including the *'flow of task execution'* (i.e. on which computational resource will be each individual task be executed) and the *'flow of data'* (i.e. which input and output data matters for the specific task). Existing tools put an emphasis on one or another.

To provide examples of this rather conventional data analysis method, the UNICORE workflow implementation [15] focusses on the computational aspects as the UNICORE middleware is primarily used in HPC-driven infrastructures (e.g. PRACE), while more recently, it is also used more and more in HTC-driven infrastructures (e.g. EGI). It allows for the specification of input and output data while in general this workflow tool implementation supports *'big data'* processing especially in the context of parallel applications using the *'Message Passing Interface (MPI)'* [16]. Within fusion science, we also worked with other workflow engines that put the emphasis more on defining the *'data flow'* such as Kepler as described in Memon and Riedel et al. [17]. Yet another example of a scientific domain-specific workflow we worked with is the WebLicht workflow tool implementation [18] used in the CLARIN infrastructure specifically optimized for analyzing data in the computational linguistic fields. While there are many other workflow implementations, the key message to take away is that this rather conventional data analysis method will still be important in the field of *'big data processing'*.

C. Databases

The use of the *'databases method'* in the context of *'big data processing'* often refers to image-based approaches. Digital (scientific) images (e.g. from the bio-medical domain) represent tremendous amounts of *'big data'* for which a *'high productivity processing'* is required to analyze and intelligently index datasets. Approaches using this method often include a typical two phase working paradigm. Firstly, scientific end-users of infrastructures (and their databases) register images and their metadata. Secondly, scientific end-users search for the right datasets with images (i.e. involving metadata) and perform queries/analysis on image contents (e.g. image similarity searches).

The rather conventional method stores metadata in SQL databases while the large 'big data' images are stored in file systems to which a reference is stored alongside the metadata. A query about image metadata is thus performed with the SQL language while the analysis of image contents often triggers automatic processing techniques (e.g. search for most similar images to the input image). Especially the latter processing-intensive part can be perfectly executed on HTC resources provided in infrastructures such as EGI. This was shown by I. Montagnat in [19] taking advantage of the wide variety of processing resources in DCIs just to provide one example of many in the context of medical image data analysis, but we argue that this method is still needed for *'big data analysis'* in the decades to come.

D. *Massively-Parallel Processing Algorithms*

Methods of this category include many traditional computing techniques and simulations that often take advantages of large-scale HPC resources. MPI-based mechanisms used in parallel computing or even OpenMP techniques [21] are often used in scientific applications. These applications are *'conventional'* computational simulations according to known physical laws or numerical methods providing solutions for approximations of the reality requiring a large processing power. More recently input data to these scientific applications is steadily growing in sizes while at the same time the output data of these simulations grow as well, because (a) simulations have more granularities of the physical models and (b) more cores are used for processing these models on modern large-scale HPC resources such as the BlueGene/Q series just to provide one example.

In this context it is important that this paper argues that *'big data processing'* cannot be done only with the *'Map-Reduce method'*, because very communication-intensive scientific applications (a) still require massively-parallel tasks with their supporting underlying low-latency networks (e.g. Infiniband) and (b) will take advantage of the wide variety of scientifically-proven physical models and libraries created over the last decade established by peer-review procedures by the HPC and scientific communities that cannot be easily transformed into MAP-REDUCE steps. To provide an example, G. Fox et al. describes in [13] why scientific applications solving partial differential equations (PDEs) and particle dynamics (e.g. n-body simulations) are not suited even for the *'Iterative Map-Reduce'* method. We further expect more and more data analysis libraries to be important for *'big data processing'* (e.g. R MPI package [22]).

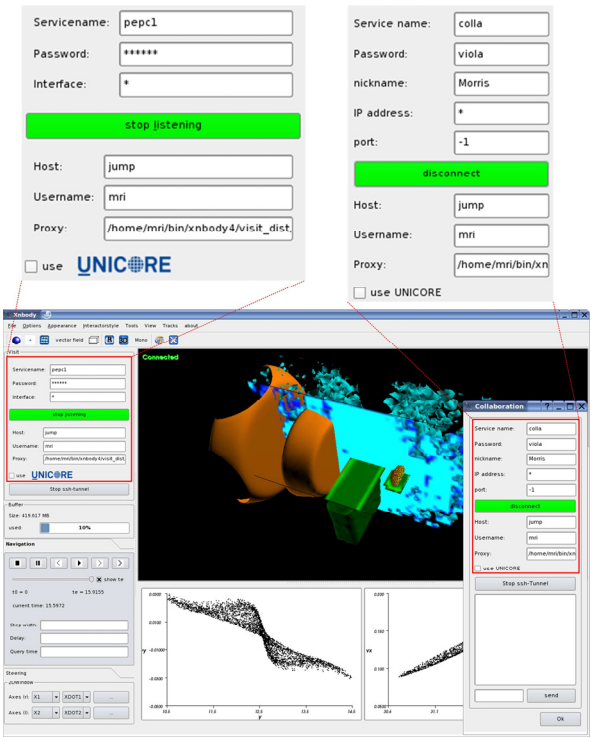


Figure 3. Scientific investigation using computational steering and visualization techniques is a valid 'big data analytics' method [20].

E. *Visualization and Steering*

The method of *'visualization of scientific data'* is a conventional data analysis method is established and intensively used since many decades. We refer here to the very processing-intensive simulations in the context of HPC simulations where each step is computed according to physical laws and visualized on the fly. To provide an example to obtain insights from computational simulations, so-called 'streamlines' and 'integral curves' are one of the important techniques that use vector fields and represent the cornerstone of visualization and data analysis across a variety of application domains as described by D. Camp et al. in [23]. This also includes the visualization techniques required for large *'big data'* currently emerging on large-scale HPC resources in scalable computational simulations. This method often implies an interactive channel from the simulation to the client desktop in order to show step-wise progress on computational simulation runs on the fly.

Adding another scientific data analysis method to the *'visualization method'* is called the *'computational steering'* method, which refers to the change of HPC simulations and their application parameters during the run-time on large HPC resources. This data analysis tool has the advantage to flexible steer the focus of a HPC application run to regions of interests dynamically understood via the *'visualization'* method. Although the *'computational steering'* method could be applied standalone (e.g. air shaft open/close within a computational fluid dynamics simulation), we argue that it is often useful to have it applied in context of visualizations. In the context of DCIs in general and the *'Distributed European Infrastructure for Supercomputing Applications (DEISA)'* [24] resources in particular, we worked with this method in various scientific application domains but especially with those having a step-wise fashion to perform simulations on large HPC resources. Hence, the step-wise approach not only enables the visualization progress as mentioned above, but also the step-wise change of parameters (aka steering) of the simulation behavior. This method thus often implies a bi-directional interactive channel to transport current visualization progress to the client desktop while on the other hand transporting steering parameters from the client desktop back to the *'steered scientific simulation'*.

To provide an example for this method, we used an implementation named as the *'Collaborative Online Visualization and Steering (COVS)'* [20] tool with n-body simulations in astrophysics and plasma-physics while the latter is illustrated in Figure 3 alongside some steering parameters and necessary connection options. As the options reveal, the COVS tool is realized with the middleware UNICORE that is available on a wide variety of DCIs today. Given its low-intrusive design where the *'client act as a server'* [20], the COVS tool is able to analyze scientific applications that make use of *'big data'*. This paper thus argues alongside others from the visualization domain (e.g. VisIt team [25]) that the methods of *'scientific visualization'* and related methods will surely play a role in the processing and analysis of scientific *'big data'* (e.g. in extreme scale simulations).

IV. Related Work

There is a wide variety of related work in the context of scientific data infrastructures and possible architectural approaches that partly also address processing methods. One of the largest distributed data infrastructure was established by the High Energy Physics (HEP) community creating the *'World-wide Large Hadron Collider Computing Grid (WLCG)'* [26] . In contrast to our approach, the WLCG data infrastructure is not multi-disciplinary and particularly focussed on HEP physics data sharing and analysis using specific algorithms, methods, and services from the HEP community (e.g. GAUDI [33]).

The EGI infrastructure [8] is more multi-discplinary oriented whereby the aforementioned WLCG represents only one user community, but EGI is rather HTC oriented thus limiting the methods to gain scientific insights that can only be done with large-scale HPC resources. In addition, EGI is moving towards a dynamic cloud-based model while the approach of EUDAT and its methods rather focus on the traditional computing paradigms such as HTC and HPC. In addition, storage space is rather specifically negotiated in EUDAT instead of freely usable by registered *'virtual organizations (VOs)'* [27]. Related work also includes PRACE [9], but in contrast to our approaches in EUDAT, PRACE is much more focussed on peta-scale HPC simulations instead of *'big data'* storages and the provisioning of *'high productivity processing methods'* as listed in Section III.

There are also a wide variety of related existing data infrastructures in the field that we can not list all due to the page restriction of this paper. One example of these data infrastructures is the *'PaNdata infrastructure for neutron and proton sources'* [28] that create with an increased set of sensor technologies more and more *'big data sets'*. In comparison to our approach, we can state that the approach is much more standards-based, but being fundamentally oriented towards one user community. Another example of a community-specific data infrastructure is the *'European Clinical Research Infrastructure Network (ECRIN)'* [29]. Several requirements of this infrastructure are similar to those in EUDAT (e.g. long-term preservation) including the challenge of archiving clinical trial data with raw data of (often anonymized) patients.

Another type of an example is the *'Integrated Carbon Observation System (ICOS)'* [30] that is specifically optimized to work with ~60 atmospheric and ecosystem measurement sites as well as ~10 oceanic sites. Hence, this related data infrastructure raise the challenge for real-time analytics of scientific measurement data that we did not mention in our survey of methods in Section III due to our limit in experience with it. Nevertheless, the work with this data in EUDAT is a reasonable future work idea in particular when processing real-time data that can be at least partly realized with the aforementioned *'Map-Reduce'* method'. Also the *'computational steering'* method might be a reasonable future work item in this context to explore *'regions of interest'* within the real-time data-streams in order to prevent the storage of potentially uninteresting measurement data originating from the multiple sensors.

In terms of access and processing of large-scale data collections (e.g. repositories, gene ontologies, proteomics data, archives of protein sequences, etc.), the European Life Sciences Infrastructure for Biological Information (ELIXIR) [31] is another interesting example of a data-oriented infrastructure. Envisioned solutions for a federated *'Authentication and Authorization Infrastructure (AAI)'* are similar to those of EUDAT and should be kept in-sync in order to enable data sharing and large-scale data processing on DCI resources using several of the methods described in Section III. In this context, also the data infrastructure for chemical safety diXA is an interesting related work since it offers a *'data pipeline'* solution [32] that could be of interest to the available solutions in the context of our surveyed *'workflow'* methods.

V. Conclusion

As mentioned in the beginning one of the often referred defintions about *'big data'* states that it cannot be processed using so-called *'conventional methods'*. We surveyed several known conventional methods (i.e. iterative Map-Reduce, scientific workflows, databases, massively-parallel processing algorithms, and visualization and steering) in the context of DCIs in general and our scientific application work experience in particular. We conclude that these conventional methods are able to process *'scientific big data'* as many of the data challenges in the past is the current work practice of today (e.g. visualization and the growing amount of CPUs/cores was a trend over decades). We thus further argue that the commercial approach of Map-Reduce for 'big data' is equally interesting for the data analysis in science and engineering, but representing only on of several methods that will play a role in the *'high productivity processing'* of large-scale scientific *'big data'* or *'data-intensive computing'*.

Another conclusion we derive from working with scientific datasets (e.g. big brain slices of 40 GB each summarizing to a whole brain with ~700 TB) is that the well-known new *'big data idea'* of *'bring the computation to the data'* instead of *'bringing the data to computation'* is at least partly wrong in our scientific DCI ecosystems. This can be explained by the fact that the large community and data centers (cf. Figure 2) have the numerous experts that work with the *'conventional methods'* over decades and are thus used to have scalable solutions for growing and new data volumes (e.g. portable IO formats, or scalable I/O libraries, etc.) or at least are working on them. *'Bringing the computation to the data'* might help in saving bandwidth for a while, but the expertise at the different centers is much more valuable that outperform the safe of bandwidth in a few data analysis cases by an order of magnitude.

Finally, we conclude that the shift of view from a traditional *'computing oriented'* focus in DCIs towards a more *'data oriented'* focus provides a more complete view on necessary services and approaches in the DCI community. This includes the provisioning of new solutions (e.g. persistent identifier service for scientific data in EUDAT) or the investigation of new scientific inquery methods (e.g. cross-disciplinary data sharing).

ACKNOWLEDGMENT

This work is partially funded by the EUDAT project that is co-funded by the EC 7th Framework Programme (Grant Agreement no: 283304).

REFERENCES

[1] O'Reilly Radar Team, *"Big Data Now: Current Perspectives from O'Reilly Radar"*, O'Reilly Media, 2011.

[2] J. Dean and S. Ghemawat, *"MapReduce: Simplified Data Processing on Large Clusters"*, in proceedings of the Sixth Symposium on Operating System Design and Implementation (OSDI), San Francisco, USA, 2004.

[3] M. Riedel, P. Wittenburg, J. Reetz, M. van de Sanden, J. Rybicki, B. von St. Vieth, G. Fiameni, G. Mariani, A. Michelini, C. Cacciari, W. Elbers, D. Broeder, R. Verkerk, E. Erastova, M. Lautenschlaeger, R. Budig, H. Thielmann, P. Coveney, S. Zasada, A. Haidar, O. Buechner, C. Manzano, S. Memon, S. Memon, H. Helin, J. Suhonen, D. Lecarpentier, K. Koski and Th. Lippert *"A Data Infrastructure Reference Model with Applications: Towards Realization of a ScienceTube Vision with a Data Replication Service",* Journal of Internet Services and Applications, Volume 4, Issue 1

[4] Wood J, et al.,*"Riding the wave - How Europe can gain from the rising tide of scientific data"*, European Union, Italy, 2010.

[5] e-IRG Data Management Task Force, *"e-IRG Report on Data Management"*, e-IRG, 2009

[6] B. Jones, D. Davies, H. Lederer, P. Aerts, Th. Eickermann, S. Newhouse, *"ESFRI Project Requirements for pan-European e-Infrastructure resources and facilities"*, European e-Infrastructure Forum, Amsterdam, Netherlands, 2010

[7] European Data Infrastructure (EUDAT) Web page, http://eudat.eu

[8] European Grid Infrastructure (EGI) Web page, http://www.egi.eu

[9] Partnership for Advanced Computing in Europe (PRACE) Web page, http://www-prace-ri.eu

[10] Common Language Resources and Technology Infrastructure (CLARIN) Web page, http://www.clarin.eu

[11] M. Bhandarkar, *"MapReduce programming with Apache Hadoop - Tutorial"*, International Parallel and Distributed Processing Symposium (IPDPS), 2010

[12] T. White, *"Hadoop: The Definitive Guide – MapReduce for the Cloud"*, O'Reilly Media, ISBN 978-0-596-52197-4, 2009

[13] G. Fox et al., *"Using HPC Systems MapReduce & Eucalyptus on FutureGrid"*, PPAM 2011, Torun, Poland, 2011

[14] Ekanayake, Jaliya, Hui Li, Bingjing Zhang, Thilina Gunarathne, Seung-Hee Bae, Judy Qiu, and Geoffrey Fox. *"Twister: a runtime for iterative mapreduce"*, in *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, pp. 810-818. ACM, 2010.

[15] A. Streit, P. Bala, A. Beck-Ratzka, K. Benedyczak, S. Bergmann, R. Breu, J. M. Daivandy, B. Demuth, A. Eifer, A. Giesler, B. Hagemeier, S. Holl, V. Huber, N. Lamla, D. Mallmann, A. S. Memon, M. S. Memon, M. Rambadt, M. Riedel, M. Romberg, B. Schuller, T. Schlauch, A. Schreiber, T. Soddemann, W. Ziegler *"UNICORE 6 - Recent and Future Advancements",* Annals of Telecommunications, Springer, Volume 65, Issue 11-12, pp 757-762, 2010

[16] P. Pacheco, *"Parallel Programming with MPI"*, Morgan Kaufmann, ISBN 1558603395, 1996.

[17] M. S. Memon, M. Riedel, A. S. Memon, F. Wolf, A. Streit, Th. Lippert, M. Plociennik, M. Owsiak, D. Tskhakaya, Ch. Konz *"Lessons Learned From Jointly Using HTC- and HPC-driven e-Science Infrastructures in Fusion Science",* 2010 International Conference on Information and Emerging Technologies (ICIET), Karachi, Pakistan, ISBN 978-1-4244-8001-2, pp. 1-6, 2010

[18] E. Hinrichs, M. Hinrichs, and T. Zastrow, *„Weblicht: web-based LRT services for German",* In Proceedings of the 10th ACL System Demonstration, Association for Computational Linguistics, Stroudsburg, USA, 2010.

[19] I. Montagnat, V. Breton, I.E. Magnin, *"Partitioning medical image databases for content-based queries on a grid"*, in methods of Information in Medicine, 44 (2005), pp. 154–160

[20] M. Riedel, Th. Eickermann, W. Frings , S. Dominiczak, D. Mallmann, Th. Düssel, A. Streit, P. Gibbon, F. Wolf, W. Schiffmann, Th. Lippert, **"***Design and Evaluation of a Collaborative Online Visualization and Steering Framework Implementation for Computational Grids"*, in Proceedings of the 8th IEEE/ACM International Conference on Grid Computing (Grid 2007), Austin, Texas, ISBN 1-4244-1560-8, pp.169 - 177, 2007

[21] R. Chandra, R. Menon, L. Dagum, D. Kohr, D. Maydan, and J. McDonald. Parallel Programming in OpenMP, Morgan Kaufmann, ISBN 1-55860-671-8, 2001.

[22] R MPI package Web page, http://www.stats.uwo.ca/faculty/yu/Rmpi/

[23] D. Camp, C. Garth, H. Childs, D. Pugmire, K. I. Joy,*"Streamline Integration Using MPI-Hybrid Parallelism on a Large Multicore Architecture",*IEEE Transactions on Visualization and Computer Graphics,November 1, 2011,

[24] W. Gentzsch, D. Girou, A. Kennedy, H. Lederer, J. Reetz, M. Riedel, A. Schott, A. Vanni, M. Vazquez, J. Wolfrat *"DEISA - Distributed European Infrastructure for Supercomputing Applications",* Journal of Grid Computing, Vol. 9 (2011) 2, 2011, pp. 259 – 277

[25] Hank Childs, Eric Brugger, Brad Whitlock, Jeremy Meredith, Sean Ahern, David Pugmire, Kathleen Biagas, Mark Miller, Cyrus Harrison, Gunther H. Weber, Hari Krishnan, Thomas Fogal, Allen Sanderson, Christoph Garth, E. Wes Bethel, David Camp, Oliver Rubel, Marc Durant, Jean M. Favre, Paul Navratil*,"VisIt: An End-User Tool For Visualizing and Analyzing Very Large Data"*, High Performance Visualization---Enabling Extreme-Scale Scientific Insight, (October 31, 2012)

[26] World-wide Large Hadron Collider Computing Grid (WLCG) Web page, http://wlcg.web.cern.ch

[27] I. Foster, C. Kesselmann, and S. Tuecke. Grid Computing - Making the Global Infrastructure a Reality, chapter, "The Anatomy of the Grid - Enable Scalable Virtual Organizations", pages 171–198. John Wiley & Sons Ltd, ISBN 0470853190, 2003.

[28] J. Bicarregui, S. Lambert, B. Matthews, M. Wilson, *"PaNdata: a European data infrastructure for neutron and proton sources"*, in Proceedings of e-Science All Hands Meeting 2011 AHM'11. York, UK.

[29] J. Demotes-Mainard, *"Towards a European clinical research infrastructure network: the ECRIN programme Therapie",* PubMed - US National Library of Medicine - National Institutes of Health 59(1):151-153

[30] Paris J, Ciais P, Rivier L, Chevallier F, Dolman H, Flaud J, Garrec C, Gerbig C, Grace J, Huertas E, Johannessen T, Jordan A, Levin I, Papale D, Valentini R, Watson A (2012) Integrated carbon observation system EGU general assembly, 2012. Vienna, Austria.

[31] ELIXIR Consortium (2012) European life sciences infrastructure for biological information Web Page, http://www.elixir-europe.org/news

[32] DIXA Consortium (2012) diXa: A new data infrastructure for chemical safety Web page, http://www.dixa-fp7.eu/home

[33] G. Barrand, I. Belyaev, P. Binko, M. Cattaneo, R. Chytracek, G. Corti, M. Frank et al. *"GAUDI—A software architecture and framework for building HEP data processing applications"*, *Computer physics communications* 140, no. 1 (2001): 45-55.

[34] T. Hoefler, A. Lumsdaine, and J. Dongarra, *"Towards Efficient MapReduce Using MPI",* Recent Advances in Parallel Virtual Machine and Message Passing Interface Lecture 5759/2009, pp. 240---249

Scaling the Performance and Cost While Scaling the Load and Resources in the Cloud

Monika Simjanoska, Sasko Ristov, Goran Velkoski and Marjan Gusev
Ss. Cyril and Methodius University,
Faculty of Information Sciences and Computer Engineering,
Skopje, Macedonia
Email: m.simjanoska@gmail.com, sashko.ristov@finki.ukim.mk, velkoski.goran@gmail.com, marjan.gushev@finki.ukim.mk

***Abstract*—Cloud computing is a paradigm that offers on-demand scalable resources with the "pay-per-usage" model. Price rises linearly as the resources scale. However, the main challenge for cloud customers is whether the performance is also scaling as the price for the resources. In this paper we analyze both the performance and the cost of a memory demanding web service. The experiments are based on measuring the performance and calculating the costs of rented CPU resources for different server loads, obtained by changing the message size and the number of concurrent messages. The results show that the lowest cost is obtained while the memory demanding service is hosted on two CPUs.**

***Index Terms*—Cloud Computing, Web Services, Performance, Resources, Cost**

I. Introduction

According to the NIST definition [1], cloud computing is a model for enabling ubiquitous, convenient, on demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. From a hardware point of view, there are three aspects in cloud computing: 1) the illusion of infinite computing resources available on demand, 2) the elimination of an up-front commitment by cloud user; and 3) the ability to pay for use of computing resources on a short-term basis as needed [2]. This usage-based pricing model offers several advantages, including reduced capital expense, a low barrier to entry, and the ability to scale up as demand requires, as well as to support brief surges in capacity [3]. However, service providers are not able to accomplish all customers' demands for services since there is a high level of alterations which have to be negotiated via Service Level Agreements (SLAs) [4].

The service provider's "pay-per-usage" model offers a pricing scheme scalable to the amount of rented resources. A natural expectation from this kind of pricing model is that the performance gain also scales to the monetary costs, i.e. the more resources are rented, the more performance is achieved. Even though, cloud service providers guarantee the availability of the rented resources to the customers by defining SLAs, a guarantee of scalable and sustainable performance misses in the SLAs [5].

For the purpose of customers' contentment, it is very important to perform a research to answer the question if the performance proportionally scales to the price paid for renting resources. This question is not so simple as it looks like. The performance depends mostly on one or several resource parameters. Increasing the parameter which is not a performance bottleneck can even reduce the performance, while increasing a parameter that causes performance bottleneck can provide a significant speedup. For example, increasing the CPU speed will improve the matrix multiplication algorithm only several percents (the computations), but increasing the number of processors will achieve maximum linear speedup according to Gustafson's Law [6]. Using the processors with dedicated cache per core can achieve even superlinear speedup (the speedup greater than the number of processors). On the other hand, scaling is limited with the serial part of the algorithm, i.e. the Amdahl's Law [7].

The cloud service provider's renting model is usually linear with constant price / performance ratio for main computing resources: CPU, main memory and storage disk. Table I presents the details about current offers in Jan 2013 for renting the following virtual machine (VM) instances: Azure [8], Google [9] and Amazon [10]. The relative scaling factor is also presented for different resources.

TABLE I
VM instance types and price simulation

Type	1 VM	2 VMs	4 VMs	8 VMs
Windows Azure (in $/h)	0,08	0,16	0,32	0,64
Google Compute (in $/h)	0,151	0,302	0,604	1,208
Amazon EC2 (in $/h)	0,065	0,13	0,26	0,52
Scaling factor	1	2	4	8

In this paper, we consider single user case whose monetary costs for resources are proportional to the amount of rented resources. Therefore, our performance analysis are based on the quantity of acquired resources. Hereupon, we defined three different cloud environments with virtual machine instances, each with different number of CPU cores, and thus, simulated different amount of rented resources. Furthermore, in order to obtain comparable results, we loaded each cloud environment with the same load, i.e., the same number of concurrent messages with the same message size. In order to simulate realistic demands for service, in each cloud environment we hosted a memory demanding ($Concat$) web service. This is a simple web service that returns a concatenation of two input

strings.

The rest of the paper is following the next organization. A brief review of the related work for scaling the performance and cost in the cloud is presented in Section II. In Section III we present the developed methodology used for testing and obtaining reliable results. The experiments and the outcomes are exhibited in Section IV. In the final Section V, we derive conclusions over the results and we present our ideas for future research.

II. Related Work

In this section we present a brief review of the recent research closely related to our field of interest.

The recent research related to a "cost-efficient" cloud computing mostly examine the cloud service providers' costs for offering cloud computing solution. For example, several papers have analyzed the cloud computing performance to find out if the cloud is energy-efficient and therefore, cost-efficient [11], [12], [13], [14]. Some of the most important cloud computing issues and challenges are discussed by Zhang [15] stating that with on-demand resource provisioning and utility based pricing, service providers can truly maximize resource utilization and minimize their operating costs.

Several papers refer to different behavior of performance when scaling. Lu et al. [16] discovered several pitfalls resulting in waste of active VMs idling. They examine several pitfalls in Windows Azure Cloud during several days of performing the experiments: Instance physical failure, Storage exception, System update. Windows Azure does not work well for tightly-coupled applications [17].

However, in this paper, we consider the customers' benefits of the on-demand resource provisioning as maximized performance with minimal costs. Similar research of this kind is presented by De Assuncao et al. [18], where the authors present several scheduling strategies for balancing between performance and usage cost, and how much they improve the requests' response times. Their results show that some of the strategies result in a higher cost under heavy load conditions, whereas some showed a good ratio of slowdown improvement to the money spent for using cloud resources. Andrzejak et al. formulated a probabilistic model that enables a user to optimize monetary costs, performance, and reliability, given the user's SLA constraints as resource availability and deadline for job completion [19]. Using their model, the users can achieve largest cost savings (for considered workload types), by using the high-CPU instance types instead of standard or high-memory instance types. The authors' contribution in [20] is developing a service that is able to perform the cost determination for scientific applications in cloud computing environments. Kondo et al. compare and contrast the performance and monetary cost-benefits of clouds for computing applications, ranging in size and storage [21].

Considering the performance, using cache intensive algorithms in both single-tenant and multi-tenant cloud environments, Gusev and Ristov show how and when cloud computing can achieve even better performance than traditional environment for certain workload [22]. Another research for cloud's performance states that the cloud achieves smaller performance degradation for greater message sizes using the memory demanding web service, and also for greater message sizes and smaller number of concurrent messages for memory demanding and compute intensive web services [23].

As we covered the state of the art related to the problem of scaling performance and cost in the cloud, we proceed with research methodology to find out if the performance rises linearly to the cost.

III. Research Methodology

In order to perform realistic experiments, we developed original methodology that includes technical details, appropriate cloud environment configuring to include all three test cases with different number of resources, and the testing procedure itself. Considering the outcomes from the experiments we present relevant mathematical relations to obtain reliable results.

A. Technical Framework

As a testing environment we used client-server architecture deployed in the open source cloud platform OpenStack [24] using Kernel-based Virtual Machine (KVM) hypervisor to instantiate VM instances. The client and server node are installed with Linux Ubuntu Server 12.04 operating system. Hardware computing resources consist of Intel(R) Xeon(R) CPU X5647 @ 2.93GHz with 4 cores and 8GB RAM. The VM instances consist of Linux Ubuntu Server 12.04 operating system and Apache Tomcat 6 as the application server. To minimize the network latency we placed the client and the VMs in the same LAN segment [25].

B. Environment Configuration

In order to simulate various number of provided resources (CPU cores), we defined three different cloud environments:

- *Test Case 1*: VM instance with 1 CPU;
- *Test Case 2*: VM instance with 2 CPUs;
- *Test Case 3*: VM instance with 4 CPUs.

Each cloud environment hosts the web service $Concat$, that accepts two strings and returns their concatenation. It is only memory demanding web service, since it does not require any processing power.

C. Testing Procedure

The client uses SoapUI [26] to generate various server loads. Each instance is loaded with N messages with parameters parameter size of M kilobytes each, with variance 0.5. This means that the number of threads will vary by $N/2$, i.e. the number of threads will increase to $3 \cdot N/2$, then decrease to $N/2$, and finally end with N within 60 seconds, i.e. the end of the test.

The range of parameters M and N is selected such that web servers in VM instances work in normal mode without replying error messages and avoiding saturation. The web service is loaded with $N = 12; 100; 500; 752; 1000; 1252; 1500; 1752$

and 2000 requests per second for each message parameter size $M = 0; 1; 2; 4; 5; 6; 7; 8$ and 9. In order to simulate different connections per core we divide the N concurrent messages in four groups of $N/4$ messages each.

D. Test Data

In the research experiments we measure the server's average response time for different parameter sizes M, and number of concurrent messages N. The experiments are realized for each test case, as defined by the environments in III-B. These measurements will express the cloud's performance. The equation (1) presents the scaled average response time T, where A denotes the average response time, n is the number of used cores, and 4 denotes the different connections per core.

$$T(n) = \frac{A(n)}{4} \tag{1}$$

Pursuant to our goal of expressing the performance through cost evaluation, in equation (2) we calculate the cost C, where $T(n)$ denotes the response time derived from equation (1) and n is the total number of processors used.

$$C(n) = T(n) * n \tag{2}$$

Our research problem checks whether the performance is proportional to the number of rented resources, therefore, the web services' total cost C is the real cost of rented CPU resources. For this purpose we will calculate relative performance and relative cost of the scaling problem, as expressed in (3), (4) and (5), as ratio of test cases with VMs 2 and 1 CPU cores; VMs 4 and 1 CPU cores; and VMs with 4 and 2 CPU cores, correspondingly.

$$R_{21} = \frac{C(2)}{C(1)} \tag{3}$$

$$R_{41} = \frac{C(4)}{C(1)} \tag{4}$$

$$R_{42} = \frac{C(4)}{C(2)} \tag{5}$$

Ideal expectation will be the proportional scaling, i.e. when $R_{21} = 2$, $R_{41} = 4$ and $R_{42} = 2$. Any deviation from these expectations will make new conclusions in this research.

IV. Experiments and Results

In this section we present the experiments and the results using the methodology specified in III.

A. Analysis of Response Time

The memory demanding *Concat* web service has been hosted in VM instances with 1, 2 and 4 CPU cores.

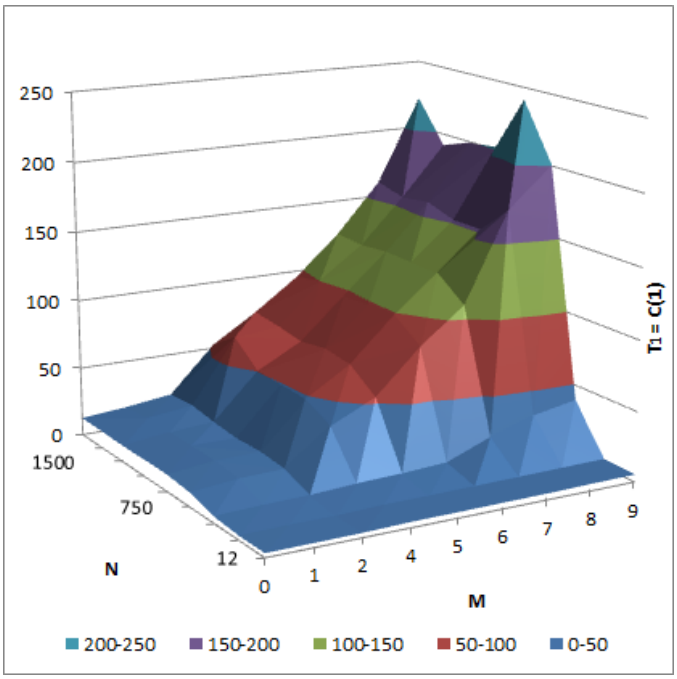


Fig. 1. Cost = response time for *Concat* web service $(1x1)$

1) Test Case 1 - VM with 1 CPU core: Figure 1 presents the results from the cloud environment with 1 VM instance with 1 CPU core. According to the results, the response time depends equally on the message parameter size, M, and number of concurrent messages, N. Figure 1 denotes that $C(1)$ is equal to $T(1)$ because the number of used CPU is 1, and if applied in equation in (2), the cost and the response time remain the same.

For a simplified presentation we denote the points in the format (M, N), where M and N refer to both the parameters we previously defined. Thus, the minimum value of 3.095 is in the point $(0, 12)$, and the maximum value of 241.35 is in the point $(9, 1000)$. Considering that the response time proportionally increases as both of the parameters M and N increase, we find the minimum value is in the expected point, but the maximum value seems to be an unexpected pick. The average value is 58.24.

2) Test Case 2 - VM with 2 CPU cores: The results from the cloud environment with 1 VM instance with 2 CPU cores are presented in Figure 2. The minimum value of 2.38 is at the point $(0, 100)$, whereas the maximum value of 59.08 is again at the point $(9, 1000)$. According to the average value of 16.24, we assume that the response time has decreased 3.6 times in comparison to the cloud with 1 VM instance with 1 CPU. In the next section we use that the performance gain of a VM with 2 cores over a VM with one core has value of 3.6.

3) Test Case 3 - VM with 4 CPU cores: Figure 3 depicts the results from the cloud environment with 1 VM instance with 4 CPU cores. The minimum value of 3.02 at $(0, 12)$ shows that for small number of concurrent messages and small message parameter size, there is minority variation in the response time

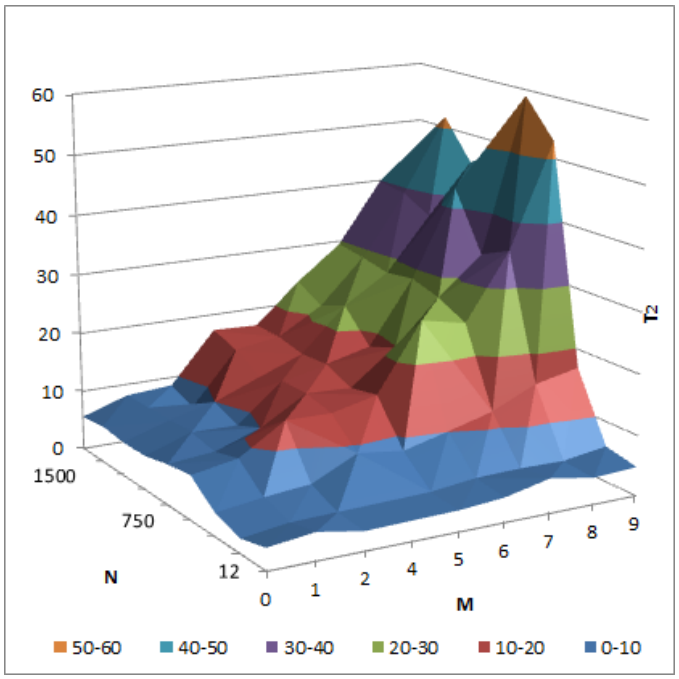


Fig. 2. Response time for *Concat* web service $(1x2)$

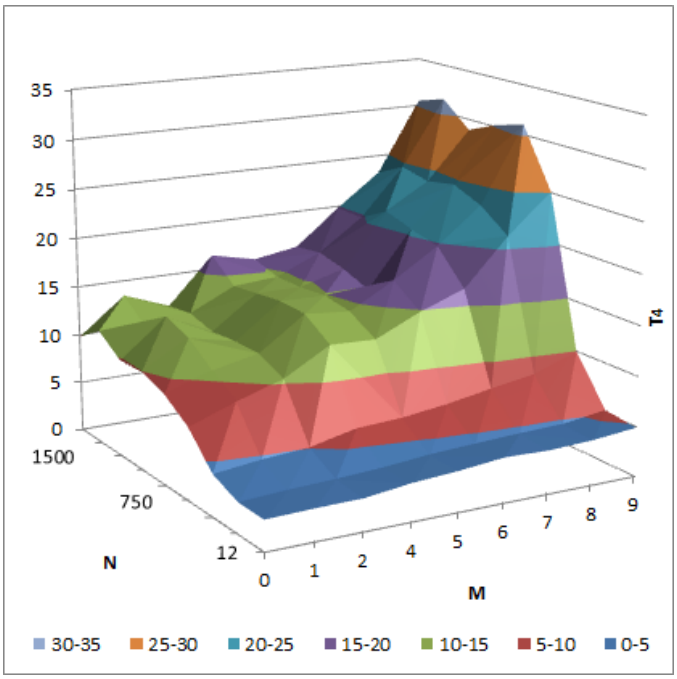


Fig. 3. Response time for *Concat* web service $(1x4)$

despite of the increased number of resources. The maximum value of 31.39 is at the point $(9, 1750)$, and the average value is 12.78 which means that the response time decreased 4.6 times compared to the cloud VM with 1 core, i.e. the performance gain of a VM with 4 CPU cores compared to a VM with 1 CPU core is 4.6. We have also calculated the performance gain of a VM with 4 CPU cores over a VM with 2 CPU cores to be 1.3.

B. Cost Analysis

In order to derive conclusions about the sufficient trade-off between the cost and the gained performance, we used the equation in (2) to calculate the customer's cost for resources using the average response time calculated when using the equation in (1).

Observing the average response time decrease in IV-A, we have concluded that scaling up the resources n times, where n is the number of cores, provides even more that n times performance gain. Hereupon, we aim to find out answer for our research problem, i.e. whether the cost for resources is equal to the performance gain, moreover, whether there is an occasion where customers pay less than they gain.

The default cost per core is the one core cloud environment, i.e. the results from the cloud environment with 1 VM instance with 1 CPU as presented in Figure 1.

The relative ratio between the costs for the cloud VM with 2 CPUs and cloud VM with 1 CPU, calculated by (3) is depicted in Figure 4. As a cost threshold we used the average performance decrease of 3.6, thus, if the cost transcend this value, a customer is considered to pay more than he gets. Interestingly, we observe that the customer's costs for

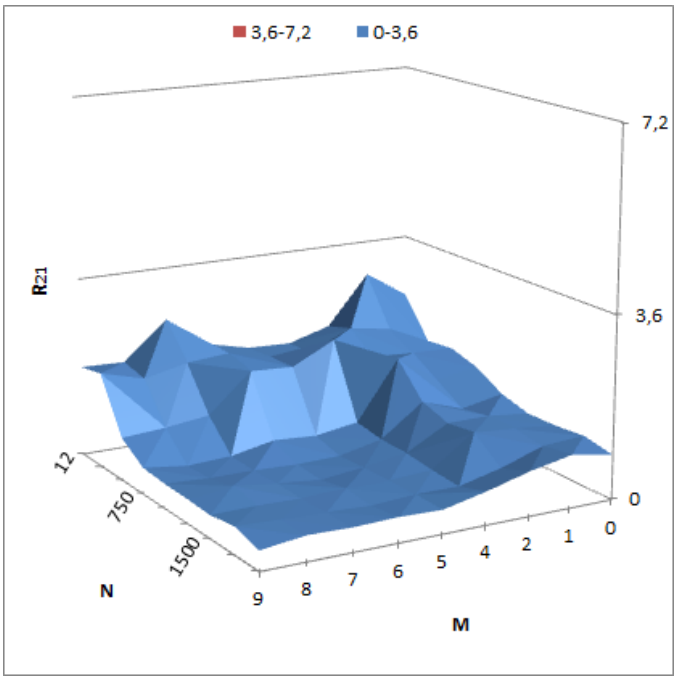


Fig. 4. Cost analysis for *Concat* web service R_{21}

resources remain far beyond the threshold value.

Figure 5 depicts the proportion of the cost for the cloud VM with 4 CPUs and the cost for a cloud VM with 1 core, calculated by (4). The results also show that for any message parameter size and number of concurrent messages, the customer's costs are much lower than the performance advantage.

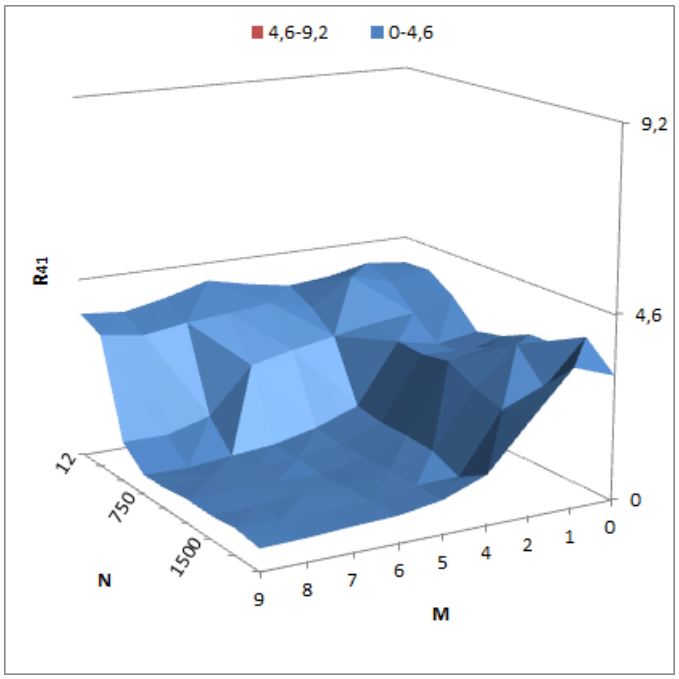


Fig. 5. Cost analysis for $Concat$ web service R_{41}

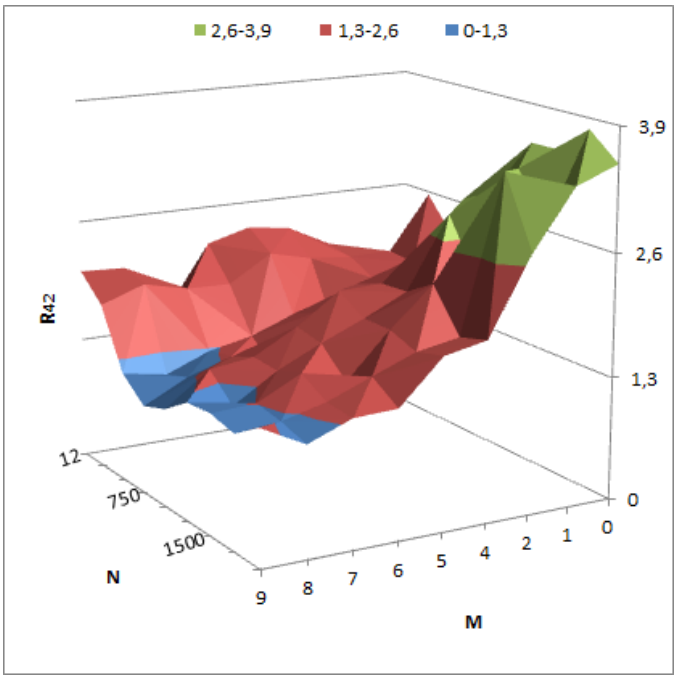


Fig. 6. Cost analysis for $Concat$ web service R_{42}

Interestingly, incrementing the number of concurrent messages, N, and the message size, M, the cost decreases at both scenarios.

Eventually we compared the costs between the cloud VM instance with 4 CPUs and cloud VM instance with 2 CPUs, calculated by (5). The reason for this analysis is realizing if the customer will make a good decision demanding 2 more CPUs once he used 2 cores. When presenting the results on Figure 6 we used 1.3 value as a threshold which we obtained from the results in IV-A where we concluded that the performance gain is 1.3 times for VM with 4 cores in comparison to VM with 2 cores. However, the results show that there are small regions where the customer pays less than he gets, thus, it is not worthwhile renting 2 more cores once the customer has 2 cores.

V. Conclusion and Future Work

Cloud computing pricing pay-per-usage model makes the paying be linearly scalable to the amount of rented resources. The most important question from a customers point of view is whether the performance also scales to quantity of resources acquired and thus, whether it proportionally varies with the customer's monetary costs.

In this paper we performed a series of experiments to answer this question and to derive conclusions for which cloud configuration, if any, the customer gets more performance than the amount paid.

Observing all scenarios, we conclude that the $Concat$ web service while hosted on 2 CPUs provides 3.6 time better performance than when hosted on 1 CPU, which is the best occasion from all test cases. Moreover, it also provides the lowest cost in comparison to the gained performance.

On the other hand, the worst results are obtained when the customer migrates from 2 CPUs to 4 CPUs, i.e. the cost is much above the performance benefit.

However, the results showed satisfying trade-off between the cost and the performance for the memory demanding web service. In our future work we aim to extend the research including a web service which is not only memory demanding, but also depends on other characteristics like computation or I/O intensive. Furthermore, we aim to perform the same analysis in a multi-tenant cloud since we expect the performance to be interfered because large number of users share the same infrastructure.

References

[1] P. Mell and T. Grance, "The NIST definition of cloud computing," National Institute of Standards and Technology, Information Technology Laboratory, Sep. 2011.

[2] A. Fox, R. Griffith *et al.*, "Above the clouds: A berkeley view of cloud computing," *Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Tech. Rep. UCB/EECS*, vol. 28, 2009.

[3] R. Grossman, "The case for cloud computing," *IT professional*, vol. 11, no. 2, pp. 23–27, 2009.

[4] P. Patel, A. Ranabahu, and A. Sheth, "Service level agreement in cloud computing," in *Cloud Workshops at OOPSLA*, 2009.

[5] D. Durkee, "Why cloud computing will never be free," *Queue*, vol. 8, no. 4, p. 20, 2010.

[6] J. L. Gustafson, "Reevaluating Amdahl's law," *Communication of ACM*, vol. 31, no. 5, pp. 532–533, May 1988.

[7] G. M. Amdahl, "Validity of the single-processor approach to achieving large scale computing capabilities," in *AFIPS Conference Proceedings*, vol. 30. AFIPS Press, Reston. Va., Atlantic City, N.J., Apr. 18-20 1967, pp. 483–485.

[8] Microsoft, "Windows azure," July 2012. [Online]. Available: http://www.windowsazure.com/pricing/

[9] Google, "Compute engine," July 2012. [Online]. Available: http://cloud.google.com/pricing/
[10] Amazon, "Ec2," 2012. [Online]. Available: http://aws.amazon.com/ec2/
[11] A. Berl, E. Gelenbe, M. Di Girolamo, G. Giuliani, H. De Meer, M. Dang, and K. Pentikousis, "Energy-efficient cloud computing," *The Computer Journal*, vol. 53, no. 7, pp. 1045–1051, 2010.
[12] J. Baliga, R. Ayre, K. Hinton, and R. Tucker, "Green cloud computing: Balancing energy in processing, storage, and transport," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 149–167, 2011.
[13] A. Beloglazov and R. Buyya, "Energy efficient resource management in virtualized cloud data centers," in *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, ser. CCGRID '10, 2010, pp. 826–831.
[14] Y. Lee and A. Zomaya, "Energy efficient utilization of resources in cloud computing systems," *The Journal of Supercomputing*, vol. 60, no. 2, pp. 268–280, 2012.
[15] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," *Journal of Internet Services and Applications*, vol. 1, no. 1, pp. 7–18, 2010.
[16] W. Lu, J. Jackson, J. Ekanayake, R. S. Barga, and N. Araujo, "Performing large science experiments on azure: Pitfalls and solutions," in *CloudCom'10*, 2010, pp. 209–217.
[17] V. Subramanian, H. Ma, L. Wang, E.-J. Lee, and P. Chen, "Rapid 3d seismic source inversion using windows azure and amazon ec2," in *Proc. of IEEE*, ser. SERVICES '11. IEEE Comp. Soc., 2011, pp. 602–606.
[18] M. De Assunção, A. Di Costanzo, and R. Buyya, "Evaluating the cost-benefit of using cloud computing to extend the capacity of clusters," in *Proceedings of the 18th ACM international symposium on High performance distributed computing*. ACM, 2009, pp. 141–150.
[19] A. Andrzejak, D. Kondo, and S. Yi, "Decision model for cloud computing under sla constraints," in *Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 257–266.
[20] H. Truong and S. Dustdar, "Composable cost estimation and monitoring for computational applications in cloud computing environments," *Procedia Computer Science*, vol. 1, no. 1, pp. 2175–2184, 2010.
[21] D. Kondo, B. Javadi, P. Malecot, F. Cappello, and D. Anderson, "Cost-benefit analysis of cloud computing versus desktop grids," in *Parallel & Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on*. IEEE, 2009, pp. 1–12.
[22] M. Gusev and S. Ristov, "The optimal resource allocation among virtual machines in cloud computing," in *CLOUD COMPUTING 2012, The Third International Conference on Cloud Computing, GRIDs, and Virtualization*, 2012, pp. 36–42.
[23] S. Ristov, G. Velkoski, M. Gusev, and K. Kjiroski, "Compute and memory intensive web service performance in the cloud," in *to be published in ICT Innovations 2012*. Springer Berlin / Heidelberg, 2012.
[24] OpenStack, "Openstack cloud software," Jan. 2013. [Online]. Available: http://openstack.org
[25] M. Juric, I. Rozman, B. Brumen, M. Colnaric, and M. Hericko, "Comparison of performance of web services, ws-security, rmi, and rmi–ssl," *Journal of Systems and Software*, vol. 79, no. 5, pp. 689–700, 2006.
[26] SoapUI, "Functional testing tool for web service testing," Jan. 2013. [Online]. Available: http://www.soapui.org/

Semidefinite optimization of High Performance Linpack on heterogeneous cluster

Draško Tomić*, Luko Gjenero**. Emir Imamagić**
*Hewlett-Packard Croatia, Zagreb, Croatia
** University of Zagreb University Computing Centre, Zagreb, Croatia

Abstract – High Performance Linpack (HPL) is an industry standard benchmark used in measuring the computational power of High Performance Clusters. In contrary to HPC clusters consisting of equal computational nodes, running HPL on heterogeneous HPC clusters, built up of a computing nodes with different computational power, showed in most cases poor efficiency. In such type of clusters, efficiency of HPL further decreases if the speed of interconnect links between computing nodes is different. In order to improve HPL efficiency on such a clusters, one needs to optimally balance HPL workload on computing nodes accordingly to their computational power, and at the same time, take into the consideration the speed of communication links between them. Our thesis is that the problem of efficiently running HPL on heterogeneous HPC cluster is solvable, and that one can formulate it as a Semidefinite Optimization of Second Eigenvalue in Magnitude (SLEM) matrix describing data-flow of HPL in a cluster. In order to test a validity of such an approach, we run a series of HPL benchmarks on Isabella HPC cluster, both optimized respective to SLEM and non-optimized. By comparing results obtained with SLEM optimization of HPL against non-optimized HPL, we were able to identify a huge improvement in HPL efficiency when using SLEM. Moreover, by taking into the consideration memory sizes of computational nodes, we were able to improve SLEM optimization of HPL further.

I. INTRODUCTION

High Performance Linpack (HPL) is an industry standard benchmark used in measuring performance of High Performance Computing (HPC) clusters [1]. Obtained Rmax value represents a number of double-floating point operations achieved during HPL run. Due to its scalability, HPL runs efficiently on clusters with nearly hundred thousand of computing nodes [2]. Scalability of standard HPL applies to homogeneous clusters only, i.e. clusters with equal computational nodes. This is due to the nature of HPL, which executes on computing nodes in a large number of iterations, and all nodes needs to synchronize before the next iteration starts. Therefore, in heterogeneous HPC clusters, slower computational nodes force faster nodes to wait until each iteration cycle is finished, and before the next iteration cycle begins. On such systems, depending on the difference in computational power of nodes, considerable degradation of HPL efficiency, defined as a rate between obtained Rmax and theoretical Rpeak values, exists [3]. In order to improve HPL efficiency, one has to deploy clever load-balancing strategies, and implement them by modifying HPL code to support heterogeneous clusters [4]. This is not always a simple task, because rewriting, debugging and testing HPL code are tasks usually requiring weeks, if not even months of work. In addition, it is hardly possible to obtain exclusive reservation of certain HPC cluster for HPL benchmarking in a period exceeding more than several weeks. Another approach of running HPL more efficiently on heterogeneous HPC clusters is to logically break cluster into smaller domains, each domain with equal resources, and run HPL replicas in domains concurrently, each domain executing one HPL replica. When HPL replicas finished their runs, one sums Rmax values over all domains. Although this approach leads more quickly towards results and brings better HPL efficiency, it does not account for communication delays and synchronization effects between domains, therefore delivering results that can be quite misleading. Because of the considerations mentioned, it is cleverer to take version of HPL that is already rewritten and ready to run efficiently on heterogeneous HPC clusters. One of such HPL versions is HPL developed at University of Frankfurt, Frankfurt Institute for Advanced Studies, which relies on standard HPL, but is rewritten and accommodated, in order to run efficiently on heterogeneous HPC clusters [5]. This rewritten HPL, in further text called heterogeneous HPL, allows balancing a workload of HPL within HPC cluster by taking into the consideration processing power of nodes. Each computational node gets a real number from interval [0, 1], depending of its processing power, with fastest node(s) valued 1, and slowest valued with lover-valued coefficient. One way of the quantization of coefficients for nodes is according to the publicly available benchmarking data, e.g. SPEC rates, or by directly running HPL on certain nodes and measuring Rmax values. Another strategy may bring even higher Rmax values, and this is to spread and run HPL replicas with resized data on all computational nodes, collect performance info of nodes executing replicas, and as a last step, run again HPL replicas on nodes, but with data resized according to the measured performance of nodes.

Performance of nodes and memory size are not always only critical parameters for the efficient HPL run. Especially in HPC Cloud systems, interconnection links between computational nodes are often of different speeds, making this approach worthless. Fortunately, it is possible to recast a problem of optimally balancing a load on nodes with different processing power and different speed of communication paths between them as a problem of minimizing Second Largest Eigenvalue in Magnitude (SLEM) of matrix representing data-flow. First formulation of such an approach appeared in [7], and then further developed in [8]. Nevertheless, its practical use was not possible before the advent of algorithm for minimizing second eigenvalue of doubly stochastic matrix [9]. After that, proposals have been done for applying SLEM in e-learning [10] and high performance computing [11] areas. Despite of this, while running HPL on Isabella HPC cluster, located at the University of Zagreb University Computing Centre, we realized that HPL does not entirely fit into the framework of SLEM. Namely, we identified inefficient use of memory as a cause for degradation of HPL efficiency. Further to this, we were able to develop and add memory requirements into already existing framework of SLEM, thus making SLEM more efficient for HPL.

II. RUNNING HIGH PERFORMANCE LINPACK ON ISABELLA CLUSTER

We performed a series of HPL benchmarks on Isabella cluster. Two parts of Isabella, hpc1 and hpc2 cluster, were available to us for benchmarking:

Server model	CPU type	Nr. of cores	RAM size	Nr. of servers
Sun Fire X4600 M2	AMD Opteron(tm) Processor 8384	32	64	8
Dell PowerEdge	Intel(R) Xeon(R) CPU X7542 @ 2.67GHz	24	128	4
HP ProLiant DL585 G1	AMD Opteron (tm) Processor 875	8	16	16
Sun Fire X4600	AMD Opteron(tm) Processor 8218	16	32	4
HP ProLiant DL580 G5	Intel(R) Xeon(R) CPU E7330 @ 2.40GHz	16	24	16

Table 1: hpc1 cluster computational nodes

Server model	CPU type	Nr. of cores	RAM size	Nr. of servers
HP ProLiant SL230 Gen8	Intel(R) Xeon(R) CPU E5-2665 @ 2.40GHz	16	64	28
Dell PowerEdge	Intel(R) Xeon(R) CPU X7542 @ 2.67GHz	24	128	4

Table 2: hpc2 cluster computational nodes

In hpc1 cluster, the interconnection between nodes was SilverStorm SDR 10Gb/s Infiniband switch, while in hpc2 cluster Dell Poweredge nodes were interconnected via the same switch and HP SL250 nodes via Qlogic QDR 40Gb/s Infiniband switch. In order to form a unique cluster from these two groups of nodes, we interconnected both switches with four 10Gb/s Infiniband links. The same operating environment was on both hpc1 and hpc2 clusters: i.e. MVAPICH 1.1.0 for Message Passing Interface system, GNU C Compiler version 4.1.2, CentOS operating system version 6.3, an implementation of Basic Linear Algebra Subroutines GotoBLAS version 1.13, and Scilab version 5.3.3, an open source software for numerical computing available from [14], as a platform for SLEM based workload optimization. For standard HPL we used a portable implementation of the High-Performance Linpack version 2.0 from [15], while for heterogeneous HPL we run version 1.1.0 from [16]. We decided to run HPL on both clusters in three different ways: First one with standard HPL, second one with heterogeneous HPL and load parameters computed according to the SPEC rates found on [12], and the last one again with heterogeneous HPL, but with load parameters obtained from SLEM. We choose such a benchmarking strategy not only to find highest Rmax of hpc1 and hpc2 clusters, but also to test both effectiveness of standard versus heterogeneous HPL and SLEM. While running heterogeneous HPL on hpc1 cluster for the first time, both with human and from SLEM obtained parameters, we obtained very low Rmax results compared to those obtained with standard HPL. After detailed inspection, we found two reasons for this. Firstly, if HPL data overflow the available RAM of certain computational node, system starts to page data from RAM to relatively slow spindles, and cluster performance degrades. In contrary, when HPL matrix does not fully utilize available RAM, then a portion of relatively slow MPI communication between nodess is bigger, thus slowing down cluster again. We also realized that optimum size of HPL matrix is about total RAM size minus size of memory region used by operating system.
Thus, in order to run heterogeneous HPL efficiently, one needs to partition and scatter HPL matrix on available nodes not only with respect to their computational power, but with respect to their available memory as well; i.e. to correlate a size of HPL matrix data with a memory available on each node.

If we denote total cluster memory size with *CM*, memory size of node *n* with M_n, and a region reserved for operating system activity within node with OS_n, then a ratio between memory available for HPL on node *n* and cluster with *N* nodes can be defined as:

$$cor_n = \frac{M_n - OS_n}{CM - N \times OS_n} \quad (1)$$

Defining matrix parameters for standard HPL is straightforward process: The three most important input parameters one needs to set in HPL.dat input file are problem size *N*, block size *NB* and process grid ratio P x Q. The good starting point for setting these parameters for optimal HPL efficiency is [13]. The amount of cluster memory used by HPL is essentially the size of the coefficient matrix. As each element of matrix is of type double-precision floating point, it occupies eight bytes of memory. Therefore, if we denote the dimension of the HPL coefficient matrix with L, and the available memory and the region sizes in bytes reserved for operating system of a certain node *n* with M_n and OS_n respectively, the condition for maximal HPL efficiency is:

$$L \approx \sqrt{\frac{\sum_{n=1}^{N}(M_n - OS_n)}{8}} \quad (2)$$

In general, HPL coefficient matrix occupies about 80% percent of the available memory size of a certain node; therefore, it can represent a good starting value. This value can vary depending of the operating system, and further refinement is necessary if one intends to obtain better HPL efficiency. In opposite to standard HPL, heterogeneous HPL has an additional input file where one can define load parameters $p_1,...p_n...p_N$ that represent computational power of nodes. As we found that memory available to HPL is of the outmost importance for its efficiency, we modified load parameters by multiplying them with correlation factors $cor_1...cor_n...cor_N$ computed from (1). Then, the following tuple *P* represents modified load parameters:

$$P = p_1 cor_1 ... p_n cor_n ... p_N cor_N \quad (3)$$

Instead of estimating p_n values by using SPEC data, and especially for larger HPC cluster, the more clever approach is to obtain them directly, i.e. by simultaneously running HPL replicas on all nodes, and collecting performance info of them. Then, this info can serve for forming *P*, i.e. input file for heterogeneous HPL.

Such an approach works fine for heterogeneous nodes interconnected with communication links of equal speed, but fails if interconnection links are of different throughput. Then, one has to perform an optimization not only according to the processing power of computational nodes, but also needs to take into the account throughput of communication links. As this is a matrix optimization problem, our idea was to use already developed semidefinite optimization algorithm for balancing a load on heterogeneous clusters, and this is SLEM. Besides, it was an excellent opportunity to test if SLEM delivers the best possible load balancing proposals. SLEM relies on semidefinite optimization of the matrix representing data-flow in a certain computing system. More precisely, this algorithm performs the minimization of the second largest eigenvalue (in magnitude) of this matrix; when it finds minimum, trace of this optimized matrix represents optimal values of load on computational nodes. In order to avoid possible confusion, it is important to stress out that with SLEM, optimal load balancing means the fastest possible execution of a certain application, and not for example maximal utilization of computational nodes. Therefore we performed HPL benchmarking on both clusters first with standard HPL, then heterogeneous HPL with load parameters based on SPEC performance data, and lastly, again with heterogeneous HPL, but with load parameters obtained from SLEM.

III. DOING SEMIDEFINITE OPTIMIZATION WITH IMPROVED MEMORY MANAGEMENT

On Isabella cluster, we implemented in [8] developed SLEM program. This program is a collection of Linux scripts and MPI procedures running on a master node outside the computational farm, with the main script able to scatter and start agents on available cluster nodes. Agents trigger performance benchmarking of their landing nodes, by running a mix of elementary double-floating operations in a large number of iterations. After that, agents mutually synchronize, and then start communication benchmarking of the interconnect links between nodes. Communication benchmark is performing a collective MPI operation, with each agent sending and receiving data to and from all other agents, via MPI send and receive operations. Finally, they exchange collected performance data with a master node, where SLEM procedure written in Scilab package computes optimal load values for cluster nodes. As we were already aware from the former research in [8] that challenging nodes with real application instead with synthetic mix of floating point operations delivers better results, we changed agents in the way they execute heterogeneous HPL replicas instead of synthetic benchmark. We also realized that heterogeneous HPL delivers the best results when memory of nodes is almost fully loaded with HPL data. For this reasons we modified load data derived from SLEM according to (3).

Although a detailed review of steps SLEM is in [11], we give next a brief overview of SLEM optimization, with emphasis on places where it deviates from original proposal [8]. As stated in the previous section, SLEM relies on the semidefinite optimization of matrix representing data-flow in a certain computer system, with respect to its second largest eigenvalue (in magnitude). As of today, we know only algorithm that works on doubly stochastic and symmetric matrices. In order to make it useful also for deterministic systems, one needs to know how to transform deterministic matrix to doubly stochastic and vice versa. Let us state the optimization problem as:

$$Minimize\ \ \mu(P) \tag{4}$$

With respect to:

$$P \geq 0, P1 = 1,\ P = P^T \tag{5}$$

Where P is doubly stochastic and symmetric matrix, its transpose is P^T, $\mu(P)$ is SLEM, and 1 is unary vector. Additionally, let us denote the performance of a certain computing node with eps_i, and speed of the communication link between processing nodes *i* and *j* with bw_{ij}. In case of *n* computational nodes, the following matrix describe effective processing speed of computing nodes and speed of communication links between them:

$$Y = \begin{bmatrix} eps_1 & bw_{12} & . & bw_{1n} \\ bw_{21} & eps_2 & . & bw_{2n} \\ . & , & . & . \\ bw_{n1} & . & bw_{n,n-1} & eps_n \end{bmatrix} \tag{6}$$

In general, communication links are bidirectional and not of equal speed in both directions. Thus, we can write symmetric form of matrix *Y*:

$$Y_s = \begin{bmatrix} eps_1 & \min(bw_{12}, bw_{21}) & . & \min(bw_{1n}, bw_{n1}) \\ \min(bw_{21}, bw_{12}) & eps_2 & . & \min(bw_{2n}, bw_{n2}) \\ . & , & . & . \\ \min(bw_{n1}, bw_{1n}) & . & \min(bw_{n,n-1}, bw_{n-1,n}) & eps_n \end{bmatrix} \tag{7}$$

On the other side, construction of a matrix describing HPL algorithm showed to be practically impossible in a short time, mostly due to the various options available in HPL. For this reason, we take an algorithm that from topological point of view is similar to HPL, and for which it is relatively easy to develop such a matrix, i.e. Laplace solver.

In general, we can construct weighted and vertex labeled directed graph, with edge weights representing number of messages exchanged between them, and vertex labels representing number of operations performed within nodes. From this graph, one can construct matrix *A*, with main diagonal elements representing number of operations computing nodes have to perform, and of-the-main-diagonal elements representing number of messages between nodes. It is important to mention here that more clever approach would be to bring additional functionality to SLEM, and let it create this matrix based on the direct observation of operations a certain HPC algorithm is performing. As a next step, SLEM computes data-flow as a Hadamard product of Y_s and A matrices:

$$P_n = \begin{bmatrix} eps_1 \cdot a_{11} & \min(bw_{12}, bw_{21}) \cdot a_{12} & . & \min(bw_{1n}, bw_{n1}) . a_{1n} \\ \min(bw_{21}, bw_{12}) \cdot a_{21} & eps_2 \cdot a_{22} & . & \min(bw_{2n}, bw_{n2}) \cdot a_{2n} \\ . & . & . & . \\ \min(bw_{n1}, bw_{1n}) \cdot a_{n1} & \min(bw_{n2}, bw_{2n}) \cdot a_{n2} & . & eps_n \cdot a_{nn} \end{bmatrix} \tag{8}$$

Matrix P_n is symmetric, but not doubly stochastic. A following recursive equation (9) developed in [8] transforms P_n to its doubly stochastic form. Experimental results in [8] show that by doing the following recursion in not more than 100 iterations, an absolute error is less than 0.1%.

$$P_{n+1} = (D_n)^{-1/2} P_n (D_n)^{-1/2} \quad ; n = 1,...,\infty \tag{9}$$

Where:

$$D_n = \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & d_n \end{bmatrix} \tag{10}$$

Where each element on the main diagonal D_n is a sum of the corresponding row of the P_n:

$$d_i = \sum_{j=1}^{n} P_{ij} \quad ;\ i = 1,...,n \tag{11}$$

We run SLEM on a master node with Scilab mathematical package installed [14], which has a program for semidefinite optimization. After optimization, one needs to obtain the deterministic form of matrix P_n.

Let us with Q_n denote the optimized version of the matrix P_n. Then, according to [8], we can transform Q_n to its deterministic form with:

$$Q_{n-1} = (D_n)^{1/2} Q_n (D_n)^{1/2} \quad ; n = \infty,...,1 \tag{12}$$

Obviously, the number of iterations needs to be limited. Experimental results in [8] show that within 100 iterations error drops below 0.1%. At the end of the recursive process in (12), Q matrix contains an optimal data-flow for a given system. In order to obtain an optimal workload distribution on the processing nodes, one has to convert an optimal data-flow in the number of operations computing nodes have to perform. One way to do this is to use trace of the matrix, representing sum of the matrix main-diagonal elements. Then, optimal workload distribution A_o is:

$$A_0 = \frac{trace(A)}{trace(Q)} Q \tag{13}$$

As mentioned already, performance of HPL is sensitive to memory: the more memory used with HPL, the better performance values are obtainable. On the other size, we experienced a huge degradation of performance on nodes when there is too much HPL data in memory, and system starts to page out blocks from RAM to spindles. Therefore, we did not use directly main-diagonal elements from (13) as input parameters for heterogeneous HPL directly, but first resized them with correlation factors computed from (1).

IV. RESULTS

Rmax to Rpeak ratio is of the outmost importance when ranking system HPC for performance. It expresses how efficiently certain cluster executes HPL. As already explained, Rmax is a number of double-floating point operations cluster achieved during HPL run, and Rpeak is a maximal theoretical value of double-floating operations cluster could achieve. If successfully finished, both standard and heterogeneous HPL fired Rmax value. Equation used on [2] to compute Rpeak values of homogeneous HPC clusters is:

$$Rpeak = N \times cores \times freq \times ops \tag{14}$$

Where *N* is a number of computational nodes in a cluster, *cores* provides number of processor cores per node, *freq* represents core frequency, and *ops* is a number of concurrent double floating operations a CPU core is theoretically able to execute in one cycle. For heterogeneous clusters, one can compute Rpeak for each server in a cluster, so Rpeak value for a cluster is a sum of Rpeak values for servers.

Server model	cores	freq	N	Ops	Rpeak
Sun Fire X4600 M2	32	2.7	8	2	1382.4
Dell PowerEdge	24	2.66	4	4	1021.4
HP ProLiant DL585 G1	8	2.2	16	2	563.2
Sun Fire X4600	16	2.6	4	2	665.6
HP ProLiant DL580 G5	16	2.4	16	4	2457.6

Table 3: Rpeak values for hpc1 cluster

Server model	cores	freq	N	ops	Rpeak
HP ProLiant SL230 Gen8	16	2.4	28	8	8601.6
Dell PowerEdge	24	2.66	4	4	1021.4

Table 4: Rpeak values for hpc2 cluster

By merging data from tables 3 and 4 together, and with Rmax obtained from standard, heterogeneous and optimized HPL runs, we can express HPL efficiencies for both clusters:

	Rpeak	Rmax	efficiency
hpc1 cluster	6090.24		
Standard HPL		2162	35.50
Heterogeneous HPL		2409	39.55
Optimized HPL		4200	68.96
hpc2 cluster	9623.04		
Standard HPL		1300	13.51
Heterogeneous HPL		4207	43.72
Optimized HPL		4354	45.25

Table 5: HPL efficiency on heterogeneous clusters

V. CONCLUSION

Series of runs of standard HPL benchmarks on heterogeneous HPC clusters showed not so good efficiency. There are two reasons for this; standard HPL structure does not allow division of a problem in smaller parts, and heterogeneous cluster were of computing nodes with significantly different performance capabilities. Consecutively, the highest Rmax we were able to obtain by running standard HPL was 2162 GFLOPS. By using the heterogeneous version of HPL, and dividing and mapping the problem on computing nodes according to their estimated performance, again we were able to obtain improvement of HPL efficiency: for cluster hpc1, the largest Rmax found was 2409 GFLOPS. At the end, we run heterogeneous HPL on cluster again, but this time letting SLEM to find optimal workload balancing.

The improvement of Rmax was huge, up to 4200 GFLOPS. We started the whole benchmarking process on hpc2 cluster in the same way, and standard HPL compared even more poorly against heterogeneous HPL. Rmax values found were 1300, 4207 and 4354 GFLOPS, with standard HPL, and heterogeneous HPL manually and SLEM optimized. We assume that big performance degradation of standard HPL on hpc2 cluster was due to the communication between nodes: As stated before, connection between Dell Poweredge nodes and the rest of the cluster was via SDR Infiniband links, while SL250 communicated with each other via QDR Infiniband links. This, together with the heterogeneity of computing nodes, brings further imbalance and degradation to standard HPL. Besides, all three benchmarks on hpc2 cluster finished with lower efficiency than on hpc1 cluster. It is still under investigation what could be the cause for this. As both clusters were with relatively small amount of nodes, we were able to obtain good results with heterogeneous HPL even with manual optimization. As the number of cluster nodes increase, this method can quickly become tedious and time consuming. Especially in Cloud, were also the communication speed between nodes can vary by an order of magnitude, it could be even impossible to apply manual optimization efficiently. Therefore, and according to the promising results gained with SLEM optimization of HPL in heterogeneous clusters, we hope one can use it efficiently in HPC Cloud systems as well. Improvement of SLEM is possible in various ways. For example, one can relatively easily write and add to it functionality that is able to automatically discover algorithm topology and build corresponding matrix. Further, it is possible to rewrite optimization code in a way it uses many nodes concurrently, instead only one. This could be of the outmost importance when running HPL on very large clusters; then, it could happened that semidefinite optimization on only one master node becomes bottleneck for successful HPL benchmarking. Besides, we hope that in the future, SLEM optimization will be not limited to HPL only: As more and more HPC applications will be Cloud aware, the need for efficient workload will increase. Eventually, this could be a chance for SLEM to become a tool for efficient workload balancing of HPC applications other than HPL.

V. Acknowledgment

We would like to thank David Rohr, from Frankfurt Institute for Advanced Studies at the Goethe University Frankfurt, whose advices were of the crucial importance for the smooth execution of rewritten version of HPL that efficiently performs on heterogeneous clusters. We would also like to thank Dobriša Dobrenić, head of Sector for Computer System at University of Zagreb University Computing Centre, who initiated the idea of HPL benchmarking of Isabella cluster and steadily supported it until its successful end.

References

[1] J. Dongarra, P. Luszczek, A. Petitet, "The linpack benchmark: Past, present and future." Concurrency and Computation: Practice and Experience, 2003, 15(9): 803-820.

[2] http://www.top500.org

[3] T. Endo, A. Nukada, S. Matsuoka, N. Maruyama, "Linpack evaluation on a supercomputer with heterogeneous accelerators", 2010 IEEE International Symposium on Parallel & Distributed Processing, 19-23 April 2010., Atlanta GA, pp. 1-8.

[4] F. Wang, C.Q. Yang, Y.F. Du, Y. Chen, H.Z. Yi, W.X. Xu, "Optimizing Linpack Benchmark on GPU-Accelerated Petascale Supercomputer", Journal of Computer Science and Technology, Vol. 26., No. 5. (1 September 2011), pp. 854-865.

[5] D. Rohr, M. Kretz, M. Bach, "DGEMM and HPL", technical report, Frankfurt Institute for Advanced Systems, University of Frankfurt, Germany, December 9th, 2010.

[7] D. Tomic, "Spectral performance evaluation of parallel processing systems", J. Chaos, Solitons and Fractals, 13 (1) (2002) 25-38.

[8] D. Tomic, O. Muftic, "Distributing workload in parallel computing environment with SLEM", internal communication, 2006-2010.

[9] S. Boyd, P. Diaconis, L. Xiao, "Fastest Mixing Markov Chains", SIAM Review, 46(4), pp. 667 – 689, December 2004.

[10] D. Tomic, O. Muftic, B. Biocic, "A novel scheduling approach of e-learning content on cloud computing infrastructure", 2011. Proceedings of the 34th International Convention MIPRO, Opatija, pp. 1443 – 1446.

[11] D. Tomic, D. Ogrizovic, "Running High Performance Linpack on CPUGPU clusters", 2012. Proceedings of the 35th International Convention MIPRO, Opatija, pp. 400 – 404.

[12] http://www.spec.org

[13] http://www.netlib.org/utk/people/JackDongarra/faq-linpack.html

[14] http://www.scilab.org/

[15] http://www.netlib.org/benchmark/hpl/

[16] http://code.compeng.uni-frankfurt.de/projects/hpl

Distributed Virtual Test Bed: an Approach to Intergation of CAE Systems in UNICORE Grid Environment

G.Radchenko and E. Hudyakova
South Ural State University, Chelyabinsk, Russia
radchenko@acm.org

Abstract - Computer-Aided Engineering (CAE) systems demand a vast amount of computing resources to simulate modern hi-tech products. In this paper, we consider the problem-oriented approach to access remote distributed supercomputer resources using the concept of distributed virtual test bed (DiVTB). DiVTB provides a problem-oriented user interface to distributed computing resources within the grid, online launch of CAE simulation, automated search, monitoring and allocation of computing resources for carrying out the virtual experiments. To support the concept of DiVTB the DiVTB technology was developed. DiVTB technology provides a solution for the development and deployment of DiVTB, integration of most common CAE systems into the distributed computing environment as grid services (based on the UNICORE grid middleware) and web access to CAE simulation process.

I. INTRODUCTION

Computer-Aided Engineering systems are now one of the key factors ensuring competitiveness of any high-tech production. The use of such systems makes it possible to conduct virtual experiments when real-life experiments are difficult or impossible. This can greatly improve the accuracy of case analysis and design decisions and shorten the way from the generation of an idea to the real industrial production [7].

The accuracy of the computer simulation depends on the level of detail of meshes used for computational experiments. The computational complexity of problems of engineering analysis keeps increasing, and their solutions require significant computing resources. The only way is to use the multiprocessor systems.

For an ordinary user the process of solution of engineering problem involving supercomputing resources is associated with certain difficulties. First, the interface and features of all software components that provide the simulation technological cycle need to be explored [2]. Second, specific knowledge and skills in the high performance computing are required. And finally yet importantly, the cost of the purchase, installation and administration of powerful supercomputer equipment and software is extremely high. All these factors complicate the process of embedding of powerful engineering methods in research and development in traditional manufacturing companies.

An alternative to creating your own supercomputer center is renting remote computing resources from the supercomputer centers that offer utility computing services. The utility computing assumes that diverse computational resources can be brought together on demand and that computations can be implemented depending on demand and service load [6]. However, there is a range of issues related to the mechanism of remote resources usage. Issues of data exchange security, transparency of allocation and use of remote resources can deter industrial users from such approach. These issues can be resolved by the grid computing concept [4] together with cloud computing concept [5], whereby the end user has a problem-oriented service that provides problem solution based on distributed computing resources.

In the past few years, the approach to supplying the CAE systems in the form of distributed computing environment resources is booming. In [3, 11] development of collaborative design system for engineering analysis by dynamically distributed development team is introduced. As members of such groups need to interact constantly during the development process, peer-to-peer computing was selected as the basic platform for this system. However, the use of this approach is not aimed at providing the end user with a finalized product accessing distributed network resources, but rather at supporting the collaborative design process for a development team.

In [9] a concept of the Virtual Test Bed (VTB) (a prototype of the virtual cooperative engineering environment) was introduced. The VTB has been designed as an architecture to facilitate the integrated execution of different simulation programs with other supported non-simulation software. The proposed solution is built on the Extensible Modeling and Simulation Framework (XMSF) platform that supports such distributed computing approaches as CORBA, SOAP, DCOM and is highly targeted at the specific field of application (the development of new generation spacecrafts).

In [1] the REST web-services architecture for distributed simulation is introduced. The proposed approach supports interoperability of collaboration of

The reported study was partially supported by RFBR, research project No. 11-07-00478-a and by the Ministry of Education and Science of Russia, task No. 8.3786.2011.

distributed computing resources and robust data exchange during simulation process.

The analysis of virtual test beds for CAE systems shows:

- the lack of a unified approach for creating virtual test beds allowing CAE simulation in the distributed computing environment on the basis of standard grid protocols;
- the lack of adaptation of the VTB problem-oriented user interface for the needs of the certain categories of users and the lack of streamlined web interface for inputting the tasks;
- lack of a problem-oriented resource broker, which would transparently provide resources of distributed computing environment (including hardware, software and license resources) in compliance with the current task requirements.

In [8] the concept of GridBean in GPE grid middleware platform was introduced. A GridBean is an object responsible for generating the job description for grid services and providing graphical user interface for input and output data. This platform introduced an easy way to integrate legacy stand-alone applications in the grid environment as grid services. Subsequently the GridBean concept was integrated into the UNICORE middleware platform [10].

In this paper, we provide a review of the Distributed Virtual Test Bed (DiVTB) concept that allows providing resources of CAE systems in the grid computing environment. Moreover, we present the DiVTB Technology — a combination of concepts and techniques used for the development and operation of DiVTB. The UNICORE grid middleware is used to implement grid services that integrate CAE systems in grid environment.

II. The Distributed Virtual Test Bed

A. Basic terms

Let a CAE-problem be assumed as the combination of:

- a geometric model of a research target and/or computational mesh which divides a simulated area into discrete sub-areas;
- boundary conditions, physical characteristics and parameters of the components interaction in a simulated area;
- a description of unknown variables which values will be obtained as a result;
- the requirements for hardware and software resources that provide a simulation process.

Let a CAE-parameter be the value, which affects the result of simulation and can vary within a certain range. Each parameter has specific semantics and describes a feature of a subject area or a solution process.

We offer a concept of a Distributed Virtual Test Bed (DiVTB) to provide access to resources of CAE systems in a distributed computing environment. DiVTB is a software solution that provides an engineering simulation in the grid computing environment for a certain CAE-problem. DiVTB includes an interface for CAE-parameters values input; a set of software tools enabling the use of grid resources for virtual experiment (*a driver*) and a set of grid services that provide the solution of a CAE-problem and implement safe standardized remote communication methods.

A combination of concepts and techniques for DiVTB development and operation is defined as *the DiVTB Technology*.

The DiVTB Technology includes:

- the DiVTB architecture and development techniques;
- the administrative concepts: the patterns of work and the
- distribution of duties between the DiVTB developers and users;
- the software solution for the DiVTB development, deployment and operation.

B. The structure of a DiVTB

CAEBean is a primary structural unit of a DiVTB. According to the DiVTB Technology, we distinguish four structural levels of a DiVTB, each represented by its CAEBean type (Figure 1):

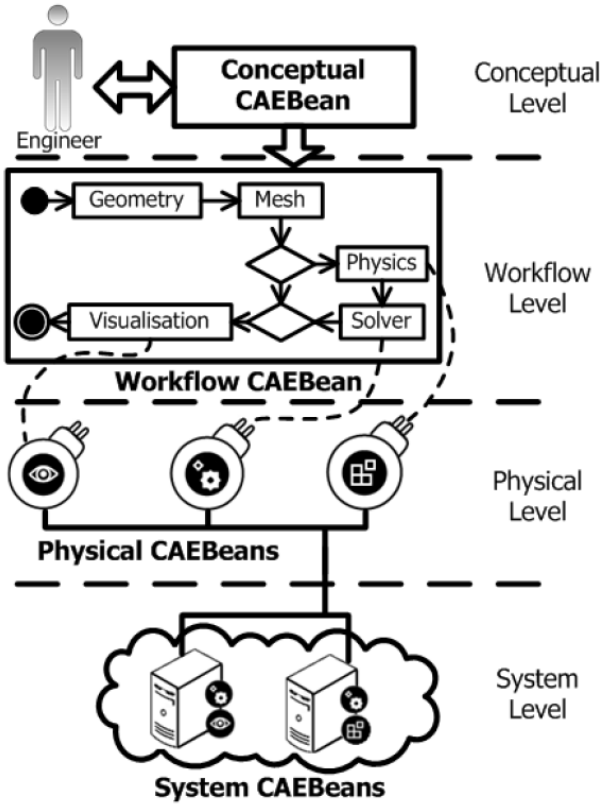


Figure 1. A generalized scheme of DiVTB levels according to the DiVTB Technnology

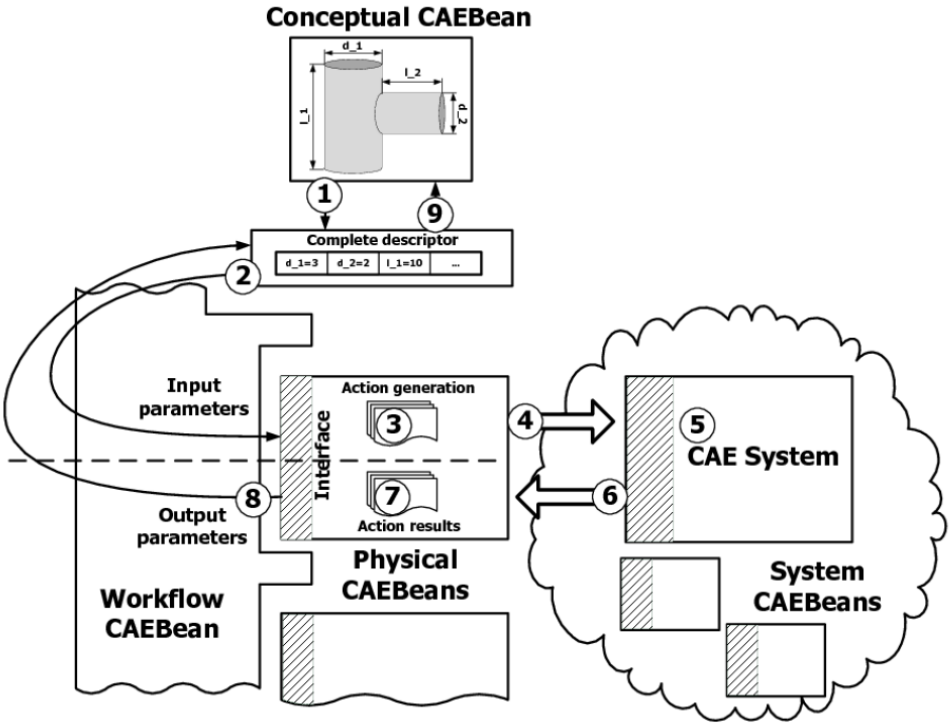


Figure 2. Simulation of a distributed virtual test bed

- Conceptual level (a hierarchy of *conceptual CAEBeans*);
- Workflow level (a *workflow CAEBean*);
- Physical level (consists of a set of *physical CAEBeans*);
- System level (consists of a set of *system CAEBeans*).

A conceptual level of DiVTB is based on conceptual CAEBeans. A conceptual CAEBean defines the *problem-oriented interface* for a DiVTB based on a set of its CAE-parameters. Through a conceptual CAEBean a user can launch the DiVTB simulation, follow a solution process and obtain the results. The DiVTB Technology provides the opportunity to design a *generalization hierarchy* of conceptual CAEBeans. The conceptual CAEBeans of lower hierarchy levels adapt a problem-oriented interface of the DiVTB for the needs of a certain category of users. These CAEBeans specify the CAE-problem defined by a parent conceptual CAEBean through assigning specified values to some CAE-parameters.

The workflow level of DiVTB is represented by a workflow CAEBean implementing a CAE-workflow for a certain CAE-problem. Let a CAE-workflow be defined as a directed graph with the following types of nodes:

- an action performed by a certain CAE system;
- a node controlling the problem solution flow.

We use the elements from the UML 2.0 activity diagram notation to form a CAE-workflow. The information required to initiate any node of a workflow is stored in a *complete descriptor* of DiVTB. Let a complete descriptor be defined as a set of CAE-parameters that uniquely describes a CAE-problem, which can be simulated by DiVTB. Any workflow node as a part of a solution can refer to a *complete descriptor* for input values and store results in the corresponding output parameters.

A physical level of DiVTB is represented by the physical CAEBeans. The main function of a physical CAEBean is a conversion of a problem-oriented description of a CAE-problem (a set of parameters in complete descriptor) into the set of files needed to launch a single *CAE-action* (mesh generation, simulation, post processing, optimization etc.) on a specific CAE system. At the end of a problem-solving process a physical CAEBean provides conversion of component-oriented results in a problem-oriented form. During the CAE-workflow execution, each physical CAEBean must be matched with a grid service to execute a CAE-action. All grid services are implemented through the DiVTB System. A workflow CAEBean together with the corresponding physical CAEBeans implement *a driver* of DiVTB.

A system CAEBean provides the functionality of a computing resource in a distributed computing environment and implements a service-oriented approach to task input and fetching the results. A system CAEBean provides the isolated workspace for each CAE-action and the programming interface that allows initial data loading, remote task input and delivery of results.

C. DiVTB Simulation

Let us consider the process of DiVTB simulation according to the DiVTB Technology (Figure 2).

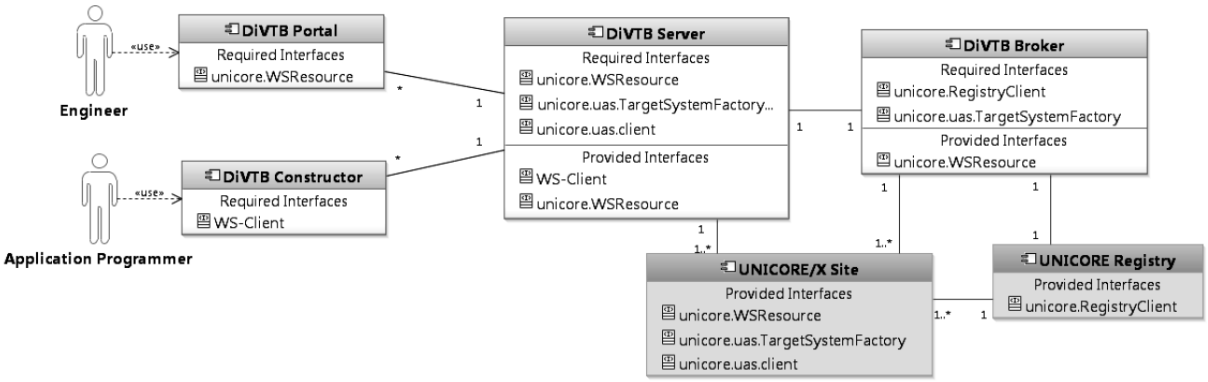


Figure 3. General diagram of DiVTB System components

1. An engineer inputs the values of CAE-parameters for DiVTB by means of user interface provided by an appropriate conceptual CAEBean. The conceptual CAEBean calculates values of CAE-parameters, which are not available to the user on the current level of generalization hierarchy. All initial CAE parameter values are stored in the complete descriptor of DiVTB. Finally, the engineer launches the simulation process.
2. The workflow CAEBean executes the CAE-actions according to the CAE-workflow. The values of initial parameters of any action (to be defined before the action starts) are taken from the complete descriptor of DiVTB and transferred to the physical CAEBean.
3. The physical CAEBean generates a set of input files for the CAE system wrapped in the corresponding system CAEBean.
4. The resource broker gives an address of the CAEBean system with the best opportunity to execute the CAE-action. The physical CAEBean transfers a set of input files to the allocated CAEBean system and starts the CAE-action.
5. The CAE system executes the CAE-action and generates a set of result files.
6. The Physical CAEBean receives result files from the system CAEBean.
7. The physical CAEBean parses the results of the action and retrieves the values of required output parameters.
8. The physical CAEBean puts the obtained values of output parameters into the complete descriptor. The workflow CAEBean triggers the next node of the CAE-workflow (step 2).
9. When the CAE-workflow is complete, all results of the DiVTB simulation are transferred to the conceptual CAEBean from the complete descriptor.

III. The DiVTB System

To implement the DiVTB Technology we created a DiVTB System — a software solution for DiVTB development and implementation. We selected the UNICORE 6 grid computing middleware [10] as a platform for implementation of grid services. The GridBeans approach included in UNICORE 6 supports the transparent integration of legacy standalone applications as grid services in the grid environment. The DiVTB System consists of the following components (Figure 3):

1. *DiVTB Constructor* — integrated development environment of DiVTB. It provides the interface for an application programmer to develop the CAEBeans of conceptual, workflow and physical levels. Accordingly, the DiVTB Constructor user interface is divided into three sections responsible for developing corresponding types of CAEBeans.
2. *DiVTB Portal* — a web application that provides a user interface for the input and solution of CAE-problems by means of an appropriate DiVTB. The web-form generator automatically builds DiVTB user interfaces based on the corresponding conceptual CAEBeans. In addition, the DiVTB Portal supports authentication and user account management.
3. *DiVTB Server* — a grid service for DiVTB storage and simulation. A process of execution of a CAE-workflow is supported by the internal workflow subsystem of the DiVTB Server. DiVTB Server provides interfaces for DiVTB Constructor and DiVTB Portal allowing to upload a DiVTB to a Server and to start a simulation of a DiVTB.
4. *DiVTB Broker* — an automated system for registration, analysis and allocation of CAE-resources for CAE-action execution. The DiVTB Broker receives requests for resources from DiVTB Server, analyzes the status of the distributed computing environment and supplies the DiVTB Server with a CAE-resource best suited to the request. DiVTB Broker interacts with an UNICORE Registry service to gain an information about resources available in the grid environment.

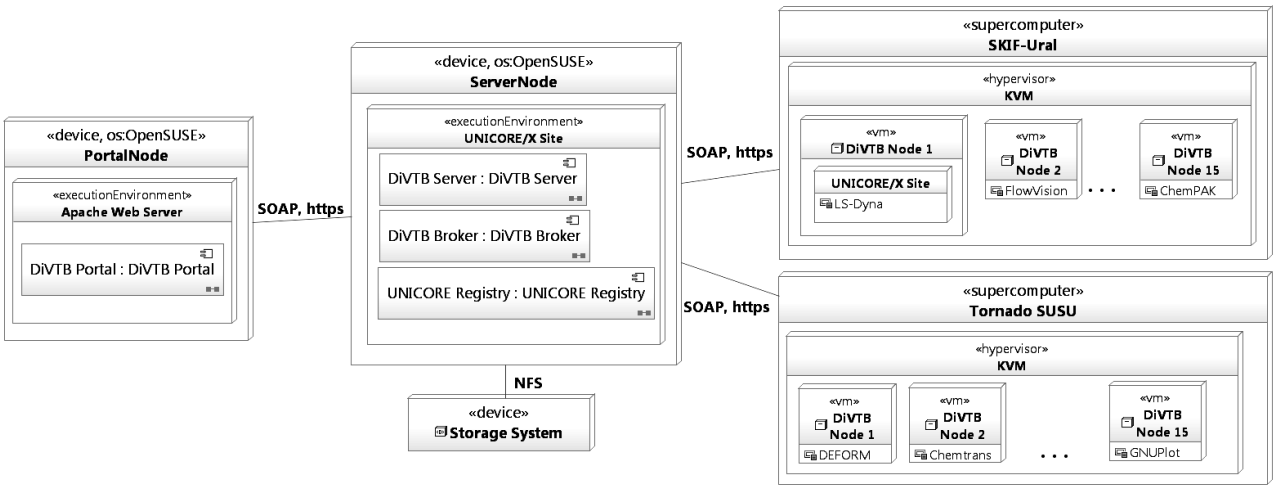


Figure 4. DiVTB System deployment diagram

5. *CAE-Resources* — grid services that implement the remote execution of CAE-Actions by means of CAE-systems installed on specific computers in a grid (instances of system CAEBeans). Each CAE-resource is a UNICORE/X site with a set of problem-oriented and system software installed on it. CAE-resource provides:
   - obtaining problem data from DiVTB Server or an external data source;
   - CAE-action initiation and automated solution;
   - transmission of results into the DiVTB Server or external data storage.

IV. DiVTB System Test

To test the DiVTB System we developed a set of DiVTB to solve problems by using common CAE systems: ANSYS CFX, ANSYS Mechanical, ABAQUS, DEFORM, LS-DYNA. In addition, a set of self-written and open-source CAE components was integrated in DiVTB System during "Young Researchers Mobility" research project, which was held at the South Ural State University in autumn 2012. DiVTB System was deployed on the supercomputer resources of South Ural State University Supercomputer Simulation Laboratory (Figure 4).

For the test we chose the issue of tempering and cooling of tubes and analysis of impact of various aspects of the quenching process on the quality of the production (*"Thermal Treatment Distributed Virtual Test Bed"*) based on DEFORM system.

Firstly, we created a model of a heat treatment process. We supplied thermal studies of the pipe hardening. We collected information about temperature fields during the quenching process and the magnitude of the initial circumference and curvature along the axis of the work piece. The "Thermal Treatment" DiVTB allows to change the flowing parameters of the heat treatment process: number of inductors, frequency and current strength, a length of the inductors, number and configuration of the jets of water, pressure of cooling flow, velocity of pipe motion, pipe steel grades (Figure 5).

Another application of DiVTB was a virtual test bed for supercomputer simulation of deformation of knitted fabrics on human body. We designed and deployed a DiVTB based on LS-DYNA CAE system that allow simulating the process of putting on knitted clothes to a torso of a human body. By varying parameters, such as density and method of weaving of knitted material, engineer can analyze the quality of the fit of the final product; view a video of the process of dressing, etc.

Testing revealed that end-users easily adapt to the solution of applied problems on remote computer infrustructures through a DiVTB System. Web-based user interface provides seamless access to grid resources without the need for "classical" techniques for working with remote and supercomputer resources, such as SSH console access to and SFTP. Advanced users also prefer to develop a simplified DiVTB to organize their everyday use of remote supercomputing resources.

In addition, the tests revealed a number of features that should be considered in the further development of DiVTB System.

First, communication between developers and end-users of DiVTB System often held directly at the workplace of the customer. In this case, the use of a DiVTB Constructor desktop application complicates the development process and interaction with the customer.

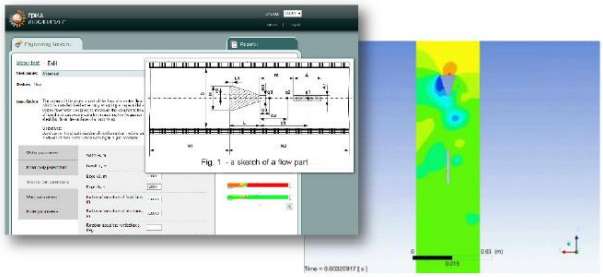


Figure 5. Example of DiVTB user interface (DiVTB Portal)

To solve this problem, we create a Web application that provides DiVTB development.

Secondly, it appears that the most challenging in terms of Application Programmer became a procedure of templating of the problem and result files (Action Generation and Action Results stages on fig. 2). Application Programmer has to create custom stand-alone applications that provide the implementation of these procedures, although in most cases he can use rather simple text generators based on certain templates. To solve this problem, we are planning to develop a DiVTB Constructor module that will provide template generation and customization of input data.

In addition, users asked us to provide pre-visualization of simulation results before loading it on a user's machine. Since the simulation results can be as big as hundreds of gigabytes, downloading this amount of data from a remote HPC system to a user's machine can take a very long period. If the results do not meet user's expectations, he would like to know about it in advance. A possible solution to this problem is to provide a remote interactive visualization of calculation results, so user can visualize simulation data on the site of its calculation. To provide this capability, we are developing a remote interactive visualization service that would provide such capabilities.

V. CONCLUSIONS

This paper presents the main aspects of CAE systems integration in the distributed computing environment. We have offered the concept of distributed virtual test bed (DiVTB) providing a problem-oriented approach for CAE-problems simulation by means of grid computing resources. We have developed a DiVTB Technology allowing development and operation of distributed virtual test beds. Based on this technology we have developed a DiVTB System and deployed virtual test beds aimed at solving CAE-problems by means of the most common CAE systems (ANSYS CFX, ANSYS Mechanical, ABAQUS, DEFORM, LS-DYNA).

REFERENCES

[1] K. Al-Zoubi and G. Wainer, "Using REST web-services architecture for distributed simulation," 2009 ACM/IEEE/SCS 23rd Workshop on Principles of Advanced and Distributed Simulation, PADS '09, Washington, DC, USA, pp 114–121, June 2009.

[2] T. M. Buriol and S. Scheer, "CAD and CAE integration through scientific visualization techniques for illumination design," Tsinghua Science & Technology, vol. 13, no. 1, pp. 26 – 33, October 2008.

[3] L. Q. Fan, A. Senthil Kumar, B. N. Jagdish, and S. H. Bok, "Development of a distributed collaborative design framework within peer-to-peer environment," Comput. Aided Des., vol. 40, no. 9, p. 891–904, Sept. 2008.

[4] I. Foster and C. Kesselman, The Grid 2: Blueprint for a New Computing Infrastructure. The Morgan Kaufmann Series in Computer Architecture and Design Series. Elsevier Science, 2003.

[5] B. Hayes, "Cloud computing," Commun. ACM, vol 51, no 7, pp. 9–11, July 2008.

[6] M. Huhns and M. Singh, "Service-oriented computing: key concepts and principles," Internet Computing, IEEE, vol. 9, no. 1, pp. 75 – 81, Feb. 2005.

[7] B. Raphael and I. Smith, Fundamentals of Computer-Aided Engineering. John Wiley & Sons, 2003.

[8] R. Ratering, A. Lukichev, M. Riedel, et al. "GridBeans: support e-science and grid applications," In Proceedings of the Second IEEE International Conference on e-Science and Grid Computing, E-SCIENCE '06, p. 45, Washington, DC, USA, 2006. IEEE Computer Society.

[9] J. Sep_ulveda, L. Rabelo, J. Park, F. Riddick, and C. Peaden, "Implementing the high level architecture in the virtual test bed," proceedings of the 36th conference on Winter simulation, pp. 1444-1451, 2004.

[10] A. Streit, P. Bala, A. Beck-Ratzka, et al, "Unicore 6 - recent and future advancements," Annals of Telecommunications, vol 65, pp. 757-762, 2010.

[11] J. Yin, W. Zhang, Y. Li, and H. Chen, "A peer-to-peer-based multi-agent framework for decentralized grid workow management in collaborative design," The International Journal of Advanced Manufacturing Technology, vol. 41, pp. 407-420, 2009.

Message Oriented Framework with Low Overhead for Efficient High-Performance Monte Carlo Simulations

E. Atanassov, T. Gurov, A. Karaivanova
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences (IICT-BAS)
1113 Sofia, Bulgaria
Emails: (emanouil,gurov,anet)@parallel.bas.bg

***Abstract*—In the recent years Bulgaria acquired a substantial amount of HPC resources of various types. The biggest procurement has been the BlueGene/P supercomputer at SAITC with 8192 CPU cores, while the Bulgarian Academy of Sciences has now two HPC clusters with Intel CPUs and Infiniband interconnection, which total more than 1000 logical cores. In addition some servers equipped with powerful GPU are available for applications that can take advantage of them. The coordinated use of such resources by one application faces significant challenges due to the heterogeneity of the resources and the networking and security constraints.**

In order to facilitate the coordinated use of all these resources where each resource is used for the parts of the application where it is most efficient, we have developed a framework that allows the researcher to interconnect resources of the above types with minimal overhead. In this paper we describe the architecture of the system and demonstrate its effectiveness for a semiconductor modeling application, showing numerical and timing results.

***Index Terms*—component; formatting; style; styling; Electron Transport, Monte Carlo simulation, Message Oriented Framework, HPC resources.**

I. Introduction

Monte Carlo applications are usually large problems demanding significant computational resources. The current situation shows significant imbalance between computational power and I/O bandwidth of current leadership-class machines. A researcher may have access to several different HPC resources at the same time. In the same time, porting an application to a new platform can be challenging.

Let us remind that the message-oriented middleware, a fast developing area (for example, [1], [2], [3], [4], [5]), is a specific class of middleware that supports the exchange of general-purpose messages in a distributed application environment with the purpose to facilitate the communication between distributed applications.

In this work we discuss some points that should be taken into consideration when porting applications to HPC recourses in Bulgaria. We suggest Message Oriented Framework with Low Overhead for efficient use of the HPC recourses available in Bulgaria.

II. Interconnection of the Existing HPC Platforms in Bulgaria

There are three types of HPC resources in Bulgaria:

- The biggest HPC resource for research in Bulgaria is the supercomputer - BlueGene/P (vendor: IBM) which is located in the building of the Executive Agency Electronic Communication Networks and Information Systems.
  The characteristics of the supercomputer IBM Blue Gene/P in Sofia are:
  - two racks, 2048 PowerPC 450 processors (32 bits, 850 MHz), a total of 8192 cores;
  - A total of 4 TB random access memory;
  - 16 I/O nodes currently connected via fiber optics to a 10 Gb/s Ethernet switch;
  - Theoretical peak performance: Rpeak= 27.85 Tflops;
  - Energy efficiency: 371.67 MFlops/W
  - 1 Gb/s Ethernet fiber optics link to Bulgarian NRENs Point-of-Presence at the IICT-BAS
  - Operating System for front-end node: SUSE Linux Enterprise Server 10 (SLES 10), Service Pack 1 (BG/P)
  - The Compute Nodes run OS Compute Node Kernel (CNK)
  - 2 file servers, 12 TB stora

  On the BG/P five networks are used for various tasks.
  1) Three-dimensional torus: for general-purpose, point-to-point message passing. A fully connected torus is only available when using jobs which occupy one or two complete midplanes, i.e. half or all of the system. When using smaller partitions of the machine, the networks topology is degraded to a 2 or 3D mesh.
  2) Global collective network is a high-bandwidth, tree based network that is used for collective communication operations, such as broadcasts and reductions, and to move process and application data from the I/O nodes to the compute nodes.
  3) Global interrupt network
  4) 10 Gigabit Ethernet network uses optical cabling

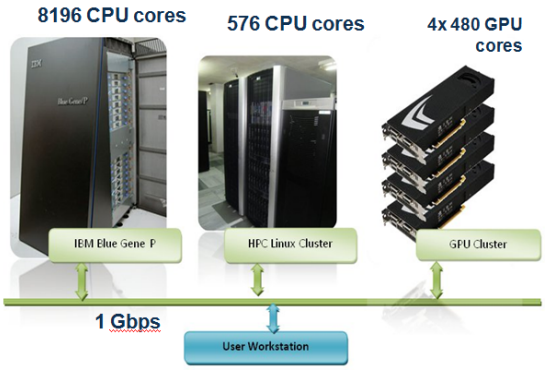


Fig. 1. Interconnection of existing HPC resources in Bulgaria for research and education.

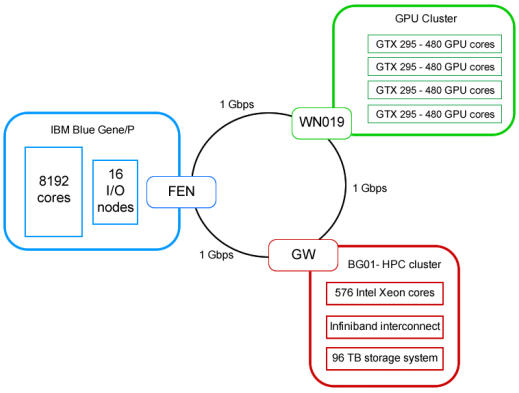


Fig. 2. Interconnection of existing HPC resources in Bulgaria for research and education.

to interconnect all the I/O nodes and the storage infrastructure.

5) Control: The control network consists of a JTAG interface with direct access to shared SRAM in every compute and I/O node. It is used for boot, monitoring, and diagnostics.

- Two HPC clusters with Intel CPUs and Infiniband interconnection located in the Institute of Information and Communication Technologies - BAS (vendor HP) and in the Institute of Organic Chemistry with Center of Phitochemistry - BAS (vendor Fujitsu)
  The HPC cluster at IICT-BAS includes:
  - HP Cluster Platform Express 7000 enclosures with 36 blades BL 280c with dual Intel Xeon X5560 @ 2.8Ghz (total 576 cores), 24 GB RAM per blade
  - 8 controlling nodes HP DL 380 G6 with dual Intel X5560 @ 2.8 Ghz, 32 GB RAM
  - Non-blocking DDR Interconnection via Voltaire Grid director 2004
  - Two SAN switches for redundant access
  - MSA2312fc with 96 TB storage, Lustre filesystem.
  - More than 92% efficiency on LINPACK (¿3 TFlops, peak performance 3.2TFlops)
  - SL 5, gLite, torque+maui, eucalyptus walrus
- In addition servers equipped with powerful GPU are available for applications that can take advantage of them. Connected to the HPC cluster at IICT-BAS extension cluster with 4 GPU cards NVIDIA GTX 295 (4x2x240 GPU cores =1920 cores, 576 MHz, 1.8 GB per card)

These resources are connected with 1 Gb/s Ethernet fiber optics links.

The Figure 1 shows the interconnection of existing HPC platforms in Bulgaria:

1) Front-end Nodes (fen, GW, WN019) provide access to the users to develop, build and submit
2) Compute Nodes run applications
3) I/O Nodes provides access to external devices. All I/O requests are routed through these nodes. (BG/P))
4) All I/O calls in the applications are forwarded to the I/O nodes.
5) High bandwidth networks for end-user applications use, i.e. point to point & collective communications.

In order to facilitate the coordinated use of all these resources where each resource is used for the parts of the application where it is most efficient, we have developed a framework that allows the researcher to interconnect resources of the above types with minimal overhead.

III. Case Study - Monte Carlo Simulation of Electron Transport in Semiconductors

We have chosen as case study the Monte Carlo simulation of electron transport in semiconductors. We consider the quantum kinetic equation describing a electron-phonon interactions in presence of applied electric field [9]. This equation can be written in the following integral form [9]:

$$f(\mathbf{k},t) = \phi(\mathbf{k}) + \int_0^t dt'' \int_G d\mathbf{k}' K(\mathbf{k},\mathbf{k}') \times \quad (1)$$
$$\left\{ \int_{t''}^t dt' S_1(\mathbf{k},\mathbf{k}',\mathbf{F},t',t'') f(\mathbf{k}',t'') + \right.$$
$$\left. \int_{t''}^t dt' S_2(\mathbf{k},\mathbf{k}',\mathbf{F},t',t'') f(\mathbf{k},t'') \right\},$$

where the kernel is separated in two terms:

$$K(\mathbf{k},\mathbf{k}') = \frac{2V}{2\pi^3\hbar^2}|g(\mathbf{q})|^2, \quad (2)$$

and

$$S_1(\mathbf{k},\mathbf{k}',\mathbf{F},t',t'') = -S_2(\mathbf{k}',\mathbf{k},\mathbf{F},t',t'') = \quad (3)$$
$$\left[(n_\mathbf{q}+1)\cos\left(\frac{\varepsilon(\mathbf{k}) - \varepsilon(\mathbf{k}') + \hbar\omega_\mathbf{q}}{\hbar}(t'-t'') - \right.\right.$$
$$\left. \frac{\hbar}{2m}(\mathbf{k}'-\mathbf{k})\cdot\mathbf{F}(t'^2 - t''^2) \right)$$
$$+ n_\mathbf{q}\cos\left(\frac{\varepsilon(\mathbf{k}) - \varepsilon(\mathbf{k}') - \hbar\omega_\mathbf{q}}{\hbar}(t'-t'') - \right.$$
$$\left.\left. \frac{\hbar}{2m}(\mathbf{k}'-\mathbf{k})\cdot\mathbf{F}(t'^2 - t''^2) \right) \right] \times exp(-\Gamma(t'-t'')).$$

Here, $\mathbf{k}$ and t are the momentum and the evolution time, respectively. $f(\mathbf{k},t)$ is the distribution function. $\phi(\mathbf{k})$ is the

initial electron distribution function. $\mathbf{F} = e\mathbf{E}/\hbar$, where $\mathbf{E}$ is the applied electric field. $n_{\mathbf{q}} = 1/(exp(\hbar\omega_{\mathbf{q}}/\mathcal{K}T) - 1)$ is the Bose function, where $\mathcal{K}$ is the Boltzmann constant and T is the temperature of the crystal, corresponds to an equilibrium distributed phonon bath. $\hbar\omega_{\mathbf{q}}$ is the phonon energy which generally depends on $\mathbf{q} = \mathbf{k}' - \mathbf{k}$, and $\varepsilon(\mathbf{k}) = (\hbar^2\mathbf{k}^2)/2m$ is the electron energy. A Fröhlich coupling is considered

$$g(\mathbf{q}) = -i\left[\frac{2\pi e^2\hbar\omega_{\mathbf{q}}}{V}\left(\frac{1}{\epsilon_{\infty}} - \frac{1}{\epsilon_s}\right)\frac{1}{(\mathbf{q})^2}\right]^{\frac{1}{2}},$$

where (ϵ_∞) and (ϵ_s) are the optical and static dielectric constants. The damping factor Γ is considered independent of the electron states $\mathbf{k}$ and $\mathbf{k}'$. This is reasonable since Γ weakly depends on $\mathbf{k}$ and $\mathbf{k}'$ for states in the energy region above the phonon threshold, where the majority of the electrons reside due to the action of the electric field. The solution of the quantum kinetic equation (1) is evaluated by a Monte Carlo algorithm suggested in [9].

The numerical tests have been performed on the High Performance cluster deployed at the Institute of Information and Communication Technologies of the Bulgarian Academy of Sciences (IICT-BAS), and on the supercomputer BlueGene/P. The scalability results on 512, 1024 and 2048 cores are shown on Fig 2. The test case uses 10 millions of trajectories to simulation 180 femtosecond evolution.

The application requires accumulation of billions of trajectories in order to achieve reasonable accuracy. Improvements in variance and execution time can be achieved with scrambled low-discrepancy sequences or hybrid sequences. The use of quasirandom numbers requires a robust and flexible implementation, since it is not feasible to ignore failures and missing results of some trajectories, unlike in Monte Carlo. GPU resources are efficient in computations using the low-discrepancy sequences of Sobol, Halton, etc. Our idea is to use exchange of messages for distributing the work between heterogeneous HPC resources independent of underlying middleware. We would like to create a framework that is sufficiently generic to be used by a wide range of applications but is efficient for the case of semiconductor modelling with MC and QMC.

IV. Message Oriented Framework

There are two patterns for exchange of information: short messages, encoding parameters and results (RPC type), and large amount of data, organized as file and using permanent or temporary disk storage.

Most of HPC computing resources are hidden behind firewalls, communications with external nodes must be forwarded through the gateway nodes, where user-level server program can be started by the user. User starts the server programs at the gateways and launches computational batch jobs. The computational jobs connect with the server and follow the pattern:

- Request work
- Return result

  until receiving marker for end of computation.

The server performs match-making of requests, for example some kinds of requests may be served only by particular resources. One server (wn019) is master, the others forward to it.

This open source middleware offers a truly cost-effective and flexible messaging platform for efficiently scaling operations, reliably executing data, and moving data and connecting processes across heterogeneous database and application environments. Message-oriented middleware provides the organization with

- High-performance data delivery
- Reliable connectivity to remote clients.

V. Conclusions

Message oriented frameworks overcome some deployment limitations like lack of common Grid middleware installed Access to cloud storage provides simple security model (signed http requests) which also offers easier deployment Robustness of the framework should be improved with regards to dealing with failures and timeouts.

Acknowledgment

The research work reported in the paper is partly supported by the project AComIn "Advanced Computing for Innovation", grant 316087, funded by the FP7 Capacity Programme (Research Potential of Convergence Regions), and by the Bulgarian NSF under Grant DVCP02/1 CoE Supper CA++.

References

[1] M. Amoretti, F. Zanichelli, G. Conte, SP2A: a service-oriented framework for P2P-based Grids, ACM New York, 2005, pp. 1-6, ISBN:1-59593-269-0, doi¿10.1145/11011499.1101508

[2] Asvija, B., K. Kalaiselvan, M. Ateet, R. Sridharan, S.R. Krishnamurthy, A performance Based QoS Aware Resource Brokering Framework, Information Technology: New Generations (ITNG), 2012, pp. 508-515.

[3] K. Czajkowski, D. Ferguson, I. Foster, et al., From Open Grid Service Infrastructure to WS-Resource Framework:Refactoring & Evolution, White Paper, 2004, http://www.globustoolkit.org/wsrf/specs/ogsi_to_wsrf_1.0.pdf.

[4] S. Hayashi, S. Teige and R. Quick, The event notification and alarm system for the Open Science Grid operations center, Journal of Physics: Conference Series, Vol. 396, 2012, doi:10.1088/1742-6596/396/3/032105.

[5] E. Souto et al., A message-oriented middleware for sensor networks, ACM New York, 2004, pp.127-134, ISBN: 1-58113-951-9, doi¿10.1145/1028509.1028514.

[6] E. I. Atanassov and S. Ivanovska, Sensitivity Study of Heston Stochastic Volatility Model Using GPGPU, LSSC 2011, LNCS, Springer, 2012, Volume 7116/2012, 439-446, DOI: 10.1007/978-3-642-29843-1_49, ISSN: 0302-9743.

[7] E. Atanassov, T. Gurov, A. Karaivanova, Cloud and Grid Computing:Security Aspects, Proceedings of BGSIAM'11, Demetra Ltd, Sofia, pp.21-25, 2012.

[8] E. Atanassov, et al., SALUTE application for Quantum Transport - New Grid Implementation Scheme, Proceedings of the Spanish conference on e-Science Grid Computing, pp. 23-32, 2007.

[9] M. Nedjalkov , T.V. Gurov, H. Kosina, and P.A. Whitlock, Statistical Algorithms for Simulation of Electron Quantum Kinetics in Semiconductors - Part II, *Lect. Notes in Comp. Sci.*, 2179, Springer, 183-190, 2001.

WRF4SG: A Scientific Gateway for the Weather Research and Forecasting Model

C. Blanco, A. S. Cofiño and V. Fernández-Quiruelas
University of Cantabria/Dep. of Applied Mathematics and Computer Sciences, Santander, Spain
e-mail: josecarlos.blanco@unican.es

Abstract - Numerical climate models such as Weather Research and Forecasting (WRF) are nowadays among the one of the most computationally demanding applications. Climate and weather research experiments performed with models such as WRF are complex tasks because of models' requirements. In order to manage that, we are going to develop a scientific gateway called WRF4SG to cover WRF users needs to manage climate and weather research experiments. The main objective of this work is to establish how WRF4SG is going to assist WRF users in their experiments.

I. INTRODUCTION

Current atmospheric science is at the forefront of advancing knowledge. Major research initiatives in this area require expertise in modeling of physics and chemistry of the atmosphere to study the climate and weather. One of the main tools used for this purpose are climate models. These numerical models simulate the interactions of the atmosphere, oceans, land surface and cryosphere given a region which can be either the whole globe or a limited area.

One of these climate models is Weather Research and Forecasting (WRF) [1] model which is a public domain software. WRF features a flexible and computationally-efficient model for operational forecasting, while offering cutting-edge advances in physics, dynamics, and data assimilation contributed by the research community. In fact, WRF is used, at least, for [2]:

- Case-study research.
- Atmospheric physics/parametrization research.
- Real-time NWP and forecast system research.
- Data assimilation research.
- Regional climate and seasonal time-scale research.
- Chemistry applications (coupled to a chemical transport model).
- Global simulations.
- Idealized simulations at very different spatial and temporal scales (e.g. convection, baroclinic waves, large eddy simulations).

For the reasons above, WRF is employed by academic scientific researchers (dynamics, physics, weather, and climate research), forecast teams at operational centers and application area researchers (e.g. air quality, hydrology, utilities, fires, …), among others.

Despite the fact that WRF is a useful tool in the climate field, when an experiment requires to run more than one simulation its complexity rises almost exponentially. Scientific researchers have to prepare the input data, monitor the execution, manage myriad of files and check the output data for each simulation which supposes to make an enormous effort for a simple climate experiment.

In order to ease the complexity of WRF climate experiments, we are going to developed a Scientific Gateway (SG) based on WS-PGRADE/gUSE [3] and WRF4G [4] frameworks. This gateway, called WRF for Scientific Gateway (WRF4SG), is going to allow WRF users to reach complex climate experiments such as weather forecast, extreme weather case studies, hindcasts and reforecasts proving time-saving features.

The main target of WRF4SG can be divided into two goals. The first goal is to give support for WRF users to access to Distributed Computing Infrastructure (DCI) concurrently due to the fact that WRF demands too many resources to tackle a normal experiment. On the other hand, the second goal is to reduce the sophistication of WRF model in order to undertake more ambitious experiments.

II. RELATED WORK

As already mentioned, configuring and executing WRF is a very complicated task. In fact, WRF is composed of several executables (see Fig. 1) which have to be ran in the right order. One typical WRF execution requires to manage multiple files scattered on various different directories. Moreover, monitoring WRF simulations demands to track some certain log files which have to be checked under specific directories for each executable. Due to the fact that all these tedious tasks consume users' time, WRF users have available different time saving tools to configure and manage WRF experiments such as WRF4G [5], WRF Portal [6] and LEAD Portal [7].

An overview of these WRF tools can be found in [8] that evaluates the main different aspects among them. According to [8], WRF4G, LEAD and WRF Portals offer workflow design and monitor systems for WRF simulations. However, only WRF4G supports complex WRF experiments with simulation dependencies. Even though all these tools are advantageous to manage WRF,

none of them gathers all the capabilities that a WRF user requires to tackle ambitious climate experiments:

- Preparation and configuring WRF geographical domain.
- Localization, visualization and management of WRF's input and output data.
- A check-pointing system for experiment simulations.
- Mechanisms to execute and monitoring large collections of tasks.
- Capacity to execute and control tasks on different DCIs at the same time.
- Graphical workflow editor.
- Workflow repository system.

In order to provide a solution for all of the above requirements, we are going to cover them with a SG. The reason we selected a SG as solution is that SGs are able to undertake more ambitious problems with higher impact in society when the scientific challenges increase [9].

Due to the fact that we want to reduce the development time and cost, we are going to adapt existent technologies to develop our SG. As a result, on the one hand, we selected WRF4G framework because it includes an advantage set of tools to describe, design and manage WRF experiments. On the other hand, WS-PGRADE/gUSE was chosen as SG framework due to its features provide an easy way to customize SGs for research communities [10]. The combination of both frameworks is going to allow us not to develop our SG from scratch.

III. WRF4G AND WS-PGRADE/GUSE FRAMEWORKS

In this section, we are going to review the main features of both WRF4G and WS-PGRADE frameworks that we are going to use to develop WRF4SG.

A. WRF4G

WRF4G stands for WRF for Grid and it is [11] a framework for the execution and monitoring WRF model on DCIs such as supercomputers, clusters or Grid resources. WRF4G offers automation management solutions to design, execute and monitor WRF climate simulations and the ability of running these simulations on heterogeneous DCIs at the same time. Due to that, the definition and execution of WRF experiments are extremely simplified by using WRF4G.

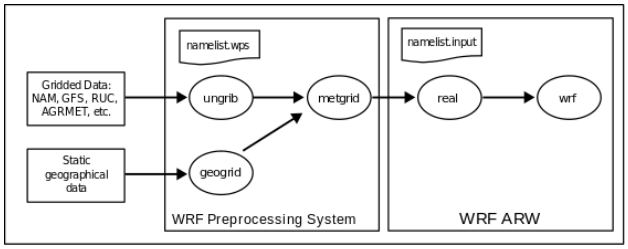


Figure 1. . The figure shows WRF workflow which is defined as a sequence of components such as ungrib, geogrid, metgrid, real and wrf. namelist.wps and namelist.input are the configurations files and the input data is represented in the left squares.

The framework separates the experiment design from the execution environment. Therefore, the experiment definitions are independent of the DCIs. Moreover, a monitor system is provided as an easy way to follow all the experiments available because the magnitude of a typical experiment exceed more than thousands of jobs.

It is important to note that the management system, which WRF4G offers, can divide simulation into chunks making checkpoints in order to create restart points. The simulation chunks can be submitted, resubmitted and checked independently. That feature is vital due to the fact that the simulations usually last several weeks or months even using MPI jobs.

B. WS-PGRADE/gUSE

WS-PGRADE/gUSE is [3] a generic DCI gateway framework and backend service stack that scientific communities can be used to customize and develop a SG according to their needs. As an open-source multi-grid application WS-PGRADE/gUSE supports development and submission of distributed applications executed on different DCIs. For instance, the last version available on [12] 3.5.2 supports DCIs such as local resources, gLite, GT2, GT4, ARC, UNICORE, PBS, LSF, web services, GEMLCA, BOINC desktops Grids and Google App Engine cloud.

WS-PGRADE portal is the user interface of gUSE (grid User Support Environment) service and both gUSE and WS-PGRADE facilitate to extend a generic gUSE Portal to a specific SG. This extension is possible via the Application Specific Module (ASM) [3] API web interface which is an easy-to-use solution to customize portlets for a gateway. In addition, WS-PGRADE/gUSE also offers a Remote API [3] to users that can use the gUSE functionalities independently from the WS-PGRADE portal interface.

In summary, the most important functional capabilities provided by WS-PGRADE/gUSE are [3]:

- Workflow support. WS-PGRADE/gUSE has a workflow editor application and different kind of tools in order to manage workflows.
- Enabling embedding workflows. WS-PGRADE/gUSE allows to use workflows as subworkflows by setting one into another. In addition, workflows can be shared among users by the SHIWA Simulation Platform [13].
- Enabling multi-DCI workflow execution. Different Virtual Organizations (VO) can be used simultaneously.
- Enabling the customization of the framework towards application-specific SGs.

IV. WRF4SG

During this section, we are going to describe in detail how WRF4SG is going to run and which solutions will be provided to WRF users. Specifically, we are going to focus on WRF4SG portlets and climate workflows linked to WRF. The objective of WRF4SG design is to simplify the configuration and execution of WRF model by a simple interface that requires a very basic user

configuration. In addition, we will also show the current features of WRF4SG prototype.

A. WRF4SG portlets

As we have mentioned previously, WS-PGRADE/gUSE offers an ASM API to customize a particular portal according to user requirements. That means that you can extend the generic WS-PGRADE/gUSE gateway to a user community by developing your own portlets. In our case, we are going to design tree portlets which are going to take advantage of WRF4G framework in their development. With these portlets, one user will be able to create, manage and execute WRF simulations by a simple Graphical User Interface (GUI).

The key features of WRF4SG portlets are explained below:

- Experiment design. WRF4SG will provide a portlet with a step-by-step process to define a WRF experiment. This design is going to be based on WRF4G experiment definitions due to the fact that it offers support to configure complex WRF experiments such as climate change, hindcast or weather forecast.
- Experiment management and monitoring. WRF4SG is going to provide elements to control all the available simulations in an experiment. The user will have a set of option buttons to submit, hold, stop, re-submit or cancel their experiments and simulations. Moreover, the portlet will also offer the necessary components in order to display dynamic experiment features such as progress or consumed resources. To this end, WRF4G will communicate that information to WRF4SG. Therefore, it is going to be developed a client-server communication framework between WS-PGRADE/gUSE and WRF4G.
- Pre and Post experiment data management. One of the biggest issues linked to WRF is to manage the input and output experiment datasets. On the one hand, before the experiment execution, WRF4G requires input datasets such as boundary and initial conditions for the geographical region which will be performed. On the other hand, the data out generated by WRF model has to be checked and post-processed. Therefore, this portlet is going to allow user to locate, select and visualize both input and output data by using WS-PGRADE LFC portlet which can interact with LFC catalog.

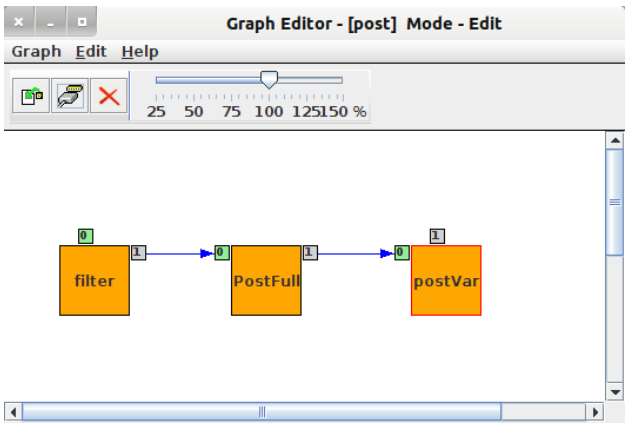


Figure 2. The figure shows the jobs of a post-processing workflow created by using WS-GPRADE/gUSE Graph. editor

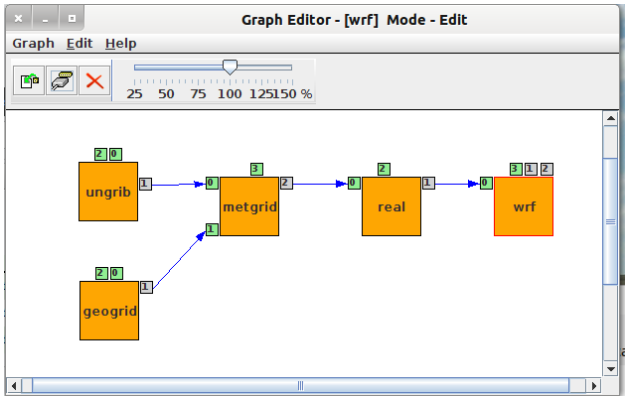


Figure 3. The figure shows the WRF workflow created by using WS-GPRADE/gUSE Graph editor.

B. WRF4SG workflows

Apart from the WRF4SG portlets, our objective is to create a set up of workflows in order to develop a wide range of them for climate researchers. In fact, we want to establish a catalog of climate workflows linked to WRF.

Due to the fact that the majority of workflows can usually run on different DCIs, they have to be generalist. Moreover, their features such as overhead, raw performance, stability, scalability and reliability on workloads have to be considered on their design [14].

The workflows available on WRF4SG will have to automate steps and tasks which WRF users usually do such as pre and post-processing data or running WRF executables. The examples below represent two typical climate workflows:

- Data post-processing. The Fig. 2 shows a workflow to post-process the WRF output data from a typical simulation. The workflow is composed of three jobs: the first job called *filter* detects the presence of corrupt files and missing data files, the second job called *postFull* performs operations over the available fields and finally the third job called *postVar* changes the meta-data files.
- WRF execution. The Fig. 3 shows the workflow of WRF executables (in this case, the executables are represented by jobs) which reproduces the Fig. 1 by using the WS-PGRADE/gUSE Graph Editor tool. This workflow does not include data assimilation and data post-processing. Due to that, both input and output data of all jobs (*ungrib*, *geogrid*, *real*, *metgrid* and *wrf*) have to be available on the same storage element.

C. WRF4SG scenario

Now we are going to examine a WRF4SG scenario in order to better illustrate the capabilities of WRF4SG. The scenario is to simulate the weather forecast over one day, on a limited area like Spain, with boundary conditions given by a general circulation model. In this scenario, the WRF4SG user is going to use the resources available in the European VO *esr*.

The use cases of this scenario are:

- User authentication. When the user wants to use the SG, they will be prompted for their credentials: username and password. After login, they will have to select which VO (in this case, *esr*) is going to be used by configuring the Security WS-PGRADE Portlet.
- Data assimilation. The input data will be located, selected and moved by using WS-PGRADE LFC Portlet. Then, the data will be checked in order to find errors.
- Experiment design. Most of the experiment configuration will be already provided for the user. They will have to fill out mandatory configuration parameters such as experiment type (in this case, weather forecast), output variables (wind, temperate, air pressure, etc.) or experiment start and end date. In addition, the user should specify the computing requirements for the experiment. Note that, the experiment will only be submitted if DCIs match the experiment requirements.
- Monitoring. The user is going to search their experiment using advantage query tools provided by WRF4SG. Additionally, the experiment will be monitored from a window which will show different experiment variables such as consumed time, data transfer or the progress.
- Data storage. As *Data assimilation* case, the user will be able to discover and view output files. Moreover, the user will be able to move the output data from one storage element to another.

D. WRF4SG prototype

The SG prototype of WRF4SG has been deployed on the font-end *ce02.macc.unican.es* and it is running on version 3.5.2 of WS-PGRADE/gUSE. The gateway has been configured to use gLite resources. In fact, there are three VOs configured on the SG such as *earth.vo.ibergrid.eu*, *esr* and *prod.vo.eu-eela.eu*. These are three of the most popular VOs at national level (Spain) and international level in Earth Science.

Even though this prototype has available climate workflows to pre- and post-process WRF simulation data, the SG does not offer the WRF4SG portlets described before.

V. Conlusions And Future Work

The aim of the WRF4SG is to allow WRF community to achieve more ambitious climate experiments by providing time saving tools such as a catalog of climate workflows or specific WS-PGRADE portlets. In addition, WRF4SG is going to be able to facilitate the access to different DCIs increasing the number of computing resources for the experiments.

As future work, we are planning to export the workflows designed to the SHIWA repository in order to promote our workflow catalog and we intend to add the portlets mentioned before to the WRF4SG prototype as well.

Acknowledgment

This work is partially funded by the Spanish PLAN NACIONAL de I+D+i 2008-2011 (Ref.# CGL2011-28864) and the European Regional Development Fund (ERDF).

References

[1] W. C. Skamarock, J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, «A Description of the Advanced Research WRF Version 2», *AVAILABLE FROM NCAR; P.O. BOX 3000; BOULDER, CO*, vol. 88, pp. 7–25, 2001.

[2] «The Weather Research&Forecasting Model Website». [Online]. Available: http://www.wrf-model.org/. [Accessed: 02-mar-2013].

[3] Kacsuk, Z. Farkas, M. Kozlovszky, G. Hermann, A. Balasko, K. Karoczkai, and I. Marton, «WS-PGRADE/gUSE Generic DCI Gateway Framework for a Large Variety of User Communities», *J Grid Computing*, vol. 10, n.º 4, pp. 601-630, dic. 2012.

[4] V. Fernández-Quiruelas, A. S. Cofiño, J. Fernández, and L. Fita, «WRF4G: enabling WRF on the Grid», presented at the Proceedings of the Second EELA-2 Conference, Madrid, 2009, pp. 149-154.

[5] V. Fernández-Quiruelas, L. Fita, J. Fernández, A. S. Cofiño, and J. M. Gutiérrez, «WRF4G: simpliying atmospheric modeling experiment design in distributed environments», presented at the IBERGRID 2011, 2011, pp. 327-332.

[6] «WRF Portal». [Online]. Available: http://www.wrfportal.org/. [Accessed: 02-mar-2013].

[7] M. Christie and S. Marru, «The LEAD Portal: a TeraGrid gateway and application service architecture», *Concurrency and Computation: Practice and Experience*, vol. 19, n.º 6, pp. 767–781, 2007.

[8] C. Blanco, A. S. Cofiño, and V. Fernández-Quiruelas, «WRF4SG: A Scientific Gateway for the Weather Research and Forecasting model community», presented at the IBERGRID 2012, Lisbon, 2012, in press.

[9] N. Wilkins-Diehr, D. Gannon, G. Klimeck, S. Oster, and S. Pamidighantam, «TeraGrid Science Gateways and Their Impact on Science», *Computer*, vol. 41, n.º 11, pp. 32 -41, nov. 2008.

[10] Z. Farkas and P. Kacsuk, «P-GRADE Portal: A generic workflow system to support user communities», *Future Generation Computer Systems*, vol. 27, n.º 5, pp. 454-465, may 2011.

[11] «WRF4G». [Online]. Available: http://www.meteo.unican.es/en/software/wrf4g. [Accessed: 02-mar-2013].

[12] «grid and cloud User Support Environment | Free software downloads at SourceForge.net». [Online]. Available: http://sourceforge.net/projects/guse/. [Accessed: 02-mar-2013].

[13] V. Korkhov, D. Krefting, T. Kukla, G. Z. Terstyanszky, M. Caan, and S. D. Olabarriaga, «Exploring workflow interoperability tools for neuroimaging data analysis», in *Proceedings of the 6th*

workshop on Workflows in support of large-scale science, New York, NY, USA, 2011, pp. 87–96.

[14] B. Plale, E. C. Withana, C. Herath, K. Chandrasekar, and Y. Luo, «Effectiveness of Hybrid Workflow Systems for Computational Science», *Procedia Computer Science*, vol. 9, pp. 508-517, 2012.

Generation of the Scrambled Halton Sequence Using Accelerators

Emanouil Atanassov and Mariya Durchova
Grid Technologies and Applications
Institute for Information and Communication Technologies
Bulgarian Academy of Sciences
Acad. G. Bonchev str. 25-A, Sofia 1113, Bulgaria
Email: emanouil,mabs@parallel.bas.bg

***Abstract*—The Halton sequence is one of the most popular low-discrepancy sequences. In order to satisfy some practical requirements, the original sequence is usually modified in some way. The scrambling algorithm, proposed by Owen, has several theoretical advantages, but on the other hand is difficult to implement in practice due to the trade-off between high memory and high computational requirements. In our work we concentrate on the case when the number of coordinates is relatively high. The use of computational accelerators and especially GPUs is increasingly relevant for such practical applications, since more and more of the resources, available through the Grid and Cloud infrastructures, provide access to such accelerators, provided that the software can make use of them. In this paper we discuss our algorithm for generation of the Halton sequence with Owen-type scrambling implemented using NVIDIA CUDA. We also show numerical results, achieved on our GPU-enabled nodes, which are equipped with NVIDIA M2090 cards.**

I. Introduction

The purpose of our paper is to describe an algorithm for generating the scrambled Halton sequences, suitable for parallel implementation using GPU. These sequences play important role in quasi-Monte Carlo algorithms. Consider an integral on the unit cube $I^s = [0,1]^s$ in s dimensions,

$$I[f] = \int_{I^s} f(x)dx, \tag{1}$$

of a Lebesgue integrable function $f(x)$, and note that this integral can be expressed as the expectation of the function f, $I[f] = E[f(x)]$, where x is a uniformly distributed vector in the unit cube.

We can approximate the integral (1) with

$$I_N[f] = \frac{1}{N}\sum_{n=1}^{N} f(x_n). \tag{2}$$

If $\{x_n\}$ is a sequence sampled from uniform distribution, the equation (2) is called (crude) Monte Carlo quadrature formula. The integration error, defined as

$$\epsilon_N[f] = |I[f] - I_N[f]|, \tag{3}$$

has a standard normal distribution, with expectation

$$E[\epsilon_N^2] = \sqrt{Var(f)}N^{-1/2}. \tag{4}$$

For any sequence $\{x_n\}$ an exact upper bound for (2) is given by Koksma-Hlawka inequality,

$$\epsilon_N[f] \leq V[f]D_N^*, \tag{5}$$

in which V[f] is the variation of f in the Hardy-Krause sense and D_N^* is the star discrepancy of the sequence [5]. Since the equation (5) is valid for any bounded function and any choice of sequence [5], it is important to consider sequences with low-discrepancy. It is believed that the best possible order of the discrepancy of an infinite sequence is $N^{-1}\log^s N$, where s is the dimensionality. Sequences with such order of discrepancy are called "low-discrepancy sequences" and are preferred in quasi-Monte Carlo methods, where they are used instead of pseudo-random numbers. When the effective dimensionality s is relatively high, the theoretical upper bounds are not so useful, but in practice the low-discrepancy sequences still yield orders of the error close to N^-1.

The Halton sequence is a popular low-discrepancy sequence, introduced by Halton in [15]. Although it provided one of the first examples of low-discrepancy sequences, the Halton sequence with some modifications is are widely used by applications of quasi-Monte Carlo methods since in many cases they offer superior results to other low-discrepancy sequences. In his work [16] Faure defines a wider class of sequences and shows how sequences with small discrepancy can be found in this class, when the dimension is one. He calls them generalized Van der Corput sequences. We give here the definitions of the generalized Van der Corput sequences and then of the Halton sequences.

Definition 1: Let $p \geq 2$ be a fixed integer, and let $\tau = \{\tau_j\}_{j=0}^{\infty}$ be a sequence of permutations of the numbers $\{0,\ldots,p-1\}$. The terms of the corresponding generalized Van der Corput sequence are obtained by representing $n \geq 0$ as

$$n = \sum_{j=0}^{\infty} a_j p^j, \qquad a_j \in \{0,\ldots,p-1\},$$

and putting

$$x_n = \sum_{j=0}^{\infty} \tau_j(a_j) p^{-j-1}.$$

The *one-dimensional Van der Corput - Halton sequence in base* p is obtained by setting $\tau_j(i) = i$.

Definition 2: Let $p_1, \ldots, p_s$ be pairwise relatively prime integers, $p_i \geq 2$. The Halton sequence

$$\sigma(p_1, \ldots, p_s) = \left\{ \left(x_n^{(1)}, \ldots, x_n^{(s)} \right) \right\}_{n=0}^{\infty}$$

is constructed by setting each sequence $\left\{ x_n^{(i)} \right\}_{n=0}^{\infty}$ to be a Van der Corput - Halton sequence in base p_i.

In the applications usually the numbers $p_1, \ldots, p_s$ are taken to be the first s primes, because it is expected that this choice offers smallest discrepancy and consequently faster convergence of the resulting quasi-Monte Carlo method.

E. Atanassov in his paper [1] introduced a modification of the Halton sequences in multiple dimensions, and proved that for these sequences an estimate of their discrepancy with a very small leading term can be obtained (see Theorem 2.1 and Theorem 2.3). His construction is based on the existence of some numbers, called "admissible", which define certain sets of permutations τ_j^i for a given set of primes. From some point of view this construction can be considered to be a specially selected case of a so-called "scrambling" procedure, i.e. a procedure that adds an element of randomness to the otherwise fully deterministic low-discrepancy sequence. One of the reasons for introducing "scrambling" in general is to add the possibility for obtaining aposteriory error estimate for the low-discrepancy sequence. In some cases "scrambling" leads to improvement of the quality of the distribution of the quasi-random number sequences.

While in Monte Carlo methods one can use the standard deviation of the result in order to estimate the probable error of a computation, in quasi-Monte Carlo methods one can not rely on this technique if the sequence is fully deterministic. There are various constructions for achieving "scrambling". A simple example would be to add a random number u_i to each coordinate of the sequence (modulo 1). Significant improvement is offered by the more complicated construction, proposed by Owen in [19], where he considers deterministic and randomized nets and shows how for smooth integrands over s dimensions, the variance is of order

$$N^{-3} (\log N)^{s-1},$$

compared to N^{-1} for ordinary Monte Carlo. Thus the integration errors are of order

$$N^{-\frac{3}{2}} (\log N)^{\frac{s-1}{2}}$$

in probability. Here we apply his construction, which is usually applied to (t, s)-sequences, to the Halton sequences:

Definition 3: Consider for each dimension k a tree of independent uniformly distributed permutations σ_{j,B_j}^k of the numbers $0, \ldots, p_k - 1, j = 0, \ldots \infty$ and for each $j, B_j \in \left[0, \ldots, p_k^{j-1} \right]$:

$$\begin{array}{lll} c_0^k = & \sigma_{0,0}^k(b_0) & \\ c_1^k = & \sigma_{1,B_1}^k(b_1), \quad where & B_1 = b_0 \\ \ldots & \ldots & \ldots \\ c_j^k = & \sigma_{j,B_j}^k(b_j), \quad where & B_j = \sum_{i=0}^{j-1} b_i p_k^i \end{array}$$

The *scrambled Halton* k *-dimensional sequence* is defined as

$$\sum_{j=0}^{\infty} c_j^k p_k^{-j-1}$$

While Owen's construction offers significant theoretical advantage over the simpler types of scrambling,it is more complicated for efficient software implementation, especially in the case of high number of dimensions. In principle the "scrambling" of the sequence can be achieved by rearranging the terms of the sequence in RAM. In this case the amount of memory that would be needed is proportional to the product of number of dimensions and number of terms that are required by the algorithm. In our research in semiconductor device modelling we need millions and even billions of random trajectories, where each trajectory can have tens of transitions. That is why we concentrated on the idea to use more computational time in order to re-compute each random permutation when it is necessary. This approach has the problem that the use of a pseudorandom generator for generating the permutations on-the-fly can reveal some weakness of the generator, which may not be evident from using the generator in Monte Carlo algorithms. That is why in our implementation we decided to use a pseudorandom function that has been proposed for use in cryptography. When we had implemented such an approach on Intel and AMD CPUs, we have found that the trade-off between speed of generation versus convergence rate improvement is not acceptable. However, with the appearance of the GPGPU computing paradigm, the situation is changed significantly.

When using the GPU for general purpose computations one has to make use of high number of parallel threads, with the limitation that the available memory and cache per thread is relatively low. Programming for the GPU has been eased significantly since the introduction of CUDA by NVIDIA. While programming using CUDA presents some challenges, it does not pose significant problems for persons with background in parallel programming using MPI.

For our algorithm the the benefit of using the GPU is that we decrease significantly the computational time for the generation of the scrambled sequence making it comparable to that for computing the underlying function on the CPU. This happens mainly because the generation is parallelizable without interdependency between the threads and it employs mostly integer operations for the most time consuming part.

In the next sections we describe the algorithm and its particular implementation using CUDA and present timing and numerical results to demonstrate its superiority versus

other possible approaches. Later on we outline an approach for parallel use of several GPUs and CPUs in a cluster using MPI.

II. Description of the algorithm

The main idea in our implementation is that if one generates p random real number $x_1, \ldots, x_s$ in $[0, 1]$, and then sorts them, this sorting defines a random permutation σ. In order to determine the place $\sigma(k)$ where the number k goes under this permutation, one has to count the number of indices i, for which $x_i < x_k$.

The ChaCha family of stream ciphers, also known as Snuffle 2008, is a variant of the Salsa20 family of ciphers. The described in [4] pseudorandom ChaCha function $h(x)$ produces a random 512-bit output from a given 128-bit x.

Let $x = 0, b_1 b_2 \ldots$ in the p-adic number system, $p < 64$ and we choose s random seeds. Applying the psuedorandom function h we interpret the output as 64 random 8-bit integers considering the first p of them. When there are cases of coincidence of the first 8-bits, then more random 8-bit integers are generated until one can distinguish between the two real numbers that are to be compared.

The ChaCha pseudorandom function h and stream cypher are designed with efficient implementation in mind and are GPGPU-friendly, since they do not use much memory. If we consider $p < 64$, then one computation of h will be enough in most cases, while with some small probability 2 or more computations of h will be required. If $64 < p < 256$ then we consider the output of 4 invocations of h as 256 random 8-bit integers and proceed in the same way. In this case 4 instead of 1 invocation of h will be required in most cases, while with smaller probability 8 or more invocations will be required.

When $256 < p < 1024$, we consider the output of 4 invocations of h as 1024 random 2-bit integers. In this case there will be more collisions, which will enlarge the total number of invocations of h, by approximately a factor of 2. Our current implementation is limited to $p < 1024$, but in principle the idea can be extended further.

We call the generation kernel as follows:

```
gener_kernel <<< grid, block >>>
                        (nonse,dim,N,x,areainfodevice);
dim3    grid(dim,1);
dim3    block(blockingsize,1,1);
blockingsize=128
```

The variable *areainfodevice* holds all the information necessary for generating the next set of 128 terms of the sequence from the previous block. The most GPU time is spent inside the pseudorandom function, that is why we consider separately the cases of $p < 64$ and the others.

That is how for each digit to scramble we compute the action of the corresponding random permutation. Effectively we compute with one kernel invocation 128 terms of the sequence, parallelising also along the dimensions, which are presumably high, e.g., 10.

In order to use the generation routines in an MPI computation, we split the terms of the sequence in blocks and generate each block in a different MPI processor. For the type of computations that we do the most effective utilization of the hardware resource is achieved if one core of the CPU and one GPU device are used for each MPI process. This is achieved during pre-processing, by choosing the appropriate GPU device for each process.

III. GPGPU-enabled target system

In our GPU-enabled computing facilities, which are part of the HPCG cluster in HP-SEE project, we have two different types of servers. The older computing nodes have GTX 295 cards, while the new servers are equipped with NVIDIA Tesla M2090 resources. Since most of our computations are done in double precision, the M2090 cards are far superiour and that is why we report results only on them.

Our M2090 cards are distributed between two servers. They can be accessed passwordless from the main login node of the cluster and jobs can be submitted using torque batch system. The M2090 cards have CUDA Capability 2.0. The older GTX 295 support only 1.3, but it is enough for our purposes. Our code was developed under previous versions of CUDA SDK, but is now ported to CUDA 5.0, deployed on Scientific Linux SL version 5.4.

The results are performed on HP ProLiant SL390s G7 server equipped with two Intel@Xeon@E5649 (6 cores, 2.53GHz, 12MB L3, 95W) processors.

IV. Numerical experiments and results

A. Numerical Integration

At first we will show the results for the next test functions:

$$F_1 = \prod_{i=1}^{s} \left({x_i}^3 + \frac{3}{4} \right) \tag{6}$$

$$F_2 = \prod_{i=1}^{s} |4x_i - 2|, \tag{7}$$

in the s-dimensional unite cube $I^s = [0, 1]^s$.

Table I shows the accuracy achieved for the test function (6) F_1 in correspondance with the number of generated points N and where s is the dimension of the integral. We compare the results by our scrambled Halton (SH) algorithm with the deterministic achieved Halton algorithm (DH).

The concrete accuracy for F_1 and $s = 8$ is given in Fig. 1.

Table II shows the measured time which is needed to compute scrambled Halton sequence for sample size N (SHalton) and comparison with the deterministic one (DHalton).

The accuracy for the function (7) F_2 with dimension $s = 6$ and $s = 16$ are given in Table III.

The accuracy for $s = 16$ is given on the Fig. 2

We compare in Table IV the computational time needed for generating the N 16-dimensional sequences.

TABLE I. ACCURACY

N		$s=4$	$s=7$	$s=8$	$s=16$	$s=32$
2^{16}	DH	3.379e-04	1.185e-03	1.829-03	1.457e-02	9.037e-02
	SH	1.236e-05	6.569e-05	8.012e-05	1.445e-03	9.185e-03
2^{17}	DH	1.615e-04	5.931e-04	1.028-03	8.652e-03	5.875e-02
	SH	5.707e-06	1.487e-05	3.4749e-05	8.055e-04	6.537e-03
2^{18}	DH	9.860e-05	3.532e-04	4.961e-04	5.698e-03	4.017e-02
	SH	1.283e-06	1.390e-05	3.145e-05	5.430e-04	2.9761e-03
2^{19}	DH	5.174e-05	1.780e-04	2.415e-04	2.929e-03	2.095e-02
	SH	1.935e-06	1.012e-05	9.857-06	3.766e-04	1.636e-03
2^{20}	DH	2.904e-05	8.944e-05	1.432e-04	1.624e-03	8.567e-03
	SH	1.763e-06	4.143e-06	7.218e-06	1.344e-04	1.014e-03
2^{21}	DH	1.405e-05	4.937e-05	7.601e-05	9.902e-04	9.495e-03
	SH	4.430e-07	3.2645e-06	4.893e-06	5.770e-05	5.376e-04
2^{22}	DH	6.467e-06	2.494e-05	4.046e-05	5.999e-04	6.341e-03
	SH	3.074e-07	9.698e-07	1.822e-06	4.306e-05	4.139e-04
2^{23}	DH	3.566e-06	1.237e-05	2.021e-05	3.571e-04	3.665e-03
	SH	2.086e-07	4.587e-07	7.903e-07	2.071e-05	4.351e-04
2^{24}	DH	1.912e-06	7.114e-06	1.116e-05	1.928e-04	1.983e-03
	SH	7.405e-08	4.204e-07	4.251e-07	1.302e-05	4.704e-04

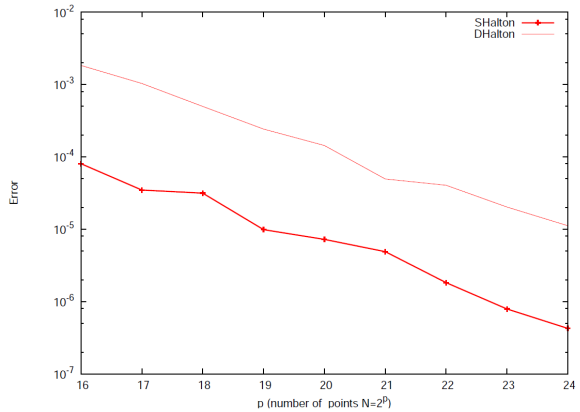


Fig. 1. Accuracy of numerical integration of **Test function** F_1 with sample size N and dimension $s = 8$ using deterministic and scrambled Halton sequence.

TABLE II. COMPUTATIONAL TIME (IN SECONDS), S=8

N	SHalton	DHalton	ratio
2^{16}	2.531	2.309	1.096
2^{17}	2.839	2.431	1.167
2^{18}	3.459	2.609	1.326
2^{19}	4.643	2.987	1.554
2^{20}	7.160	3.760	1.904

TABLE III. ACCURACY

N	2^{16}	2^{18}	2^{19}	2^{20}
SH,$s=6$	3.37e-04	1.44e-04	7.94e-05	2.81e-05
DH	2.53e-04	3.90e-05	5.68e-05	4.72e-05
SH,$s=16$	2.81e-02	7.83e-03	5.58e-03	5.49e-03
DH	9.82e-01	2.34e-01	1.16e-01	5.56e-02

B. Geometric Asian Options

Geometric mean definition of d assets can be interpreted as:

$$G = \left(\prod_{i=1}^{d} S_i \right)^{\frac{1}{d}},$$

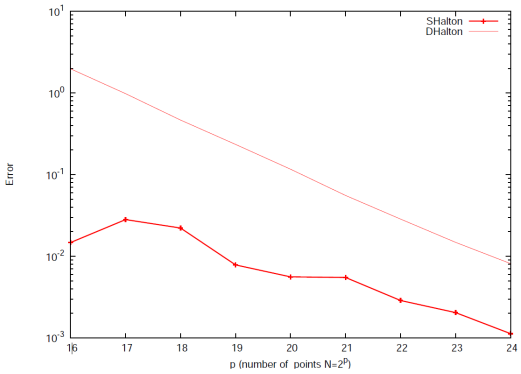


Fig. 2. Accuracy of numerical integration of **Test function** F_2 with sample size N and dimension $s = 16$ using deterministic and scrambled Halton sequence.

TABLE IV. COMPUTATIONAL TIME (IN SECONDS), $s = 16$

N	2^{15}	2^{16}	2^{17}	2^{18}	2^{19}	2^{20}	2^{21}	2^{22}	2^{23}
SH	2.65	3.06	3.92	5.58	8.92	15.58	28.95	55.63	109.10
DH	2.33	2.46	2.67	3.10	3.97	5.69	9.15	16.07	29.90
ratio	1.13	1.24	1.47	1.8	2.27	2.73	3.16	3.46	3.64

where S_i - i^{th} asset price.

The payoff of the call option at the maturity date T is defined as $max\{0, G-K\}$, where K is the strike price at the date T.

An analytical solution for pricing a geometric Asian option is described in [18]. The basic idea followed there is that the product of lognormally distributed variables is also lognormally distributed because the behavior of S_i follows geometric Brownian motion.

To evaluate European call option it is used the Black-Scholes equation [17] as follows:

$$C_T = S * Norm(d_1) - K * e^{-r(T-t)} * Norm(d_2),$$

$$d_1 = \frac{ln(\frac{S}{K}) + (r+\sigma^2)(T-t)}{\sigma\sqrt{T-t}},$$

$$d_2 = d_1 - \sigma\sqrt{T-t},$$

where t is the current time, r is risk-free rate of interest (constant) and $Norm(d_2)$ is cumulative normal distribution.

In our simulation we consider the described in [6] set of parameters:

- Number of assets $d = 30$
- Initial asset prices, $S_i(0) = 1$, $i = 1 \ldots d$
- Volatilities, $\sigma_i = 0.3$
- Correlations, $\rho_{ij} = 0.5$, for $i < j$
- Strike price, $K = 1$
- Risk-free rate, $r = 10\%$
- Time to maturity, $T = 1$ year

It should be noted that in such an simulation the order $d = 30$ is comparatively high, which require more computational resource for generating the 2^{21} 30-dimensional sequences.

Applying our algorithm for generating the scrambled Halton sequence for this option simulation we achieve the results, showed in Table V and in Fig. 3.

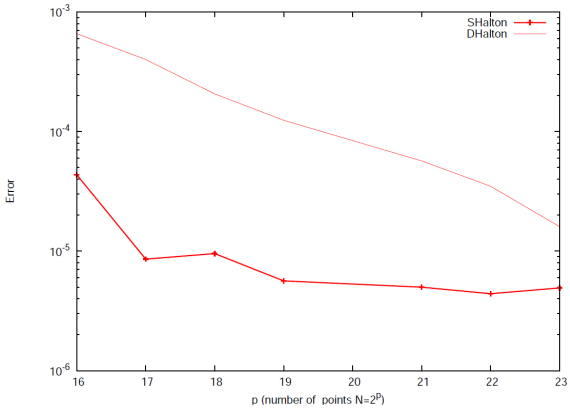


Fig. 3. Geometric mean of 30 stock prices using Halton and scrambled Halton sequences

TABLE V. SIMULATION TIME t (IN SECONDS), $s = 16$

N	2^{15}	2^{16}	2^{17}	2^{18}	2^{19}	2^{20}	2^{21}
SH	2.77	3.28	4.30	6.38	10.5	18.76	35.16
DH	2.53	2.83	3.42	4.63	7.00	11.76	21.27
ratio	1.09	1.16	1.26	1.38	1.50	1.60	1.65

The results demonstrate that the trade-off between computing time and accuracy is acceptable for the sequence with Owen scrambling applied. When more realistic computations, like those in Asian option pricing, are to be performed, the situation is even better for the scrambled sequence, because the relative weight of the generation code in the total computational time is decreased further.

We do not compare here with CPU-based implementation of the algorithm, because it is unfeasible to use such implementation in real-world computations due to relatively high cost of computations of the Cryptographic hash function h.

V. CONCLUSION

We have developed and algorithm for generating the Halton sequences that implements full Owen scrambling targeting GPU computing resources, as available in the HP-SEE project. The algorithm achieves acceptable trade-off between use of memory and computing resources and is usable in situations, where high number of coordinates are required. The overhead of scrambling is also acceptable considering the improvement in the numerical results. The algorithm can successfully utilize the available parallelism in GPU cards.

ACKNOWLEDGMENT

This work was supported by the National Science Fund of Bulgaria under Grants DCVP 02/1 and by the European Commission under EU FP7 project HP-SEE under contract number 261499.

REFERENCES

[1] E. Atanassov, A New Efficient Algorithm for Generating the Scrambled Sobol' Sequence, *Numerical Methods and Applications, Springer*, **2542**, 2003, 83-90.

[2] H. Chi, P. Beerli, D. W. Evans, M. Mascagni, "On the Scrambled Sobol Sequence", Lecture Notes in Computer Science, Springer, vol. 3516, 2005, pp. 775-782.

[3] S. Ivanovska, E. Atanassov, A. Karaivanova, A Superconvergent Monte Carlo Method for Multiple Integrals on the Grid, *LNCS, Springer-Verlag, Berlin-Heidelberg*, **3516**, 2005, 735-742.

[4] D. J. Bernstein, " ChaCha, a variant of Salsa20", http://cr.yp.to/chacha.html

[5] R. Caflisch, Monte Carlo and quasi-Monte Carlo methods, *Acta Numerica*, **7**, 1998, 1-49.

[6] H. Chi, P. Beerli, D. W. Evans, M. Mascagni, "On the Scrambled Sobol Sequence", Lecture Notes in Computer Science, Springer, vol. 3516, 2005, pp. 775-782.

[7] H. Chi and E. Jones, Generating Parallel Quasirandom Sequences by using Randomization, *Journal of distributed and parallel computing*, **67(7)**, 2007, 876-881.

[8] H. Chi and M. Mascagni, Efficient Generation of Parallel Quasirandom Sequences via Scrambling, *Lecture Notes in Computer Science, Springer*, **4487**, 2007, 723-730.

[9] B. C. Bromley, Quasirandom Number Generation for Parallel Monte Carlo Algorithms, *Journal of Parallel Distributed Computing*, **38(1)**, 1996, 101-104.

[10] W. Schmid, A. Uhl, Techniques for parallel quasi-Monte Carlo integration with digital sequences and associated problems, *Math. and Comp. in Sim.*, **55**, 2001, 249-257.

[11] A. Owen, The Dimension Distribution and Quadrature Test functions, *Statistica Sinicai*, **13**, 2003, 1-17.

[12] S. Chaudhary, Acceleration of Monte Carlo Methods using Low Discrepancy Sequences, *Dissertation, University of California, Los Angeles*, 2004.

[13] G. Okten and B. Tuffin and V. Burago, A central limit theorem and improved error bounds for a hybrid-Monte Carlo sequence with applications in computational Finance, *Journal of Complexity*, **22(4)**, 2006, 453-458.

[14] MPI website, http://www-unix.mcs.anl.gov/mpi/

[15] J. H. Halton, On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals, Numerische Mathematik, 1960, Volume 2, Issue 1, pp 84-90

[16] H. Faure, Discrpance de suites associes un systme de numracion (en dimension un). H. Bull. Soc. Math. France, 109 (1981) 143182.

[17] F. Black, M. Scholes, The pricing of option and corporate liabilities, Journal of Political Economy,81:637-659, 1973

[18] P. Boyle, "New life forms on the option landscape", Journal of Financial Engeneering, 2 (3), 1992, pp. 217-252.

[19] A. B. Owen, Scrambled Net Variance for Integrals of smooth Functions, Annals of Statistics, 1997, Vol. 25, No. 4, pp. 1541-1562

Intel vs AMD: Matrix Multiplication Performance

Nenad Anchev, Marjan Gusev, Sasko Ristov and Blagoj Atanasovski
Ss. Cyril and Methodius University
Faculty of Information Sciences and Computer Engineering
Skopje, Macedonia
Email: nenad_ancev@hotmail.com, {marjan.gushev,sashko.ristov}@finki.ukim.mk, blagoj.atanasovski@gmail.com

***Abstract*—Matrix-Matrix multiplication (MMM) is widely used algorithm in today's computations and researches. Many techniques exist to speed up its execution. In this paper, we analyze the performance of MMM varying matrix size in order to determine its behavior and the region where it provides the best performance. We also determine the best speedup and efficiency in parallel implementation for different CPU architectures since cache architecture and organization is very important for MMM performance. Intel i7 and AMD Opteron CPUs are used as an environment. Several achieved results are expected, but there are also many unexpected. Superlinear speedup (speedup greater than the number of used threads) and the efficiency greater than 100% are achieved for each parallel implementation only on AMD Opteron. We observe regions with performance discrepancy for speed, speedup, and efficiency for both CPUs.**

***Index Terms*—HPC, CPU, Cache, Memory, Superlinear Speedup**

I. Introduction

MMM is a linear algebra algorithm which has been a subject of research in industry and the science for a long time. There are many techniques to speedup the algorithm: using faster processors, modifying the algorithm for less operations, using different hardware architecture. Since it is quite well for parallelization (granular and scalable), huge speedup can be achieved if it is scaled on many cores.

Various CPU vendors produce different CPU architectures. In this paper we realize a systematic approach to analyze the performance of two different CPUs, i.e. Intel i7 and AMD Opteron using the same MMM algorithm. We choose these CPUs since both have similar cache architecture. Both use four cores with private L1 and L2 caches, and all cores share the last level L3 cache.

Our goal is not to improve the "slow(est)" dense MMM, but we use it intentionally since it generates a lot of cache misses and is a good benchmark for many cache parameters: cache size, cache levels, replacement policy, set associativity, cache line size, cache inclusivity / exclusivity etc. We measure the response time for each test case and calculate the speed, and speedup and efficiency in parallel executions.

The rest of the paper is organized as follows: Section II describes the related work. In Section III we describe the testing methodology used in our experiments. The next Section IV elaborates the results using three parameters: speed, speedup and efficiency for both platforms. We discuss the results in Section V. Finally, Section VI concludes our work and Section VI presents our plan for further work.

II. Related Work

MMM algorithm is the most common used algorithm in computations, as well as in research. Many authors analyze their performance trying to speedup its execution with different techniques classified in two main categories: improving the algorithm by reducing the operations or improving the algorithm according to available multiprocessors, especially its cache architecture and organization. Both techniques are important since MMM is cache intensive algorithm [1], i.e. each element is reused N times, where N is the the matrix size. In this paper we try to understand the behavior of the MMM algorithm presented with the three most important parameters: speed, speedup and efficiency, while the MMM algorithm is executed on two different single-chip multi-core multiprocessors with the same cache structure.

A. Speed Analisys

We found many papers that improve the algorithm according to the available hardware. Hennessy and Patterson [2] present 2D blocking matrices techniques as one of several cache optimization that enormously decreases cache misses and thus the execution time. Although this techniques increase the operations, it reduces the high cache level misses and accesses to main memory since block sizes are chosen as the matrices can be stored in L1 cache. Gusev et al. even improve the 2D blocking matrices by using rectangles instead of squares. They reduce the performance drawbacks due to associativity [3] for the processors with small associativity (AMD) by reducing the height of the second matrix B blocks. Williams et al. [4] used padding to the first element of each submatrix to land on equidistant cache sets and thus reduce the drawbacks due to cache set associativity.

B. Speedup Analisys

Achieving linear speedup when the algorithm is scaled is imperative. However, even superlinear speedup is achieved by many authors, but without detailed explanation [5], [6]. Al-Jaroodi et al. [7] found superlinear speedup for matrix multiplication explaining that more processors have more cache and thus parallel execution will generate smaller number of cache misses. However, they do not explain why superlinear speedup is achieved just for 2 and 3 processors, and not for 4 and more processors where more cache capacity is available. Kolberg et al. [8] found superlinear speedup for MMM on 16 processors using MPI, but not for 32 or 64, despite increased

cache memory. More detailed but also incomplete analysis is performed in [9] and [10].

Ristov and Gusev [11] introduced more detailed analysis about superlinear speedup. They proved experimentally that superlinear speedup region exist for different matrix sizes and scaling the MMM on more cores choosing the matrix size from the superlinear region will lead to superlinear speedup regardless of the number of processors. Superlinear speedup is achieved in cloud virtual environment and even on Windows platform in Windows Azure [12], despite the virtualization layer. It can be achieved on multi-GPU implementation [13] due to configurable cache memory of Fermi architecture, as well. Jenks [14] found superlinear speedup with parallel execution of matrix multiplication algorithm using MPI and transposing one source matrix, thus reducing the cache misses. Adding parallel overheads in order to increase both cache reuse and fine-grained parallelism can lead to superlinear speedup [15].

We determine superlinear speedup for MMM as well. But we found another reason, i.e. implicit prefetching due to shared last level L3 cache.

C. Efficiency Analisys

We have not found any paper that analyzes the efficiency while scaling the MMM by number of processors. In this paper we will try to understand if scaling the processor resources will increase or decrease the efficiency on both multiprocessors. We have set a hypothesis that the efficiency decreases when the resources are increased for the same problem size.

III. Testing Methodology

This section presents the testing methodology used for the experiments in order to provide reliable results.

A. Testing Algorithm

We use Dense MMM algorithm with squared matrices $C_{N\cdot N} = A_{N\cdot N} \cdot B_{N\cdot N}$ as test data. Matrix elements are stored as double precision numbers with size $ME = 8$ bytes each. Each element c_{ij} of matrix C is calculated as inner product of row i of matrix A and column j of matrix B, for each $i, j = 0, 1, \cdots, N-1$. Cache miss occurs if an element is not present in the cache and the processor needs to load it from main memory, which is much more expensive operation. Increasing the matrix size N will occupy the cache faster, thus increase the cache miss ratio and the total execution time.

We are not interested in speeding up the algorithm by reducing the number of operations, but we use this algorithm since it is cache intensive.

One thread in sequential implementation multiplies the whole matrix $A_{N\cdot N}$ and matrix $B_{N\cdot N}$. For parallel implementation, each thread multiplies the row block matrix $A_{N/P\cdot N}$ and the whole matrix $B_{N\cdot N}$, where $P \in \{2, 3, 4\}$ denotes the total number of parallel threads and used CPU cores. Both implementations are executed without any optimization or adaptation to a certain CPU architecture.

A single thread in sequential implementation executes N^3 sums and N^3 products, or total $2 \cdot N^3$ operations. Each thread in parallel implementation executes average $2 \cdot N/P \cdot N^2$ operations. This means that the algorithm performs $2 \cdot N^3$ operations, both for sequential and parallel execution.

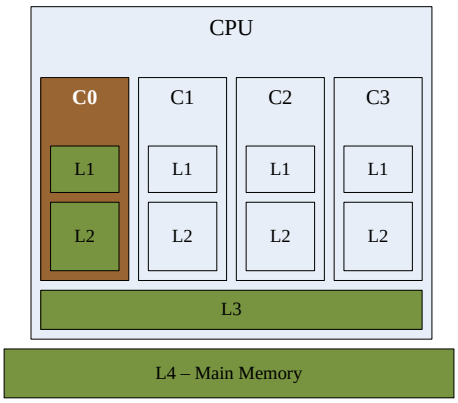


Fig. 1. CPUs cache architecture

CPU	L1D	L1I	L2	L3
Intel i7	32KB	32KB	256KB	8MB
AMD Opteron	64KB	64KB	512KB	2MB

TABLE I
Cache size for i7 and Opteron CPUs

B. Testing Environment

Our testing environment consists of two different vendor CPUs, i.e. Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz and Quad-Core AMD Opteron(tm) Processor 8347, each with 4 cores, as depicted in Fig. 1. Each core has its own private L1 and L2 caches and all 4 cores share L3 cache. Table I presents the details about cache size of both CPUs, where L1D and L1I denote for L1 data and instruction cache, correspondingly.

Both servers are configured with the same runtime environment: Linux Ubuntu 10.10 operating system is installed on both machines and C++ with OpenMP is compiled with gcc compiler.

C. The Experiments

We realize 4 experiments on both platforms. Each experiment consists of series of test cases varying matrix size from $N = 1$ to $N = 1100$.

1) Sequential Implementation - Experiment 1: Sequential execution is realized with one thread on one core C0 as depicted on Fig. 1. This experiment uses C0 core with its private L1 and L2 caches, as well as the whole L3 Cache.

2) Experiments 2, 3 and 4 - Parallel Implementations: Three experiments are realized using 2, 3 and 4 processors. Each processor has its own private L1 and L2 caches, but now share L3 cache.

D. Test Data

We measure the execution time $T(P)$ for each test case (different matrix size N) in each experiment (different number

of processors P). We use $T(P)$ to calculate the *Speed* $V(P)$, the *Speedup* $S(P)$ and the *Efficiency* $E(P)$ using well known definitions (1), (2) and (3).

We calculate the speed $V(P)$ by dividing the number of operations and measured execution time as defined in (1), both for sequential and parallel execution, i.e. for $P \in \{1, 2, 3, 4\}$.

$$V(P) = \frac{2 \cdot N^3}{T(P)} \tag{1}$$

The speedup $S(P)$ for parallel execution (for $P \in \{2, 3, 4\}$) is calculated as ratio of sequential and parallel execution time by (2).

$$S(P) = \frac{T(1)}{T(P)} \tag{2}$$

Finnaly, the efficienacy $E(P)$ is calculated for parallel execution (for $P \in \{2, 3, 4\}$) as ratio of the speedup and the number of processors (3). We will use the efficiency to determine which parallel implementation provides the best speedup per core.

$$E(P) = \frac{S(P)}{P} \tag{3}$$

E. Test Goals

The test experiments have two goals. The first goal is to model the multiprocessor's behavior, i.e. speed, speedup and efficiency as a function of matrix size for cache intensive MMM algorithm on both environments. The second goal is to determine which platform provides better performance.

F. Testing Regions

Gusev and Ristov [1] define different cache regions L_1, L_2, L_3 and L_4 where matrix multiplication behaves differently. The L_1 region is determined with those matrix sizes that can be stored in L1 cache. Then we expect the highest processor speed and the smallest number of cache misses. L_2 region is determined when the matrices size is such that the matrices can not be stored in L1 cache, but fit in the L2 cache generating cache misses in L1, but not for L2. Respectively in L_3 region L2 generates cache misses but data fits in L3. The region after this point is called L_4 region presenting main memory or L4 cache.

IV. Results of the Experiments

In this section we present the results of the experiments realized in two CPUs with the same cache distribution, but different cache size, as explained in Section III-B.

A. Results on Intel i7 CPU

We measure the execution time for each number of processors $P \in \{1, 2, 3, 4\}$. In this section we analyze the results of calculated Speed $V(P)$, Speedup $S(P)$ and Efficiency $(E(P)$ for each experiment defined in Section III-C while executed on Intel i7 processor.

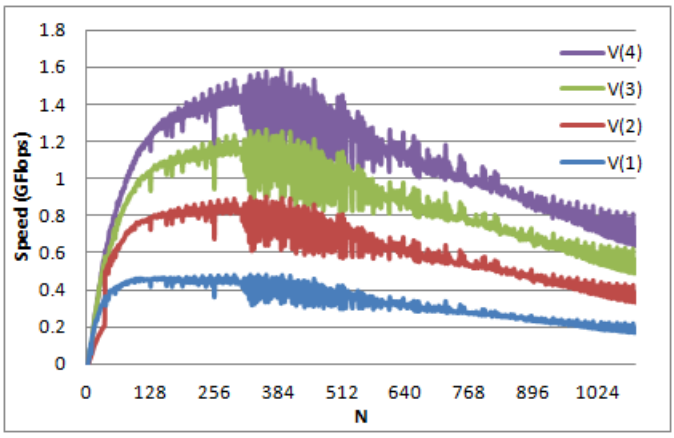


Fig. 2. Speed for MMM as a function of matrix size while executed on different number of Intel i7 cores

1) Speed on Intel i7 CPU: The speed for each test case of all four experiments while executed on Intel i7 processor is depicted in Fig. 2.

We observe three regions (A, B, and C going from left to the right) with different speed behavior for each experiment. The speed increases in region A while N increases until region B where it saturates. Then it decreases in region C which is more emphasized for the experiments with greater number of processors. However, the three regions are different for each experiment and their ranges are presented in Table II.

TABLE II
Different Speed Behavior

Exp.	Region A	Region B	Region C
1	[1,102]	[102,380]	[380,1100]
2	[1,135]	[135,380]	[380,1100]
3	[1,275]	[275,380]	[380,1100]
4	[1,295]	[295,380]	[380,1100]

Speed drawbacks are observed for $N = 128$ and more emphasized for $N = 256$ for all experiments due to cache set associativity [3]. The drawback is more emphasized while using more processors since all processors force for the same set and more cache misses are generated.

Another unexpected speed behavior is in the region $N = [312, 580]$ where its value is discrepant. Another discrepant region is for matrix size $N > 932$. We believe that this discrepancy happens due to cache line, L2 and L3 cache capacity and associativity. That is, memory access time is smaller if matrix and vector elements fulfill the cache lines more efficiently and reading one element will load the whole cache line. But, this pattern suffers of cache associativity since more elements will be stored in the same cache set and more cache misses will be generated. We will analyze these unexpected results in more details in our further research analyzing the cache misses, the address of the first element of matrices and matrix pattern in the cache and main memory.

We can conclude that the maximum performance (speed) can be achieved when the matrices A and B can be fit in L3

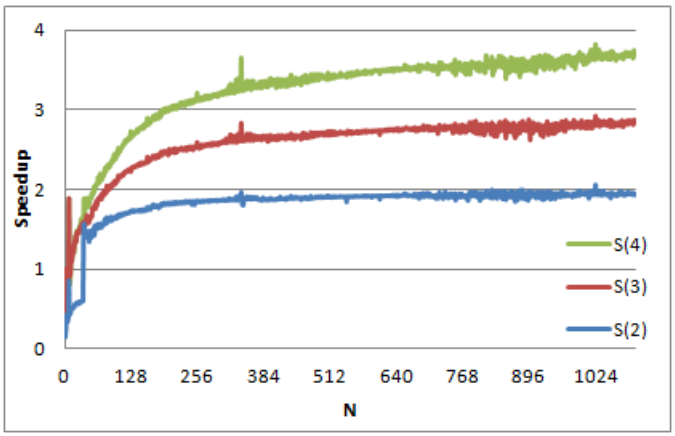


Fig. 3. Speedup for MMM executed as a function of matrix size while executed on different number of Intel i7 cores

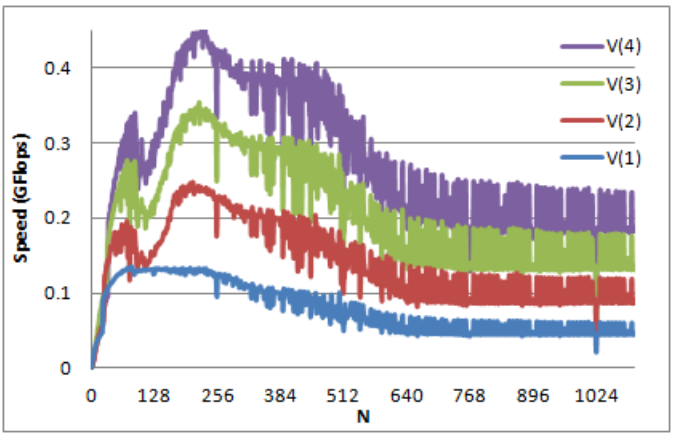


Fig. 5. Speed for MMM as a function of matrix size while executed on different number of AMD Opteron cores

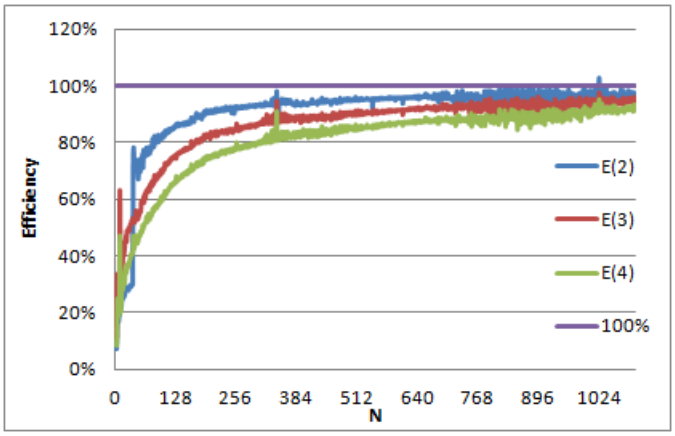


Fig. 4. Efficiency for MMM executed as a function of matrix size while executed on different number of Intel i7 cores

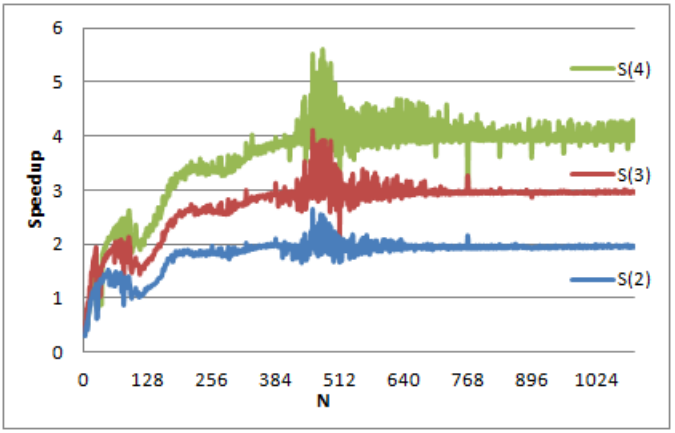


Fig. 6. Speedup for MMM executed as a function of matrix size while executed on different number of AMD Opteron cores

cache, i.e. around $N = 380$. After this point, the speed starts to decrease for each experiment.

2) Speedup on Intel i7 CPU: The achieved speedup for the three experiments with parallel implementation compared to sequential execution is depicted in Fig. 3.

The speedup in the three experiments with parallel implementation satisfies Gustafson's Law [16], i.e. it is sublinear in each test case. The speedup rises and saturates until its limit, i.e. $S(P) \rightarrow P$ for each experiment.

Two unexpected results, i.e. a small positive peak for $N = 342$ and the discrepant speedup in the region for matrix size $N > 736$, will be analyzed in more details in our future work.

3) Efficiency on Intel i7 CPU: Fig. 4 depicts the efficiency for the three experiments with parallel implementation. All three experiments have similar curves for the efficiency, where experiment 2 is the most efficient implementation in front of experiment 3 and 4, i.e. $E(2) > E(3) > E(4)$ for each test case. We can conclude that adding more resources reduces the efficiency, and the efficiency increases for greater matrix size N. However, the best choice should be the maximum speed.

B. Results on AMD Opteron CPU

In this section we analyze the results of calculated Speed $V(P)$, Speedup $S(P)$ and Efficiency $(E(P)$ for each test case for all experiments executed on AMD Opteron processor.

1) Speed on AMD Opteron CPU: The speed for each test case of all four experiments while executed on AMD processor is depicted in Fig. 5. We observe that the speed for sequential execution behaves different than those for parallel execution.

We can model the $V(1)$ behavior in four regions, i.e. increasing, maximum, decreasing and saturating, while increasing the matrix size N. But very unexpected behavior appears for speed for parallel execution. There are two ranges $N = [81, 84]$ and $N = [217, 218]$ with local maximums. After the second maximum, the speed starts to decrease and then saturates.

Starting from $N = 325$, the speed is also discrepant for AMD CPU, and even more than Intel i7 CPU.

2) Speedup on AMD Opteron CPU: Fig. 6 depicts the achieved speedup for the three experiments with parallel implementation.

We determine a phenomenon for the speedup. Apart of the

TABLE III
SPEEDUP REGIONS AND MAXIMUM VALUES

Parameter	Exp. 2	Exp. 3	Exp. 4
Superlinear region	[435, 517]	[435, 657]	[435, 1100]
Max. speedup S_{max}	2.64	4.07	5.58
N_{max} (S_{max} point)	456	456	476

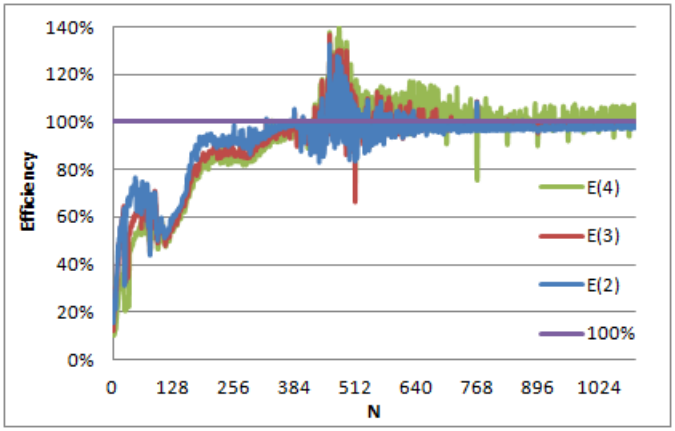


Fig. 7. Efficiency for MMM executed as a function of matrix size while executed on different number of AMD Opteron cores

Gustafson's Law that the maximum speedup is linear, we observe a superlinear speedup regions for each experiment with parallel implementation. Table III presents the regions and maximum values of the achieved speedup. All superlinear regions begin for the same matrix size $N = 435$, but they are greater when more cores are used. Even more, the superlinear speedup region for 4 cores is infinite (until measured $N = 1100$).

The same local maximums exist as for the speed since $V(1)$ is constant in those regions, and another local speedup maximum appears in superlinear region as presented in Table III.

We can conclude that maximum speedup can be achieved in the region around N_{max} for particular number of cores.

3) Efficiency on AMD Opteron CPU: The efficiency for the three experiments with parallel implementation is depicted in Fig. 7. Two interesting results are achieved. The first, we achieved a regions where the efficiency is greater than 1 (100%) because of achieved superlinear speedup. The second, there are two regions where the efficiency behaves differently. That is, the experiment 2 is the most efficient implementation in front of experiment 3 and 4, i.e. $E(2) > E(3) > E(4)$ for test cases in the region $N < 416$. For $N \geq 416$, the maximum efficiency is achieved for the experiment 4, in front of experiment 3 and 2, i.e. $E(4) > E(3) > E(2)$.

We can conclude that adding more resources reduces the efficiency, and the efficiency increases for greater matrix size N. However, the best choice should be the maximum speed.

V. DISCUSSION

The results show superlinear speedup region for MMM without any optimization on AMD Opteron processor using $P \in \{2, 3, 4\}$ cores.

Let's explain in more details the reasons that lead to superlinear speedup, and only on AMD Opteron CPU. Since both CPUs are on different frequency, we will use CPU clocks. The total number of clocks TC can be expressed as a sum of the clocks that CPU spends on arithmetic operations (CC) and the clocks for memory access (MC), i.e. $TC = CC + MC$. CC is constant for both CPUs while using the same number of cores and therefore it satisfies the Gustafson's Law. The sum of all clocks spent by all cores for parallel execution will be greater than the sequential due to extra operations required for creating the threads and CPU idle state because not all the cores start and finish the execution at the same time.

Let's analyze the MC now. Each matrix element should be accessed N times. The best performance can be achieved if each element is loaded from the main memory only the first time and other $N - 1$ times from L1 cache. The hypothesis set by many authors about superlinear speedup for matrix multiplication is that the reason for superlinear speedup is due to greater capacity of private cache memory per core in parallel execution will generate less cache misses than the sequential execution. However, our results show different reason for superlinear speedup.

We claim that more cache capacity in parallel execution is only one of the possible conditions for superlinear speedup achievement. In this paper, we present that the superlinear speedup appears when the problem size can not be stored completely in L3 cache, that is, it does not appear due to greater cache capacity in parallel execution, but the main factor for its existence is the shared L3 cache. This allows a kind of implicit prefetch of the matrix elements. This explains the larger speedup in the main memory region in the shared cache scenario, especially for the experiment with 4 cores. In parallel execution, when one of the cores loads a chunk of a matrix in the cache, it makes an implicit shared cache prefetch to other processors. This scenario is more important in a LRU cache replacement policy, where the matrix B takes more recent hits than the first one A. Therefore, the additional speedup than expected is achieved since the whole matrix B is shared among the cores.

Although we determined another possibility for superlinear speedup existence, it has been achieved only on AMD Opteron CPU, and not on Intel i7 processor, despite the same cache structure (the same number of cores, private L1 and L2 caches and shared L3 cache).

But, let's analyze in more details the speed for Intel i7. According to Gusev and Ristov [1] and the L3 cache size of i7 presented in Table I, L_3 region should be until $N = 724$ (two matrices can be stored in L3 cache size of 8MB). However, Fig. 2 depicts that the L_3 region ends around $N = 380$, which seems like L3 cache is not shared among all cores, but only a quarter of 8MB, i.e. 2MB is private per core. This could be the reason why superlinear speed is not achieved on Intel i7.

Despite the nonexistence of superlinear speedup on Intel i7, the speed ratio is much greater than the frequency of the AMD Opteron CPU. This implies that the Intel i7 processor achieves better speed than the AMD Opteron.

VI. Conclusion

This paper analyzes the MMM algorithm executed on two different vendor CPUs, i.e. Intel i7 and AMD Opteron, each with 4 cores with the same cache structure (private L1 and L2 caches per core, and shared L3 cache), but different cache capacities and set associativity. The three most important parameters for MMM are measured and analyzed: the speed, speedup and efficiency by varying the matrix size in order to determine the different behaviors of the algorithm for different cache regions: L_1, L_2, L_3 and L_4.

The speed as a function of matrix size has similar curves on Intel i7 CPU. All of them have three regions with different speed behavior: increasing, saturating and decreasing. Two regions are detected where the performance is discrepant. However, the achieved speed on AMD Opteron behaves different. Sequential execution is characterized with four regions: increasing, saturating, decreasing and again saturating. The speed for parallel implementations has two local maximums in the first saturating region for sequential execution. The speed is discrepant in the whole region for matrix size greater than 325 for AMD Opteron, and it is even greater than the discrepancy for Intel i7 CPU. The maximum speed for Opteron can be achieved at the end of L_2 region and for Intel i7 in L_3 region.

The speedup as a function of matrix size for each parallel execution complies with the Gustafson's Law for Intel i7. It is greater for greater matrix size and saturates towards linear speedup.

Superlinear speedup is achieved on AMD Opteron for all parallel executions. The main reason is not only due to greater cache capacity to store the matrix elements in parallel execution, because it appears in shared L_3 region instead of private L_2 region. We determine that superlinear speedup appear because the shared last level cache allows implicit prefetching of the matrix elements in parallel execution, i.e. when one of the cores loads a chunk of a matrix in the cache, it makes an implicit shared cache prefetch to other processors since they will access those elements in cache instead of main memory. Additionally, the superlinear region is wider for parallel execution with greater number of cores, being the infinite (in the analyzed region) while executed with 4 cores.

Both CPUs have different behavior for the efficiency also. Adding more resources on Intel i7 processor increases the speedup, but reduces the efficiency, i.e. the efficiency is greater when using smaller number of cores in parallel execution. The same holds for AMD Opteron but only in the left - sublinear region for smaller matrix size. Despite our hypothesis and Gustafson's Law, in superlinear region (for greater matrix size until the analyzed $N = 1100$) both the speedup and efficiency are saturated, but now adding the more resources will increase the efficiency as well.

VII. Future Work

Many unexpected results are observed: speed, speedup and efficiency discrepant regions, local maximums, speedup peaks, etc which will be the subject of our further research. In this paper we used the "slow" dense MMM without optimization to determine different vendor CPU performance on "slow" cache intensive algorithm. Our future work will be directed towards analyzing specific optimization techniques on different processor architectures, and their contribution of achieving even greater speeds and resolving specific bottlenecks. Moreover, other compute and cache intensive algorithms may be analyzed to achieve better performance on specific CPU cache architectures.

In this paper we determine the region where maximum performance (speed) is achieved and the best scaling (speedup and efficiency) on single chip multicore shared memory multiprocessor. We will try to model the performance on other multiprocessors like multichip multicore or multichip singlecore, also with different cache organization.

References

[1] M. Gusev and S. Ristov, "Matrix multiplication performance analysis in virtualized shared memory multiprocessor," in *MIPRO, 2012 Proceedings of the 35th International Convention, IEEE Conference Publications*, 2012, pp. 264–269.

[2] J. L. Hennessy and D. A. Patterson, *Computer Architecture, Fifth Edition: A Quantitative Approach*. MA, USA: Elsevier, 2012.

[3] M. Gusev and S. Ristov, "Performance gains and drawbacks using set associative cache," *Journal of Next Generation Information Technology (JNIT)*, vol. 3, no. 3, pp. 87–98, 31 Aug 2012.

[4] S. Williams, L. Oliker, R. Vuduc, J. Shalf, K. Yelick, and J. Demmel, "Optimization of sparse matrix-vector multiplication on emerging multicore platforms," *Parallel Comput.*, vol. 35, no. 3, pp. 178–194, 2009.

[5] D. J. Lee and T. J. Downar, "The application of posix threads and OpenMP to the U.S. nrc neutron kinetics code parcs," in *Proceedings of the International Workshop on OpenMP Applications and Tools: OpenMP Shared Memory Parallel Programming*, ser. WOMPAT '01. London, UK, UK: Springer-Verlag, 2001, pp. 90–100.

[6] R. Blikberg and T. Sørevik, "Nested parallelism: Allocation of threads to tasks and OpenMP implementation," *Sci. Program.*, vol. 9, no. 2,3, pp. 185–194, Aug. 2001.

[7] J. Al-Jaroodi, N. Mohamed, H. Jiang, and D. Swanson, "An agent-based infrastructure for parallel java on heterogeneous clusters," in *Proc. of the IEEE Int. Conf. on Cluster Computing*, ser. CLUSTER '02. Washington, DC, USA: IEEE Computer Society, 2002, pp. 19–19.

[8] M. Kolberg, G. Bohlender, and D. Claudio, "Improving the performance of a verified linear system solver using optimized libraries and parallel computation," in *High Performance Computing for Computational Science - VECPAR 2008*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 13–26.

[9] G. Goumas, K. Kourtis, N. Anastopoulos, V. Karakasis, and N. Koziris, "Performance evaluation of the sparse matrix-vector multiplication on modern architectures," *J. Supercomp.*, vol. 50, no. 1, pp. 36–77, 2009.

[10] X. Martorell, J. Labarta, N. Navarro, and E. Ayguad, "A library implementation of the nano-threads programming model," in *Euro-Par, Vol. II'96*, 1996, pp. 644–649.

[11] S. Ristov and M. Gusev, "Superlinear speedup for matrix multiplication," in *Information Technology Interfaces, Proceedings of the ITI 2012 34th International Conference on*, 2012, pp. 499–504.

[12] M. Gusev and S. Ristov, "Superlinear speedup in windows azure cloud," in *2012 IEEE 1st International Conference on Cloud Networking (CLOUDNET) (IEEE CloudNet'12)*, France, Nov 2012, pp. 173–175.

[13] D. P. Playne and K. A. Hawick, "Comparison of gpu architectures for asynchronous communication with finite-differencing applications." *Concurrency and Computation: Practice and Experience*, vol. 24, no. 1, pp. 73–83, 2012.

[14] S. Jenks, "Multithreading and thread migration using MPI and myrinet," in *Proceedings of the Parallel and Distributed Computing and Systems*, ser. PDCS'04, 2004.

[15] A. M. Castaldo and R. C. Whaley, "Scaling lapack panel operations using parallel cache assignment," *SIGPLAN Not.*, vol. 45, no. 5, pp. 223–232, Jan. 2010.

[16] J. L. Gustafson, "Reevaluating Amdahl's law," *Communication of ACM*, vol. 31, no. 5, pp. 532–533, May 1988.

Web service for separating components in the exponential decay process

Alan Martinović*, Damir Arbula and Željko Jeričević
*student's paper
Engineering Faculty/Department of Computer Engineering, Rijeka, Croatia
zeljko.jericevic@riteh.hr

Abstract - Exponential decay describes multitude of physical, chemical and biological processes. It is very often the case that the summary signal is measured and individual components have to be extracted by computation. Due to its universality, the problem attracted a lot of interest in the last fifty years from researchers working in various fields ranging from nuclear physics to biology. One of the first developments was in chemistry in order to analyze chemical kinetics data. The construction of web service for analysing exponential decay data is described. The service operates in two modes, one for registered, and another for non-registered users. Registered users can save the results including graphic output in odt file and ask the service to email results to them. Unregistered users can do all calculations and observe the results through web browser.

I. INTRODUCTION

Although the separation of exponentials looks deceptively simple, it is actually a difficult problem because of the non-orthogonality of exponentials. Attempts to separate exponentials with close half times by nonlinear least squares usually lead to a huge number of iterations with no convergence. Our service uses the least squares methods with linearization step based on numerical integration. Numerical integration for this type of signal has smoothing properties because the positive and negative errors cancel each other. It could be seen from the Fourier transform of numerical integration operator that the high frequencies are suppressed in the integration result. After the integration, solution of multi-exponential problem is obtained by solving over-determined system of linear equations followed by finding the roots of polynomial. The number of exponentials in the signal dictates the degree of polynomial, rank of the linear system and multiplicity of numerical integration. The advantage of accurate linearization with noise cancellation properties is that the exponentials separation becomes a one step procedure and condition number of the linear system can be used to control the quality of solution. The procedure can be completely generalized and apriori assumptions about the solution are not necessary, unless user wants to use them as constrains.

II. THEORY

The basic idea of the method is first proposed by Foss [1], and was later used by Matheson [2] in chemistry and Jericevic [3] in biology. In all above mentioned papers the method was developed and applied to special cases, no general solution was developed, although Foss claimed it is possible to construct one. The general solution was finally developed by Jericevic [4], and it does represent the foundation on which here described web application is based. The application consists of numerical kernel capable of separating arbitrary number of exponentials and user interface which will be described later.

The approach outlined here offers unprecedented flexibility addressing important problems of accurate and fast (real time) processing for multiple models (for arbitrary number of exponential terms) using the same software. Previous implementations required that specialized subroutines for each individual cases (mono-exponential, bi-exponential, tri-exponential, etc) have to be written..

This flexibility further leads to possibility of expanding the methodology by combining the signals from different measurements in order to achieve superior resolution of exponentials. In other words, the idea is to combine fluorescence decay curves from the same process at different wavelengths into single system as suggested in the global method by Knutson [5], but it is modified for out purposes because we are dealing with linear problem while Knutson applied it in nonlinear regression.

The major steps in our methodology are:

1. Linearization by numerical integration method
2. Solution of over-determined system of linear equations with or without non-negativity constrains [6]
3. Determining coefficients and roots of polynomial equation(s) [7] based on linear system solutions. Multiple approaches possible here, only the simplest and numerically most stable are mentioned [8]. This step yields non-linear parameters, i.e. decay constants.
4. Reformulating original problem using now known decay constants in order to compute pre-exponential terms. Two possible approaches here, one based on solving square system of linear equations, another based on over-determined system. The results coming from both approaches are in agreement.
5. Testing and verifying the results by detailed error analysis.

The detailed development of general solution is presented elsewhere in [4] and here we are giving only the final result:

General solution for $y(t)=\sum_{i=1}^{N} A_i e^{-k_i t}$

$$p_1 = -\sum_{i=1}^{N} k_i$$

$$p_n = -\sum_{i=1}^{\frac{N!}{(N-n)!n!}} \prod_{1}^{n} k_m \qquad m \in ({}^{N}C_n)$$

$$p_N = -\prod_{i=1}^{N} k_i$$

$$p_{N+1} = \sum_{i=1}^{N} A_i$$

$$p_{N+2} = \sum_{i=1}^{i=N} A_i \sum_{j=1;\, j\neq i}^{N} k_j$$

$$p_{N+n} = \frac{1}{(n-1)!} \sum_{i=1}^{N} A_i \sum_{j=1;}^{\frac{(N-1)!}{(N-n)!(n-1)!}} \prod_{1}^{n-1} k_m$$

$$m \in ({}^{N-1}C_{n-1};\ m \neq i)$$

$$p_{2N} = \frac{1}{(N-1)!} \sum_{i=1}^{N} A_i \prod_{j=1;\, j\neq i}^{N} k_j$$

III. MOTIVATION

The currently existing general solution algorithm developed by Jeričević [4] has only a CLI (*command line interface*) and must be run locally on Unix/Linux systems. The basic idea is to extend the interface of the application by using it as a computational core while having a public service for exponential regression via an easy to use web interface.

The web interface offers the possibility to upload the file with data that need to be exponentially fitted and specify the parameters of the fitting through a graphical interface after which results with basic details are generated. Additional details are available on the result if the user chooses to see them.

The web service operates in registered and anonymous mode. There are no difference between the fitting options and results between modes; the only advantage of the registered usage is that it gives an option to receive an .odt report on the chosen result through email.

IV. USE CASE

At first the user is presented some generic options about the service and an option to choose in which mode of operation he wants to work with. Both modes are briefly explained to him on the same page.

Next step depends on the user choice.

Users who have chosen to register are taken to the page where they are to submit registration information. The only required information is the email address the reports will be sent to. The registration is not permanent, and the user is forgotten once he stops using the service and has to register again with whatever information he chooses.

The next step is the same regardless of the working mode. It requires the user to submit the data in a form of a text file, and choose the parameters on for the fitting. Each parameter is briefly explained when hovered over the parameter name with a mouse pointer.

After correct data has been submitted, and all required parameters selected, the fitting takes place. Once it is over the results are displayed for the user to see and decide further action. If the user wants additional information, he clicks on the 'details' button supplied with each result. Additional information depend on the previously chosen parameters.

In registered mode, the user is given the additional option to have the detailed information of the result emailed to the address supplied during registration. The report is generated as a editable .odt file.

The use cases explained above can be visualized through the activity diagram below. [Figure 1 Use case activity diagram]

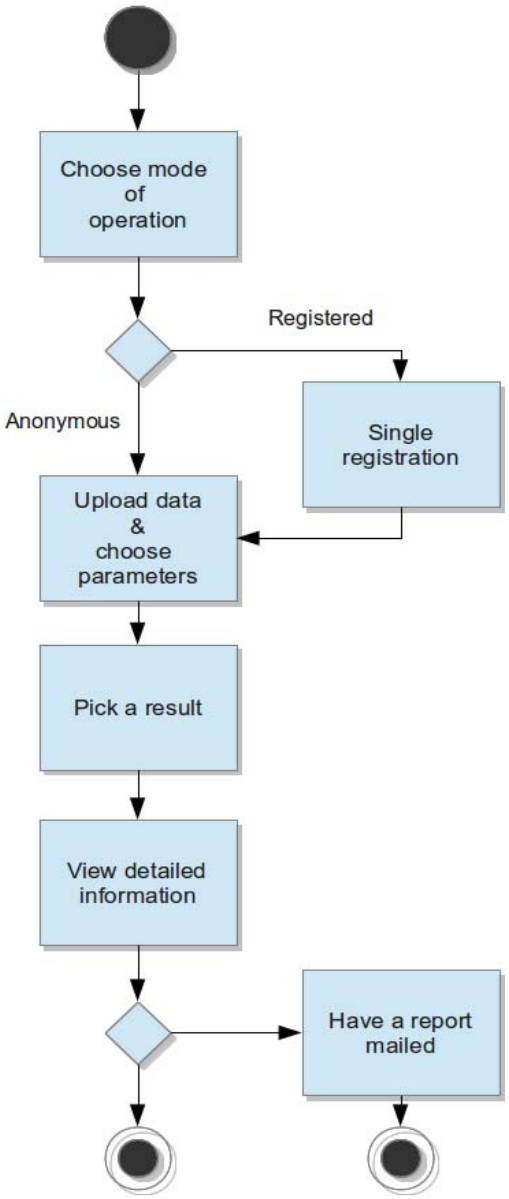


Figure 1 Use case activity diagram

V. ARCHITECTURE

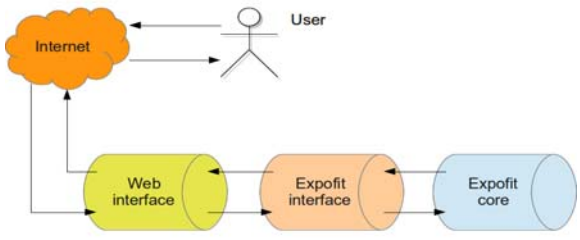


Figure 2 System architecture

A. Web interface

Renders HTML files that are served as the GUI (*Graphical User Interface*) to the final user through the browser. It is also handles all the web related activities, which include:

- session management
- user file uploads
- database management
- email sending
- interface styling

On the other side, it passes the user submitted information to the Expofit interface, from which, as a response, it receives result data that needs to be formatted an presented in a user friendly manner.

Technologies used:

- Django - a modular open source Python web framework

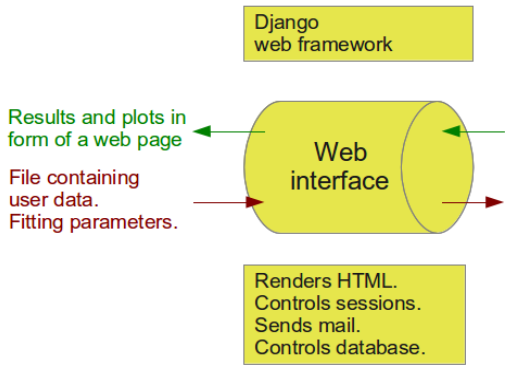


Figure 3 Web interface

B. Expofit interface

Receives the data and parameters that the user has submitted and prepare it for the Expofit core. That includes reformatting the data in the form the CLI expofit can understand. After preparation the data is sent to the Expofit core.

Once the Expofit core returns the results as raw text, Expofit interface structures them and forwards to the Web interface.

The structuring process includes:

- parsing raw data and forming an object structure from it
- generating plots
- generating .odt reports

Technologies used:

- Numpy – Python library for numerical computation
- Sympy – Python library for symbolical computation
- Matplotlib – Python library for data visualization

- Appy.pod Python library for generation of .odt reports

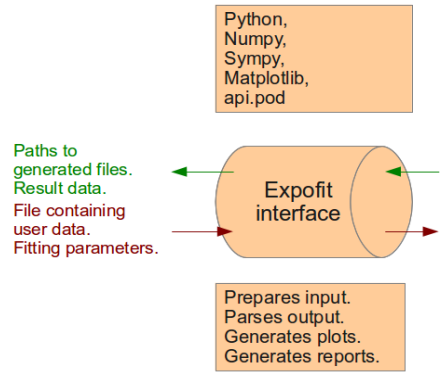


Figure 4 Expofit interface

1) Class diagram

Shown below [Figure 5 Class relation diagram**Error! Reference source not found.**] is the class diagram of the Expofit interface. Since there is work being done with lots of data and file paths, the architecture is formed around a Repo object (instance of the Repo class) which serves as a data center.

a) Repo

Contains all the data used in the process of receiving data from the Web interface, to returning structured data to it. There is a single instance per calculation, which is available to other objects.

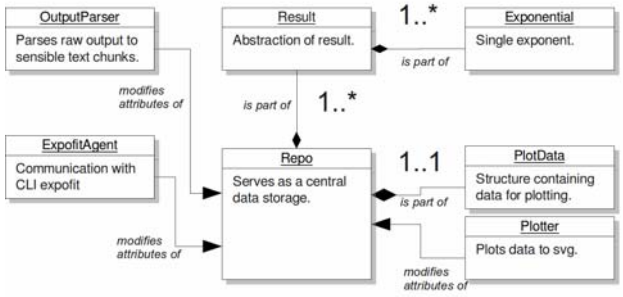


Figure 5 Class relation diagram

The data it contains:

- folder and file paths for both raw and processed data
- raw data and processed data
- processed data in form of objects

b) OutputParser

Processes the raw data received from the Expofit core to form data structures. This is achieved using regular expressions.

c) ExpoFitAgent

Provides the connection to the Expofit core, regardless of where it is located.

d) Plotter

Creates data suitable for plotting from the parsed results and creates .svg images.

e) Result

It receives a raw chunk of data containing a single result. Using its methods, it dissects the chunk and fills up its data structures and attributes, forming a fully populated object representation of the result received from the data chunk.

C. Expofit core

Expofit core is the component that performs the computation. The program input data, and gives back the parameters a_i and k_i of the result function

$$f(t) = C + \sum_{i=1}^{n} A_i e^{-k_i t} \qquad (1)$$

It can work as a standalone CLI application in which case it accepts the input file and parameters in the following form:

```
expofit    \
           input_type\
           [task]\
           [nnlstsq]\
           [weight]\
           [norm_col]\
           [norm_row]\
           [compute_errors]\
           [time_scale]\
           [print_flg]\
           [grh_out]\
           [comb]\
           < input_file
```

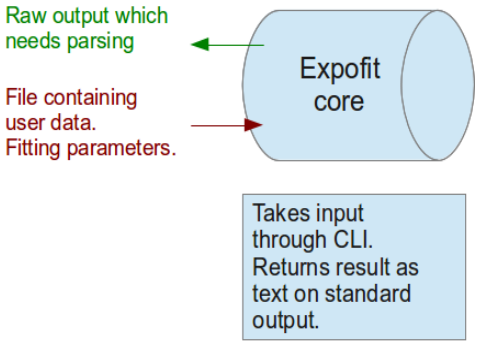


Figure 6 Expofit core

1) Input parameters

The parameters in square brackets [] are optional. Most of the parameters described here are inserted through the web interface via a HTML form. The parameters are presented in the following way:

Parameter name [value type], default value

a) Input type [INT], no default

The program can only fit the relaxation phase of the measured process. With the input type parameter we specify the starting point of the relaxation phase in the data file.

b) task [integer][0 or 1], default 51

Combination of two parameters *a* and *b*. Parameter *a* defines the number of exponentials which will be used to fit the data. It represents the *n* in (1**Error! Reference source not found.**. For positive *a* the program multiple results starting from for *n=1* an incrementing by one until *n=a*.

Parameter *b* defines if the existence of constant *C* should be considered in the calculation. For *b = 0* no constant is included in the result, for *b = 1* has a nonzero value which will be determined from the fitting.

c) nnlstsq [0 or 1], default 0

Defines if a_i and k_i can be non-negative values.

d) weight flg [0 or 1], default 0

Specifies if the input data is weighted. For weight flg = 1 the program will expect to have a weight parameter for every data row in the input data file.

e) norm col norm row, [C or R or 0], default norm col=C norm row=R

Defines if the state matrix will be normalized by rows or columns.

f) compute errors [0 or 1], default 0

Defines if approximation errors for a_i and k_i should be calculated.

g) time scale [nonzero non-negative float], default 1.

This parameter defines the time scale of the returned parameters. It is the factor which multiplies *t* in (1).If we want to change the units to minutes, we specify the time scale to be 1/60.

h) print flag [0 or 1], default 0

Defines if a verbose output of the fitting process should be printed. This parameter can't be set by the web interface.

i) grh out [0 or 1], default 1

Defines if the computed results should be plotted.

j) comb [0 or 1], default 1

Checks if there is enough available memory for the computation. This parameter can't be set by the web interface.

2) Input data file

The input file has to follow a form to be recognized [Table 1 Input file form] .

Depending on the input parameters, the data file should have from one to three columns of data.

If data is distributed in equal time steps, one column representing values is sufficient.

In case of irregular time intervals between data sampling, two columns are needed where one represents time and the other values. Since each data has its time, time_step has to be set to 0.

Three columns are used when the data is weighted. In that case the third column represents the weight of individual values. In this case the *weight flg* also has to be set.

TABLE 1 INPUT FILE FORM

number_of_points	starting_time	time_step
t_1	y_1	w_1
t_2	y_2	w_1
t_3	y_3	w_1
…	…	…
Required if time_step set to zero		Required if *weight flg=1*

VI. CONLUSION

The paper has shown the mechanisms, technologies and use cases of the expofit web service. We believe that the public availability of the service will serve as a valuable tool for research in science and engineering regardless of the field.

REFERENCES

[1] Foss, S.D., "A Method of Exponential Curve Fitting by Numerical Integration", Biometrics, 26 (1970) 815-821.

[2] Matheson, I.B.C., "The Method of Successive Integration: A General Technique for Recasting Kinetic Equations in a Readily Soluble Form Which is Linear in the Coefficients and Sufficiently Rapid for Real Time Instrumental Use", Anal. Instr., 16 (1987) 345-373.

[3] Jericevic, Z. et al., "Digital Imaging Fluorescence Microscopy: Statistical Analysis of Photobleaching and Passive Cellular Uptake Processes", Adv. Cell Biol., 3 (1990) 111-151.

[4] Jericevic, Z. "Method for Fitting a Sum of Exponentials to Experimental Data by Linearization Using Numerical Integration Approximation, and Its Application to Well Log Data", USP #7,088,097, 2006

[5] Knutson, J.R. et al, "Simultaneous Analysis of Multiple Fluorescence Decay Curves : A Global Approach", Chem. Phys. Lett., 102 (1983) 501-507.

[6] Lawson, C.L. and Hanson, R.J. "Solving Least Squares Problems", 1974, Prentice-Hall, Englewood Cliffs, pp340

[7] Morgan, A. "Solving Polynomial Systems Using Continuation for Engineering and Scientific Problems", 1987, Prentice-Hall, Englewood Cliffs, pp546

[8] Press, W.H. et all, "Numerical Recipes in C", 1994 Cambridge University Press, Cambridge, pp994

Performance Drawbacks for Matrix Multiplication using Set Associative Cache in GPU devices

Leonid Djinevski, Sime Arsenovski
FON University
1000 Skopje, Macedonia
Email: leonid.djinevski, sime.aresnovski@fon.edu.mk

Sasko Ristov and Marjan Gusev
Ss. Cyril and Methodius University
1000 Skopje, Macedonia
Email: sashko.ristov, marjan.gushev@finki.ukim.mk

***Abstract*—Performance of shared memory processors show negative performance impulses (drawbacks) in certain regions for execution of the basic matrix multiplication algorithm. In this paper we continue with analysis of GPU memory hierarchy and corresponding cache memory organization.**

We give a theoretical analysis why a negative performance impulse appears for specifics problem sizes. The main reason is the cache storage organization, i.e. the negative performance peak appears caused by mapping of matrix elements onto one cache set, instead of using the whole cache. The obtained experimental results prove our theoretical analysis. We also propose a method to avoid situations where performance drawbacks appear.

***Index Terms*—Cache Memory, SIMD, GPGPU.**

I. Introduction

Matrix multiplication is widely used algorithm in many computing applications. Its execution directly depends on cache memory architecture and organization. The performance depends on the following cache parameters: cache size, replacement policy, cache levels, cache-line size, cache inclusivity, cache associativity etc. Many hardware architectures exist for faster matrix multiplication algorithm execution, such as: supercomputers, grids, data-flow computing, cloud computing, GPUs, etc. Deeper understanding of their organization can lead to significant performance improvement.

GPU devices have recently provided a massive acceleration and together with their low cost have brought a significant processing power inside regular PCs. Their architecture is intended to maximize the throughput without concerning the latency of certain thread. GPUs are more appropriate for applications with regular data access patterns, and not for more complicated scatter/gather access patterns [1].

Three main cache types exist: direct map, fully associative or set associative. In this paper we focus on GPU architecture with its set associative cache memory storage pattern in order to understand the performance drawbacks for particular matrix size and to improve the matrix multiplication algorithm.

The rest of the paper is organized as follows. In Section II we give overview of related work in the area of the research problem. Section III briefly presents the GPU memory architecture. Theoretical analysis of possible performance drawbacks is presented in Section IV and a description of methodology used in the experiments in Section V. The results of the experiments are elaborated in Section VI. Finally, we conclude our work followed by recommendations in Section VII.

II. Related work

The matrix elements storage pattern has a strong impact on the matrix multiplication performance on GPU [2]. Likun and Dingfang [3] propose a mechanisms how to avoid cache race and cache split in order to improve the GPU GFlops. Tang et al. [4] optimize the cache locality on GPU modeling the cache miss analysis. However, they ignore the data reuse among the concurrent thread blocks on the SM. Volkov and Demmel [5] present detailed benchmarks of the GPU memory system, kernel start-up costs, and arithmetic throughput on dense matrix operations.

Cache set associativity can provide huge performance drawbacks in matrix multiplication for particular matrix size. Reactive mechanisms (selective displacement and feedback) [6] and way prediction [7] can improve set-associative cache access times. Greater set associativity will reduce the cache misses, but still will not improve the performance since this will increase the cache hit access time. Padding the first element the second matrix will amortizes the performance drawback due to cache associativity [8]. Hongil [9] selects dynamically an optimized replacement policy for each cache set via workload speculation mechanism to improve the cache performance.

Gusev and Ristov [10] proved both theoretically and experimentally that CPU Cache memory storage pattern can significantly reduce the performance of matrix multiplication increasing the generation of last level cache misses due to usage of set associative cache. Using their theorems one can determine the matrix sizes where maximum cache performance drawback in the matrix multiplication algoriothm will be generated due to matrix storage pattern in a n-way associative memory. In this paper we have used the theorems and experimentally proved that they hold for set associative cache in GPU architectures.

The latest GPUs have also two level cache hierarchy organized with set cache associativity. Matsumoto et. al [11] determine huge performance drawbacks of DGEMM for matrix size that are in multiples of 1024 without deeper explanation.

Two problems are exposed with usage of the caches, Cache capacity problem refers to the lack of the resources, while the cache associativity problem refers to inefficient usage of the cache. In this paper we are focused on performance analysis of the cache associativity problem.

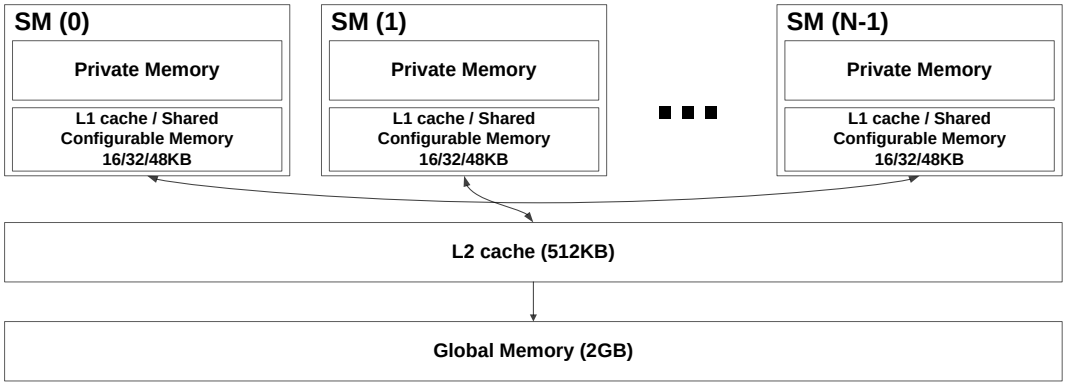


Fig. 1. Memory hierarchy of NVIDIA Kepler architecture

III. GPU Memory Architecture

The origins of General-Purpose computing on Graphics Processing Units (GPGPU) come from utilization of the enormous computational performance of the GPU devices, which was utilized for the first time using OpenGL or DirectX API calls [12]. Since this was very difficult and time consuming programming approach even for the small number of developers that were familiar with native graphics operations, Buck et al. [13] have developed Brook for GPUs which laid down the foundations for the NVIDIA's CUDA programming model [14]. CUDA programs are accelerated by data-parallel computations of millions of threads, which in this context mean instance of a kernel, where kernel is the program running on the GPU device. This approach characterizes the GPU architecture as a SIMD parallel machine [15] with convenient memory hierarchy.

Today GPUs have evolved into massively parallel, many-core architectures which contain an array of Streaming Multiprocessors (SM), each containing 8 Scalar Processors (SP) for the Tesla architecture [16], 32 SPs for the Fermi architecture [17], and 192 SPs for the latest Kepler architecture [18]. However, CUDA in particular is a Single Instruction Multiple Thread (SIMT) programming model [19], where all threads execute in step the same instruction, but within one SM. On the other hand, threads in different SMs are executing instructions independently from each other, thus providing scalability.

The memory hierarchy of NVIDIA Kepler GPU architecture is presented in Fig. 1. The last level in the hierarchy is the global memory, which is an off-chip memory and is the slowest. The first level in the memory hierarchy is 64KB, and its size is configurable by the L1 cache memory and shared memory during runtime, therefore obtaining 3 possible configurations for with the following sizes 16/32/48KB. These memories are shared by a number of threads organized in thread blocks. It can be accessed almost as fast as register memory, which is the private memory and is exclusive to a single thread.

Another off-chip memory is the L2 cache memory. It is shared to all threads in any SM. This memory is between the L1 cache memory and the global memory. Unlike the L1 cache memory, the size of the L2 cache memory is not configurable. However, the cache-line (cache block) size is configurable during compilation, with 64B or 128B.

There is a lack of information from the manufacturer when it comes to cache associativity of L1 and L2 cache memories. However, some micro-benchmarks like [5], [20], [21] contribute towards understanding the architecture of the GPU devices, where set-associative mapping is suggested for the cache associativity.

IV. Theoretical Analysis

In this section we present the matrix multiplication and its organization in the cache memory system. We also give a theoretical analysis why the negative performance peaks appear when using a n-way set associative cache and explain theoretically where we should expect negative performance peaks.

Table I presents the variable abbreviations used in this paper.

TABLE I
VARIABLE ABBREVIATIONS FOR BETTER PRESENTATION

Abbr.	Variable
CS	cache memory size in bytes
cbs	cache line (block) size in bytes
n	cache set associativity
S	total number of sets in associative cache
N	matrix size
ME	matrix element size
l	number of elements that can fit in a cache line
m	number of blocks necessary to store one matrix row
d	number of matrix rows that can fit in different sets

A. Matrix Multiplication Algorithm

We use matrices with equal matrix size $N \cdot N$ for simplicity. The result product matrix $C = [c_{ij}]$ is defined in

(1) by multiplying matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ where $i, j = 0, 1, \ldots N-1$.

$$C_{N \cdot N} = A_{N \cdot N} \cdot B_{N \cdot N}, \quad c_{ij} = \sum_{k=0}^{N-1} a_{ik} \cdot b_{kj} \tag{1}$$

The cache requirements for matrix multiplication algorithm are to store a row of matrix A and the whole matrix B in cache to enable efficient processing [22]. This research is not focused on the cache capacity problem, but on the cache associativity problem [23].

B. Storing the Matrix Elements in Set Associative Cache

This section presents an analysis of the organization of matrix elements in the cache. Efficient matrix multiplication algorithm where the cache capacity problem is not exposed will require such a cache organization where a complete row of matrix A will be stored in the cache along with whole column of matrix B.

Interesting phenomenon appears when loading the element a_{ik}. This action will result with transfer of the whole cache line (block) in the cache, i.e. the elements $a_{i,k+1}, a_{i,k+2} \ldots a_{i,k+l-1}$. Luckily, these elements will be used in the following operations and there will be benefit from early prefetching of the row elements of matrix A.

However, this is not the case for column elements of the matrix B. The processor loads the elements $b_{k,j+1}, b_{k,j+2} \ldots b_{k,j+l-1}$ instead of desired $b_{k+1,j}, b_{k+2,j}, \ldots$. This phenomenon will have to be used to efficiently realize matrix multiplication, i.e. to organize multiplications of the same row in matrix A with several consecutive columns from matrix C.

Matrix row elements will occupy consequently the cache blocks. However, the next matrix element will not occupy the next consecutive cache block, but a block which will be mapped by a specific modulo function used in the n-way set associative cache. This will produce the cache associativity problem, since some cache blocks will be mapped onto the same cache set. The worst case will happen when all cache blocks from a matrix column will map onto a same set and generate a lot of cache misses. The following section gives an overview of theoretical results about the cache associativity problem [10].

C. Performance drawbacks in a n-way set associative cache

In this section we use the theoretical analysis for performance drawbacks for matrix multiplication algorithm using a n-way set associative GPU cache, as generalization of a shared cache memory multiprocessor organization. Gusev and Ristov determine that maximum drawbacks for matrix multiplication algorithm appear for matrix size N if (2), (3), (4) and (5) are satisfied [10].

$$\exists d > 0, \qquad d = \frac{S \cdot cbs}{N \cdot ME} \quad and \quad d < \frac{N}{n} \tag{2}$$

$$N = \frac{CS}{n \cdot d \cdot ME} \tag{3}$$

$$\exists d > 0, \qquad d < \sqrt{\frac{S \cdot cbs}{n \cdot ME}} \tag{4}$$

$$N > \sqrt{\frac{S \cdot cbs \cdot n}{ME}} \tag{5}$$

In this research we were eager to check if those relations are also valid for GPU cache architecture and bring relevant conclusions. The next section presents theoretical results about this phenomenon.

D. Performance peaks in GPUs using n-way set associative cache

Using (2), (3), (4) and (5) for L2 cache, we determine that maximum performance drawbacks for matrix multiplication algorithm execution will be achieved for $d = 1, 2, 4, 8$ and 16, i.e. for $N = 8192, 4096, 2048, 1024, 512$.

Table II presents the cache parameters of the GPU model GeForce GTX 680.

TABLE II
CACHE VARIABLES FOR THE EXPERIMENTS

Variable	Value in L1	Value in L2
CS	32KB	512KB
n	4	16
ME	4B	4B
cbs	128B	128B
l	32	32
S	642	256
$N \cdot d$	2048	8192

Table III presents the results of calculations for the GPUs presented in Table II. There we can observe the conditions whether performance drawbacks will appear for critical values of N (as defined in (2)).

TABLE III
CONDITIONS FOR PERFORMANCE DRAWBACKS

N	d	N/d	n
64	128	0.5	16
128	64	2	16
256	32	8	16
512	16	32	16
1024	8	128	16
2048	4	512	16
4096	2	2048	16
8192	1	8192	16

Analyzing these results we can conclude that performance drawbacks will appear for $N = 512, 1024, 2048, 4096$ and 8192. This paper aims to confirm these theoretical results by experimental research.

V. TESTING METHODOLOGY

This section describes the testing methodology based on a series of experiments realized in order to determine the performance drawbacks due to cache set associativity problem.

All experiments were executed on the hardware infrastructure presented in Table IV installed with Ubuntu 12.04 LTS operating system. The implementations were compiled with the NVIDIA compiler nvcc from the CUDA 5.0 toolkit.

We choose the configuration of 32KB L1 cache memory and 128B for the cache-line size. Additionally, we assume that the cache memories are set-associative, such that the number of sets $S = 2^k$, where k is integer.

We execute 6 experiments of sequential matrix multiplication algorithm implementation described in Section IV-A in the regions around the critical matrix sizes: $N = 64, 128, 256, 512, 1024$ and 2048. The sequential implementation runs one thread per only one active SM [24]. Therefore, the whole L1 cache is dedicated to the thread.

Each experiment consists of 10 test cases for problems in the area around the critical points, for example, for five values of N, which are -10, -8, -6, -4, and -2 below the expected N, where the drawback is expected and for five values of n, which are +2, +4, +6, +8, and +10 above the expected N, where drawback is expected.

Each test case is executed 11 times. We exclude the first iteration and we measure the average execution time of the rest 10 iterations in order to achieve reliable results of each test case.

VI. Results of the Experiments

This section presents the results of the experiments on the GPU with cache parameters presented in Table II. Our focus are the areas around the problem size for the main drawbacks points. In the beginning we will compute the values where L1 and L2 caches will have enough capacity to store the matrices. Table V defines these values and determines the L_1, L_2 and L_3 regions, as region where matrix size N will fit in cache L1, L2 and main memory correspondingly [25].

Figure 2 depicts the speed expressed in MFLOPS for execution of the matrix multiplication algorithm within the defined L_1, L_2 and L_3 regions [25].

A. Experiment 1 - Range around $N = 64$

Experiment 1 covers the area of $N = 64$ depicted in Fig. 3. Matrices A and B can be stored completely in L1 cache.

As Table III shows, the performance drawbacks are not observed in this region since (2) is not satisfied, i.e. and the elements of a column of matrix B will not spill out of a particular set. Therfore, the speed in this region has positive trendline.

TABLE IV
Hardware used for the performed tests and experiments

Device	Model
CPU	Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz
RAM	Kingston 32GB @ 1.60GHz
GPU	NVIDIA GeForce GTX 680

TABLE V
Determination of L1, L2 and L3 regions

max N to fit in L1	max N to fit in L2
50	350

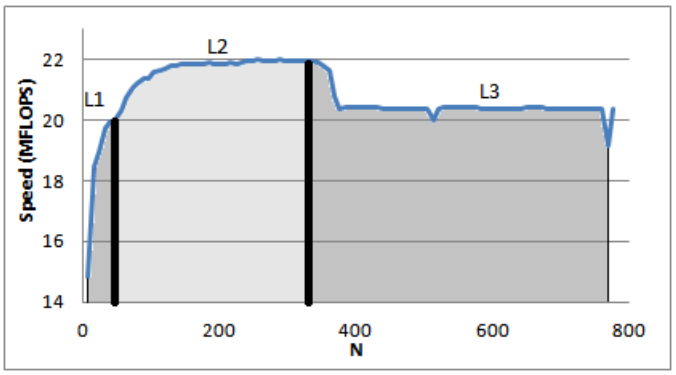


Fig. 2. Speed and regions L_1, L_2 and L_3 for matrix multiplication algorithm

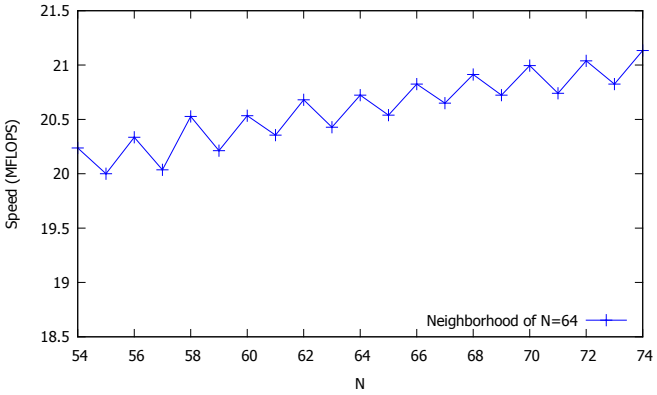


Fig. 3. Speed in the area around $N = 64$

B. Experiment 2 - Range around $N = 128$

Experiment 2 refers to the area of $N = 128$ depicted in Fig. 4. Matrix B cannot be stored completely in L1 cache and drawbacks appear due to insufficient L1 cache and L2 cache set associativity.

The results also prove the theoretical analysis summirized in Table III, i.e. $n > N/d$. Speed in this region also has a positive trendline, but lower than Experiment 1.

C. Experiment 3 - Range around $N = 256$

Experiment 3 analyses the area of $N = 256$, which is in L_2 region. Matrix B also cannot be stored completely in L1 cache and drawbacks appear due to insufficient L1 cache and L2 cache set associativity. The speed in this region has also a positive trend, as depicted in Fig. 5.

Performance drawbacks are not observed in this region as presented in Table III since still $n > N/d$ is satisfied. Additionally, we observe that the performance for $N = 256$ is even greater than near values for N in that region. We explain this observation with the fact that the whole matrix row can

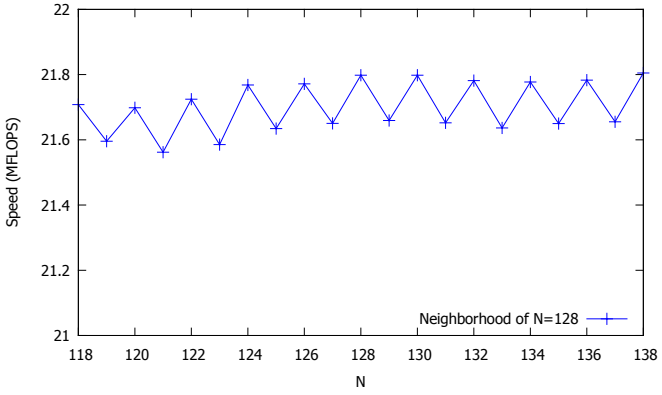


Fig. 4. Speed in the area around $N = 128$

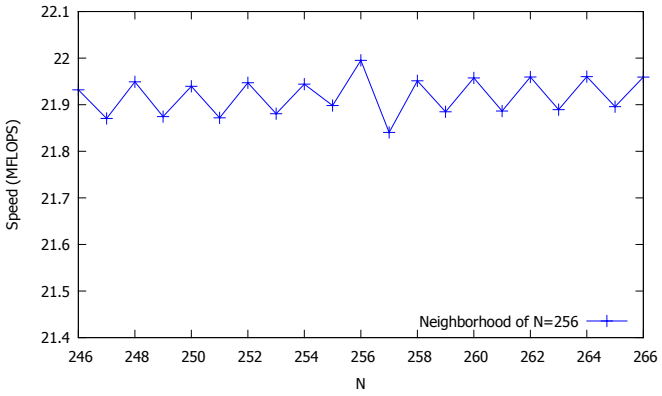


Fig. 5. Speed in the area around $N = 256$

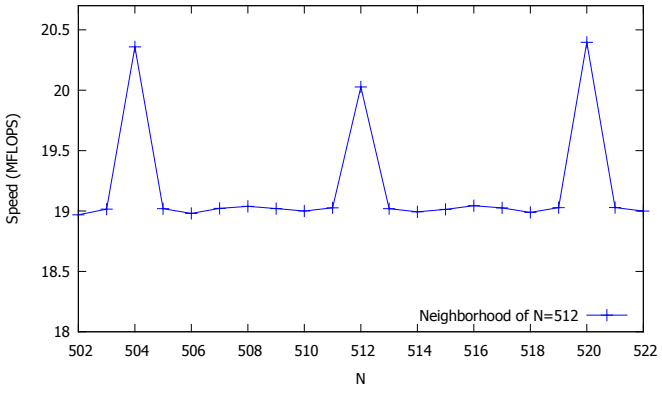


Fig. 6. Speed in the area around $N = 512$

be stored in the exact number of cache blocks. Therefore, the average access time is reduced for a matrix element stored in particular cache block.

D. Experiment 4 - Range around $N = 512$

Experiment 4 covers the area around $N = 512$ depicted in Fig. 6. Matrix B cannot be stored completely neither in L1 nor L2 cache, and thus drawback is mainly due to their size and associativity.

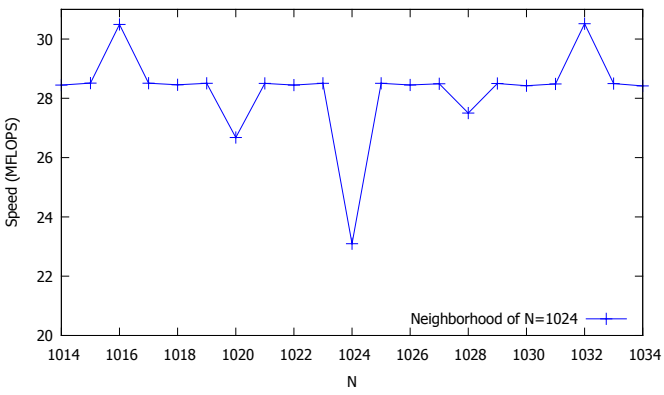


Fig. 7. Speed in the area around $N = 1024$

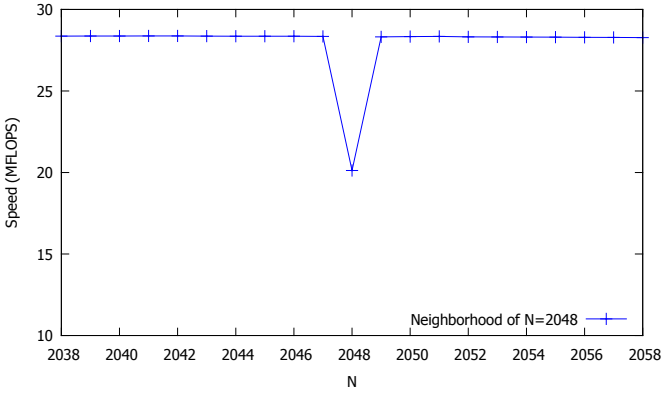


Fig. 8. Speed in the area around $N = 2048$

We observe interesting results in this region. The speed for $N = 512$ is greater than the other problem sizes in the region, except for $N = 504$ and $N = 520$. $N = 512$ has performance drawback compared to these points. The observation consists of two parts: greater speed and performance drawbacks. The former appears for the same reason as explained for $N = 256$. The latter appears compared due to cache associativity and condition of Table III, compared to points $N = 504$ and $N = 520$.

E. Experiment 5 - Range around $N = 1024$

Experiment 5 analyses the area around $N = 1024$. Matrix B cannot be stored completely in L2 cache and drawbacks appear due to L2 cache size and associativity. Fig. 7 presents the speed drawbacks.

Significant performance drawback appears as stated in Table III, but also positive peaks are observed in the points $N = 8$ and $N - 8$ as in the region around $N = 512$.

F. Experiment 6 - Range around $N = 2048$

Experiment 6 refers the area of $N = 2048$ and matrix B cannot be stored completely in L2 cache. Drawback is shown for speed in Fig. 8.

VII. Discussion and Conclusions

Set associative cache in GPU can seriously degrade the performance. In this paper we provide theoretical analysis why performance drawbacks appear for specific problem sizes of matrix multiplication algorithm. We have realized series of experiments in the areas of critical problem sizes which prove the theoretical analysis.

A diagram of achieved speed for essential matrix multiplication is presented in Fig. 2. Six experiments were described in details for points where theoretical results show expectation of negative performance peaks with analysis of speed diagrams presented in Fig. 3 - Fig. 8.

The results show that significant performance drawbacks appear for $N = 1024$ and 2048 due to L2 cache set associativity. Performance drawbacks are not observed for other critical problem sizes $N = 64, 128$ and 256 as expected. Cache set size is enough to fit the cache storage requirements for these problem sizes. The performance for $N = 256$ is even greater than near values for N in that region.

Significant negative performance impulses appear as stated in Table III. Neither positive, nor negative peaks are observed in the other points of the observed regions.

Interesting results are observed in the region around $N = 512$, i.e. the speed for $N = 512$ is greater than the other problem sizes in the region, except for $N = 504$ and $N = 520$. $N = 512$ has performance drawback compared to these points. Future work will cover further research on this phenomenon.

References

[1] D. Tarjan, J. Meng, and K. Skadron, "Increasing memory miss tolerance for simd cores," in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, ser. SC '09. New York, NY, USA: ACM, 2009, pp. 22:1–22:11.

[2] A.-K. Cheik Ahamed and F. Magoules, "Iterative methods for sparse linear systems on graphics processing unit," in *High Performance Computing and Communication 2012 IEEE 9th International Conference on Embedded Software and Systems (HPCC-ICESS), 2012 IEEE 14th International Conference on*, june 2012, pp. 836 –842.

[3] Z. Likun and C. Dingfang, "Accelerate your graphic program with gpu/cpu cache," in *Proceedings of the 2008 International Conference on Cyberworlds*, ser. CW '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 667–671.

[4] T. Tang, X. Yang, and Y. Lin, "Cache miss analysis for gpu programs based on stack distance profile," in *Proceedings of the 2011 31st International Conference on Distributed Computing Systems*, ser. ICDCS '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 623–634.

[5] V. Volkov and J. Demmel, "Benchmarking gpus to tune dense linear algebra," in *Proceedings of the 2008 ACM/IEEE conference on Supercomputing*. IEEE Press, 2008, p. 31.

[6] B. Batson and T. N. Vijaykumar, "Reactive-associative caches," in *Proceedings of the 2001 International Conference on Parallel Architectures and Compilation Techniques*, ser. PACT '01. Washington, DC, USA: IEEE Computer Society, 2001, pp. 49–60.

[7] B. Calder, D. Grunwald, and J. Emer, "Predictive sequential associative cache," in *Proceedings of the 2nd IEEE Symposium on High-Performance Computer Architecture*, ser. HPCA '96. Washington, DC, USA: IEEE Computer Society, 1996, pp. 244–. [Online]. Available: http://dl.acm.org/citation.cfm?id=525424.822662

[8] S. Williams, L. Oliker, R. Vuduc, J. Shalf, K. Yelick, and J. Demmel, "Optimization of sparse matrix-vector multiplication on emerging multicore platforms," *Parallel Comput.*, vol. 35, no. 3, pp. 178–194, Mar. 2009.

[9] H. Yoon, T. Zhang, and M. H. Lipasti, "Sip: Speculative insertion policy for high performance caching," Computer Sciences Department University of Wisconsin-Madison, Tech. Rep. 1676, 2010.

[10] M. Gusev and S. Ristov, "Performance gains and drawbacks using set associative cache," *Journal of Next Generation Information Technology (JNIT)*, vol. 3, no. 3, pp. 87–98, 31 Aug 2012.

[11] K. Matsumoto, N. Nakasato, and S. Sedukhin, "Implementing a code generator for fast matrix multiplication in opencl on the gpu," in *Embedded Multicore Socs (MCSoC), 2012 IEEE 6th International Symposium on*, sept. 2012, pp. 198 –204.

[12] D. Kirk, W. Wen-mei, and W. Hwu, "Programming massively parallel processors: a hands-on approach," USA, 2010.

[13] I. Buck, T. Foley, D. Horn, J. Sugerman, K. Fatahalian, M. Houston, and P. Hanrahan, "Brook for gpus: stream computing on graphics hardware," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 777–786, Aug. 2004.

[14] NVIDIA. (2013, Jan.) Cuda programming guide. [Online]. Available: http://developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/CUDA_C_Programming_Guide.pdf

[15] A. Grama, G. Karypis, V. Kumar, and A. Gupta, *Introduction to Parallel Computing (2nd Edition)*, 2nd ed. Addison Wesley, Jan. 2003.

[16] E. Lindholm, J. Nickolls, S. Oberman, and J. Montrym, "Nvidia tesla: A unified graphics and computing architecture," *IEEE Micro*, vol. 28, no. 2, pp. 39–55, Mar. 2008.

[17] P. Glaskowsky, "Nvidias fermi: the first complete gpu computing architecture," NVIDIA, Tech. Rep., 2009, white Paper.

[18] NVIDIA, "Next generation cuda compute architecture: Kepler gk110," 2012.

[19] J. Nickolls and W. Dally, "The gpu computing era," *IEEE Micro*, vol. 30, no. 2, pp. 56–69, 2010.

[20] M. Papadopoulou, M. Sadooghi-Alvandi, and H. Wong, "Micro-benchmarking the gt200 gpu," *Computer Group, ECE, University of Toronto, Tech. Rep*, 2009.

[21] R. Meltzer and C. Zeng. (2013, Jan.) Micro-benchmarking the c2070. [Online]. Available: http://people.seas.harvard.edu/~zeng/microbenchmarking/

[22] J. L. Hennessy and D. A. Patterson, *Computer Architecture, Fifth Edition: A Quantitative Approach*. MA, USA: Elsevier, 2012.

[23] S. Ristov and M. Gusev, "Achieving maximum performance for matrix multiplication using set associative cache," in *Computing Technology and Information Management (ICCM2012), 2012 The 8th International Conference on*, ser. ICNIT '12, vol. 2, 2012, pp. 542–547.

[24] L. Djinevski, S. Ristov, and M. Gusev, "Superlinear speedup in gpu devices," in *ICT Innovations 2012*. Springer Berlin / Heidelberg, 2012.

[25] M. Gusev and S. Ristov, "Matrix multiplication performance analysis in virtualized shared memory multiprocessor," in *MIPRO, 2012 Proceedings of the 35th International Convention, IEEE Conference Publications*, 2012, pp. 264–269.

Defending the need for a new global software approach: a literature review

M. Trkman, S. Vrhovec, D. Vavpotič and M. Krisper
Faculty for computer in information science, University of Ljubljana, Ljubljana, Slovenia
{marina.trkman, simon.vrhovec, damjan.vavpotic, marjan.krisper}@fri.uni-lj.si

Abstract – Software engineering has been affected by the globalization trend. Software companies embracing global software engineering can benefit from it by wisely approaching its challenges. This paper gives a literature review on general GSD's challenges by examining how the focus changed in the last decade. There exist some global software development approaches that focus on overcoming these challenges. For example, the loosely coupled team approach helps to overcome the coordination issues that the virtual teams have to deal with by modularization of development work. However, none of these approaches is focused on supporting the development project with scattered individual developers.

I. INTRODUCTION

Global collaboration and the ways to improve its results are of increased importance [17]. Global Software Engineering (GSE) is characterized by a highly distributed environment which tends to create specific challenges [4, 14] of design, development, operation and maintainance of the software. Interestingly, many challenges which emerged in 2001, re-emerged in 2006 and again in the recent years. A contribution of this paper is a literature review on those evergreen challenges.

The structure of the paper is as follows. First, we introduce Global Software Development (GSD) and its challenges. Next, we discuss why many challenges present in 2001 are still a problem today. We conclude the paper with directions for new GSD approaches affected by the stattered individual developers model.

II. GSD APPROACH AND ITS VARIATIONS

GSD is an approach for outsourcing local in-house software development to developers at remote sites . Software companies choose it for several potential benefits : 1) reduced development cost, 2) reduced development duration, 3) new opportunities for cross-site modularization of development work, 4) better-skilled developer pool, 5) greater innovation, lerning and transfer of best practices, 6) closer proximity to markets and customers.

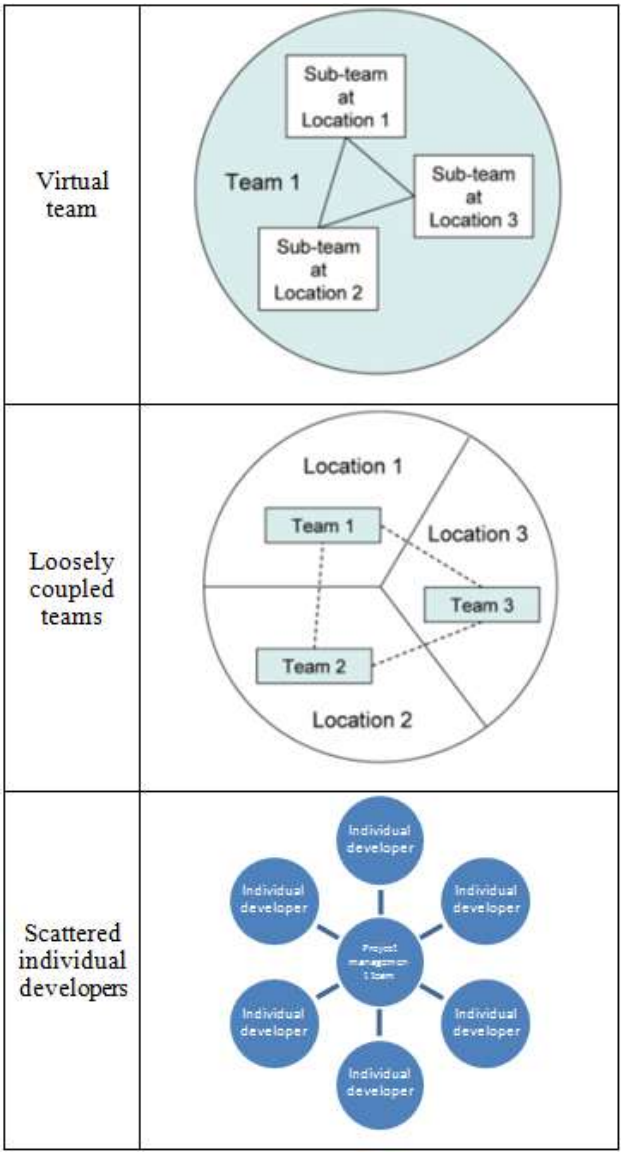


Figure 1. GSD models of possible team structures in GSD (adapted from [4])

Conchúir et al. [4] described various GSD approaches regarding the modularization of work (see Figure 1). First, the virtual team model treats all team members as members of one large team even though they may be separated by long distances. Second, the loosely-coupled teams model treats one set of co-located colleagues as one team. This model promotes modularization of tasks by feature. In this model, the granularity of modularization is

very important. The model only works if the development tasks are not too small to be meaningfully distributed.

In the next section, we describe and discuss several risks associated with GSD approaches [11]. For the project's success, it is essential for the practitioners to decide how to deal with them.

III. GSD CHALLENGES – A LITERATURE REVIEW

Cultural differences are frequently appearing as one of the key challenges in GSD. Different corporate cultures have diverse working habits, norms, values, patterns of behavior, work ethics, types of communication, quality standards, terminology, types of hierarchy, etc. To alleviate the negative effects of cultural distance between global development sites, two prominent strategies have emerged. One strategy promotes the use of the same processes, methodology and terminology on all development sites. Standardization of project-related artifacts is vital in order to achieve this [8]. However, some authors argue that this strategy may not be appropriate in GSD except under crisis-management circumstances [7]. As mentioned, corporate culture defines what is acceptable at some development site. Cultural differences between them in turn impede the ability of software companies to swiftly transfer best practices between development sites [4]. Following these claims, the other strategy advocates learning about different cultures and accepting them. Development sites have open hands regarding the way they organize their work as long as they provide the expected results.

Coordination complexity increases when developers are distributed [4]. It is significantly affected by temporal, geographic and cultural distance between development sites [18]. The base development site could quickly become overloaded with work while remote development sites could remain nearly idle [17]. To some extent, the absence of personal contact can be replaced by formalization and discipline [12]. A formal definition of roles and their responsibilities at the start of the project is essential [2]. Even when development sites organize their work according to their corporate culture, it is important to synchronize the inter-organizational processes at cooperating development sites. These processes can be aligned by defining common work products at the start of the project. For example, a common vocabulary and templates for requirements specifications, test scenarios, and functional designs can provide project managers with an easier insight into work progress at different development sites [5].

Centralization, the use of tools and a reliable infrastructure are also beneficial. It is a common standard in software development to use a central repository and a bug-tracking tool. In addition, documentation, reports, templates, and tests can also facilitate coordination [13]. Good documentation is vital for GSD. In contrast to collocated development, it is essential that the documentation is current, accurate, and more detailed. This minimizes ambiguity and misunderstandings that may arise due to inhibited communication channels [10]. Communication is hindered by several issues. The most evident issues are language, terminology and culture. It is essential for the development sites to agree on a language that everybody satisfactorily understands before the project begins [3], [6].

This may seem obvious in collocated development. However, it can become quite a challenge to assure the use of the same terminology in a globally distributed environment, especially when changes occur during the project [8]. Differences in corporate cultures, such as different communication patterns and hierarchy types, might spur unnecessary conflicts and misunderstandings [5]. It is helpful to know how developers on other sites communicate and react in certain circumstances.

The lack of spontaneous and informal **communication** may inhibit the ability to establish good relationships and trust. It is difficult for people to trust someone they have never personally met. Therefore, several authors recommend face-to-face meetings [1]. These meetings may however be unaffordable for software companies in transition economies. In agile software development, informal communication is considered more useful than formal communication [12]. However, the opportunities for informal and spontaneous communication across development sites are rare [19]. In addition, this communication is often less effective due to cultural differences, language barriers, and organizational boundaries [12]. To alleviate these drawbacks, formal and rich communication channels should be defined at the start of the project [10]. Furthermore, peer-to-peer communication on various levels, such as management, project, and team, is encouraged [5].

A. Statistics of the litterature review

In 2001, IEEE Software published a special issue on GSD. In Table 1 we present relevant GSD challenges that were exposed.

TABLE I. NUMBER OF TREATEMENTS FOR EACH GSD'S CHALLENGE IN YEAR 2001

GSD challenge, 2001	**No.**	**References**
Communication	5x	[10], [3], [9], [1], [13]
Coordination	6x	[10], [3], [9], [1], [13], [8]
Geo-location	3x	[1], [3], [9]
Relationships & trust	3x	[1], [9], [13]
Job security / resistance	2x	[8], [10]
Culture	4x	[1], [8], [9], [10]
Control	1x	[3]
Language	1x	[8]
Knowledge dissemination	2x	[9], [10]
Informal communication	2x	[9], [10]
Tools	1x	[10]
Time zones	2x	[1], [3]
Indirect communication	0x	
Risk assessment	2x	[9], [13]

In 2006, there were two special issues on GSD. The first was published by IEEE Software and the second by Communications of the ACM. Table 2 shows the challenges focused in these special issues.

TABLE II. NUMBER OF TREATEMENTS FOR EACH GSD'S CHALLENGE IN YEARS 2007-2012

GSD challenge, 2006	**No.**	**References**
Communication	1x	[12]
Coordination	0x	
Geo-location	2x	[6], [16]
Relationships & trust	1x	[16]
Job security / resistance	0x	
Culture	3x	[16], [6], [7]
Control	0x	
Language	1x	[6]
Knowledge dissemination	2x	[7], [12]
Informal communication	0x	
Tools	1x	[7]
Time zones	2x	[6], [16]
Indirect communication	1x	[2]
Risk assessment	0x	

TABLE III. NUMBER OF TREATEMENTS FOR EACH GSD'S CHALLENGE IN YEAR 2006

GSD challenge, 2007-2012	**No.**	**References**
Communication	2x	[4], [5]
Coordination	3x	[17], [4], [5]
Geo-location	3x	[15], [17], [4]
Relationships & trust	1x	[5]
Job security / resistance	2x	[15], [4]
Culture	4x	[15], [17], [4], [5]
Control	1x	[4]
Language	1x	[15]
Knowledge dissemination	2x	[4], [5]
Informal communication	0x	
Tools	1x	[5]
Time zones	3x	[15], [17], [4]
Indirect communication	1x	[5]
Risk assessment	0x	

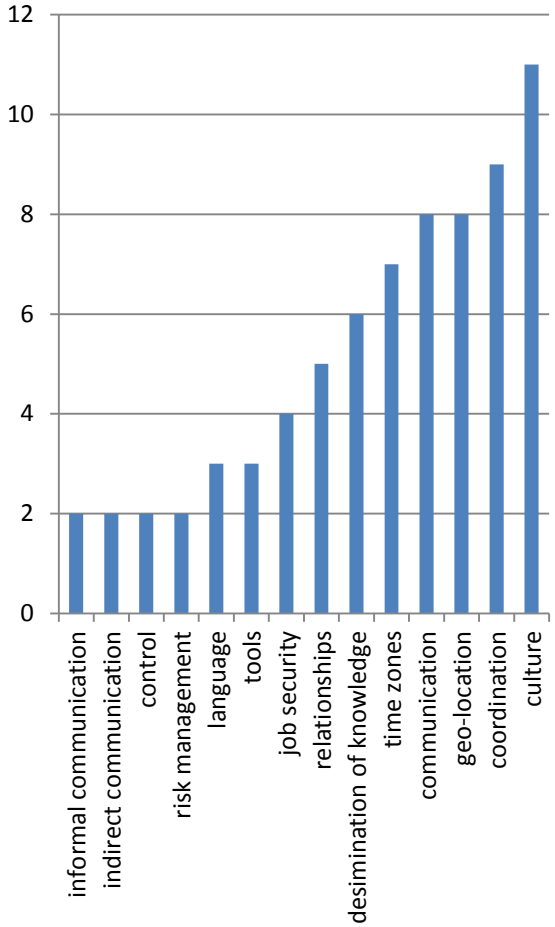


Figure 2. Number of occurances for each GSD challenge from 2001 till 2012

IV. DISCUSSION AND CONCLUSION

In 2001, there was a lot of discussion about GSD challenges, especially about coordination, communication, and culture (see Table I). In 2006, culture, geographical location, and knowledge dissemination were just a few of authors' interest. In the following years, we can see a growing trend of solving the issues that appeared in 2001. For example, there was a big focus on solving coordination issues in 2001. These issues later disappeared from authors' focus. However, they later re-emerged (see Table III). We can observe similar focus switching with communication and cultural issues.

In addition to the two GSD approaches discussed above (the virtual team model and the loosely-coupled teams model), a model focused on supporting the coordination of scattered individual developers is beeing called for.

The graph on Figure 2 shows the overall popularity of GSD's challenges. By far the most popular discussion in this period is about the challenge of the culture, which is followed by the coordination etc.

This paper provided a literature review on GSD's challenges by examining the shifts in research focus in the last decade. Even though some issues seem to have disappeared from reserchers' focus, they are still present today. They might only be hiding under different terminology.

REFERENCES

[1] Battin, R. D., Crocker, R., Kreidler, J., Subramanian, K. Leveraging resources in global software development. IEEE Software. 2001. 18 (2). pp. 70-77.

[2] Bhat, J. M., Gupta, M., Murthy, S. N. Overcoming Requirements Engineering Challenges: Lessons from Offshore Outsourcing. IEEE Software. 2006. 23 (5). pp. 38-44.

[3] Carmel, E., Agarwal, R. Tactical approaches for alleviating distance in global software development. IEEE Software. 2001. 18 (2).

[4] Conchúir, E. Ó., Ågerfalk, P. J., Olsson, H. H., Fitzgerald, B. Global software development: where are the benefits? IEEE Software. 2009. 52 (8). pp. 127-131.

[5] Damian, D. Stakeholders in Global Requirements Engineering: Lessons Learned from Practice. IEEE Software. 2007. 24 (2). pp. 21-27.

[6] Damian, D., & Moitra, D. Guest Editors' Introduction: Global Software Development: How Far Have We Come? IEEE Software. 2006. 23 (5).

[7] Desouza, K. C., Awazu, Y., Baloh, P. Managing Knowledge in Global Software Development Efforts: Issues and Practices. IEEE Software. 2006. 23 (5). pp. 17-19.

[8] Ebert, C., DeNeve, P. Surviving global software development. IEEE Software. 2001. 18 (2). pp. 62-69.

[9] Heeks, R., Krishna, S., Nicholsen, B., Sahay, S. Synching or sinking: global software outsourcing relationships. IEEE Software. 2001. 18 (2). pp. 54-60.

[10] Herbsleb, J. D., Moitra, D. Global software development. IEEE Software. 2001. 18 (2). pp. 16-20.

[11] Lacity, M. C., Khan, S., Yan, A., Willcocks, L. P. A review of the IT outsourcing empirical literature and future research directions. Journal of Information Technology. 2010. 25 (4). pp. 395-433.

[12] Lee, G., DeLone, W. H., Espinosa, J. A. Ambidextrous Coping Strategies in Globally Distributed Software Development Projects. Communications of the ACM. 2006. 49 (10). pp. 35-40.

[13] Mockus, A., Weiss, D. M. Globalization by chunking: a quantitative approach. IEEE Software. 2001. 18 (2). pp. 30-37.

[14] Portillo-Rodríguez, J., Vizcaíno, A., Piattini, M., Beecham, S. Tools used in Global Software Engineering: A systematic mapping review. Information and software technology. 2012. 54 (7). pp. 663-685.

[15] Richardson, I., Paulish, D., Verner, J. Research directions for global software engineering-where to next? Journal of Software: Evolution and Process. 2012. 24 (2). pp. 115-117.

[16] Sinha, V., Sengupta, B., Chandra, S. Enabling Collaboration in Distributed Requirements Management. IEEE Software. 2006. 23 (5). pp. 52-61.

[17] Šmite, D., Wohlin, C. A Whisper of Evidence in Global Software Engineering. IEEE Software. 2011. 28 (4). pp. 15-18.

[18] Šmite, D., Wohlin, C., Gorschek, T., Feldt, R. Empirical evidence in global software engineering: a systematic review. Empirical Software Engineering. 2010. 15 (1). pp. 91-118.

[19] Vlaar, P. W. L., Van Fenema, P. C., Tiwari, V. Cocreating understanding and value in distributed work: How members of onsite and offshore vendor teams give, make, demand, and break sense. MIS Quarterly. 2008. 32 (2). pp. 227-255.

Parallelization of NAPOM implementation

Matjaž Depolli*, Gregor Kosec*, Janez Ugovšek*, Vlado Malačič**
* Jožef Stefan Institute, Ljubljana, Slovenia
** Marine biology station, Piran, Slovenia
matjaz.depolli@ijs.si gkosec@ijs.si janez@ugovsek.info vlado.malacic@mbss.org

Abstract - In this paper, the code for the North Atlantic Princeton Ocean Model (NAPOM) used by the Marine Biology Station (MBS) is parallelized and optimized. The FORTRAN source code and the hardware architecture of MBS cluster are examined and analyzed to determine the behavior of the NAPOM execution with bottlenecks identified on both ends. Based on the analysis, the most effective optimization and parallelization actions are planned. Most time consuming modules of the NAPOM package are optimized to achieve maximal performance on the hardware architecture. The pre-process modules are distributed on more computational nodes while all independent complex operations are parallelized with the shared memory principles. The resulting parallelized implementation of the NAPOM package executes nearly four times faster than the original one with only a minimal additional load to the MBS cluster.

I. INTRODUCTION

The task of this work is to speedup the execution of a simulation tool used by the Marine Biology Station. The existing linux-based computer cluster as well as the existing bash scripts and FORTRAN codes along with proprietary compiler are used. The options for speeding up the simulation tool include parallelization on the level of operating system, general optimizations of code and parallelization using OpenMP 5 and MPI 5, which are both supported by the available compiler.

The simulation tool is based on Princeton Ocean Model (POM) 5. POM is a 3D primitive formulation model. The model considers fluid flow through Navier Stokes momentum equation coupled with non-linear equations of state together with energy and salinity transport. POM uses two level Euler temporal discretization and Finite Differences Method (FDM) 5 for spatial discretization. The horizontal velocity components, heat, salinity and kinetic energy transport are solved explicitly (internal mode), while pressure, density and vertical velocity component are solved implicitly (external mode). The internal and external modes are solved at different temporal resolutions. The splitting rate is set based on the integration stability criteria (CFL conditions) 5. The dynamics of the internal mode is much less intense and thus it uses longer time step in comparison with the external mode. The internal mode is computed on a 3D domain, while the external considers a 2D domain.

The POM model stands for standard numerical tool for circulation forecast in several countries. The augmented variant of POM, referred as NAPOM (North Adriatic POM) is used for academic and operative forecast at Marine Biology Station in Piran as well as at the Slovenian Environment Agency (SEA). The NAPOM is derived from a POM2k, a variant of POM that undergone last changes in 2006. The NAPOM incorporates local bathymetry, products form local meteorological models (boundary and initial conditions), and daily measurements of rivers discharge and temperature. The whole numerical package consists of two major parts; first, the pre-process where all the required data is gathered and adequately transformed, and second, the main code, where numerical simulation takes place. The original POM is written in FORTRAN 77; however, the augmentations are written in a mixture of FORTRAN 77 and FORTRAN 90. The code uses single precision data type and is assembled mainly in a single source file. The current code is sequential.

MBS uses a computer cluster composed of five computing nodes. Each node is built around a quad-core 2.83 GHz CPU (Intel Xeon E5440) with 64 kB of L1 cache per core. Two pairs of cores share 6 MB of L2 cache and all cores share 8 GB of the main memory. Two Gigabit Ethernet ports are bonded in a load balancing (round-robin) mode and connected to a switch. The bonding should double the interconnect bandwidth. The nodes are running Linux Ubuntu 8.04 operating system.

II. SYSTEM ANALYSIS

Before starting an optimization, the system to be optimized must be analyzed, to separate the modules that are suitable for optimization from those that are not.

A. Suitability for optimization

To identify places that are suitable for optimization we performed several tests of the MBS hardware and of the NAPOM software package. From hardware point of view, we tested network bandwidth to determine the suitability of parallelization over several machines via Message Passing Interface (MPI); and memory bandwidth to see the importance of caching. From software point of view, we performed several profiling runs of unoptimized NAPOM, and several runs with perturbed input and modified compiler parameters to test the numerical stability of the software. We also determined the sections of code that were the most time demanding by profiling the original main code.

Measured network bandwidth between two MBS cluster nodes is shown in 2 along with a reference measurement performed on a similar cluster with a single Gigabit Ethernet interconnect over a network switch. The MBS network performs almost twice worse than reference measurement for all block sizes, although theoretically it

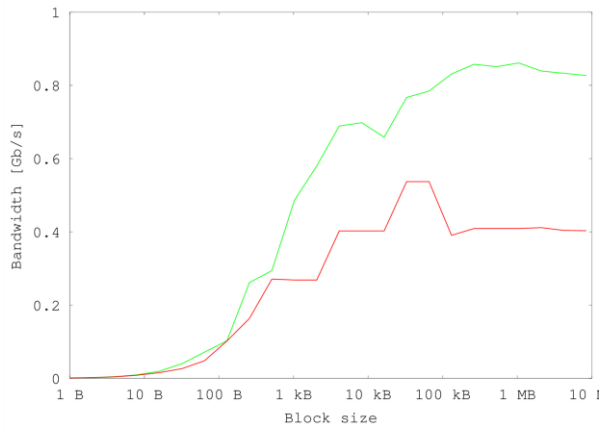


Figure 1. Comparison of MBS cluster network bandwidth (red line) and a reference Gigabit Ethernet bandwidth (green line).

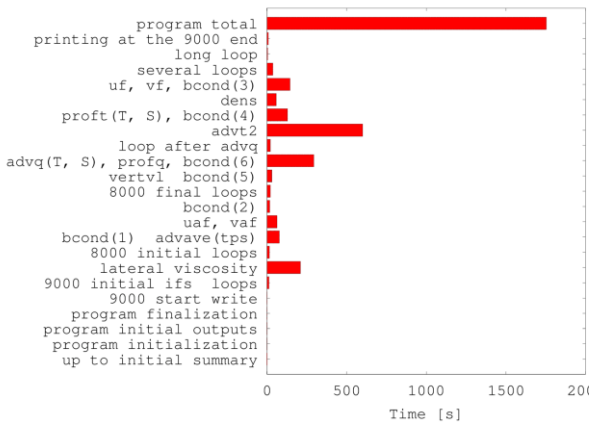


Figure 3. Profiling results of NAPOM main simulation program. Program total (topmost entry) is the sum of all other code segments, named by their most complex functions.

should perform twice better. Therefore, instead of trying MPI parallelization, which can bring good results [6], we tried other options first.

Measured memory bandwidth of a single cluster node is shown in 2. Several plateaus are visible from the plot, corresponding to the levels of the memory hierarchy. Measurement also discloses a weak point of the system, which is the main memory bandwidth of only 3.5 GB/s (the rightmost measured points on 2). When compared with typical workstations that are available in our lab, and easily achieve bandwidths of over 10 GB/s, the MBP cluster main memory bandwidth proves to be a serious bottleneck. Therefore optimization tasks will focus on reducing memory footprint of the NAPOM code as well as increasing memory temporal and spatial locality, thus reducing the number of main memory accesses through increased cache hit rate.

The profiling of NAPOM source code was performed manually by adding timers to important parts of the code, since most of the code is not split into functions and thus manual profiling offers better precision regarding code segments than automated profiling tools. 2 shows segments of code and their execution times as well as the total execution time. Code segments are named by their most complex features. More information about the code can be found in [7]. The most time critical code segments

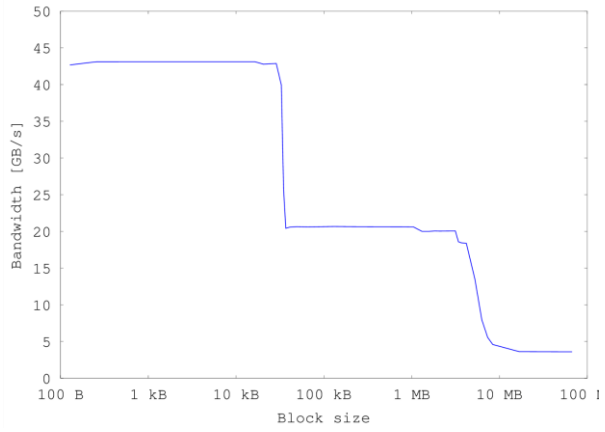


Figure 2. Memory bandwidth as a function of block size, for a single computer from MBS cluster.

that can be identified from profiling are: *advt2*, *advq*, *profq*, and *lateral_viscosity*. These segments will be given the most attention during the optimization and parallelization phase. As can be seen from their names, some code segments are composed of multiple functions, e.g. *advq(T, S)* comprises the calls of function *advq* on two varibles: *S* and *T*. Although those segments could be further divided they were left grouped because of the good parallelization options that they offered as a group.

B. *Numerical stability*

The NAPOM model consists of highly coupled and non-linear problems (fluid flow coupled with energy and salinity transport, turbulence models, etc.) and is of hyperbolic nature because of the advective dominated salinity transport. Consequently, the numerical integration of the model could be unstable, i.e. small perturbations could cause large differences in the final fields. The problem could be even more pronounced as the code is implemented in a single precision mode. In the first stages of optimization, we noticed that changing the compiler flags influences the results. Similar effects are noticeable when the code is vectorized or when minor changes should not cause any side effects are introduced, such as reordering numerical operation sequences. This poses a problem since it is almost impossible to maintain binary compatibility of results between the original and the optimized NAPOM. Simple analysis is performed to examine NAPOM behavior.

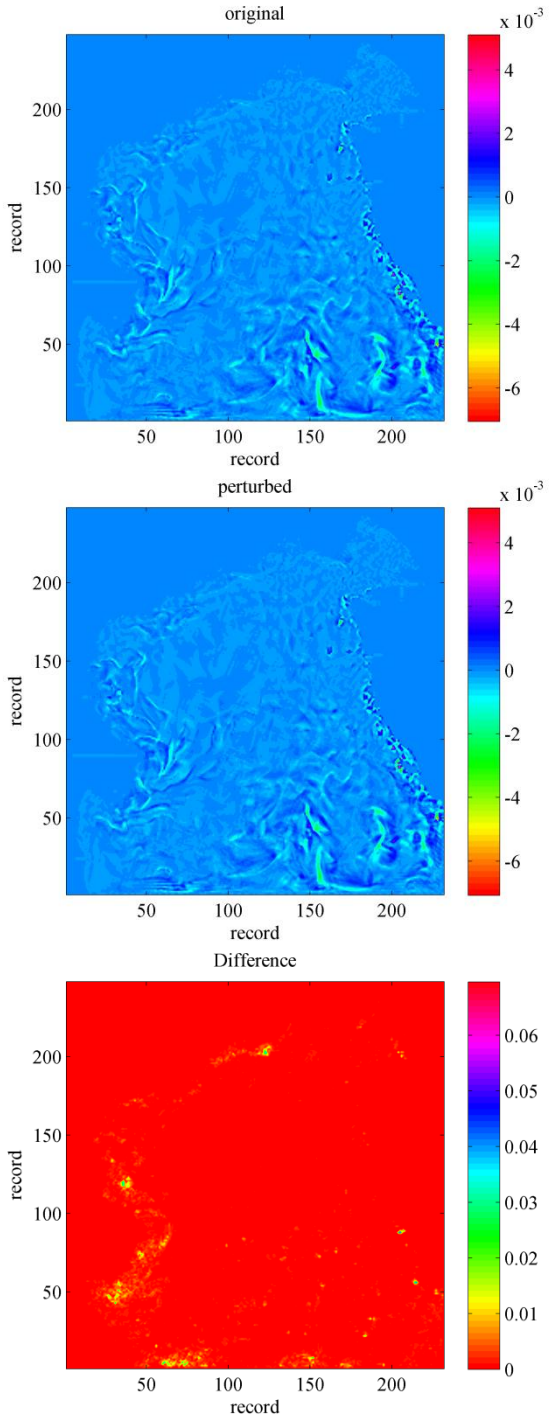


Figure 4. Comparison of results obtained with the original NAPOM (top), the optimized NAPOM (middle), and the difference between the two (bottom).

Two simulations are performed where either the non-perturbed input data is used or the input data perturbed on the least significant bit. The results obtained from the perturbed input are compared against the results obtained from the non-perturbed input. The difference is calculated, for all problem fields, as an absolute difference between the original and the perturbed field, normalized with the maximum value within the considered field. The maximal differences are substantial (up to 9 %). In 3, the results of the original and the perturbed case together with the difference for *w* velocity component on layer 4 are shown. This result serves as an estimate of the differences that are allowed between the original and optimized NAPOM implementations.

III. Optimization Strategy

Although the MPI parallelized POM implementations exist, which could utilize the whole MBS cluster, the current NAPOM code is too tightly integrated with POM2k code to change POM implementation. Replacing the POM engine would require more time that was available for the work on this project. The efficiency of such parallelization is also under question after the initial hardware tests that show network interconnect is far from optimal.

Based on the code and hardware analysis, we performed several actions aimed to parallelize and optimize the NAPOM package. First approach with OpenMP parallelization of independent sections of code and large loops did not prove effective. The effected portions of code experienced only minor speedup, thus giving another clue that NAPOM is memory bound rather than compute bound. Further parallelization approaches were therefore based on the code that was first optimized for cache utilization.

Optimizations of the main NAPOM code included the following actions: The compiler flags were modified to allow for the use of vector (SIMD) instructions supported by the targeted processor and for automatic vectorization of loops. SIMD instructions work on the level of single core, using the multiple floating point units of the core concurrently thus speeding-up execution. Within multilevel loops, vectorization only modifies the innermost level, allowing further parallelization of the other loop levels with OpenMP paradigm.

Unexpectedly, also a correction of the program code was required. There was a violation of matrix bounds when accessing an element in one of the functions that can be traced back to POM2k. Although when compiled for single-threaded use, the violation was not fatal and did not influence results, but when compiled with OpenMP support, it caused the program to crash.

Several unnecessary operations, e.g. $\sqrt{x}^4$ and $x + a + \ldots - a$ were found in code. Such operations produce computational overhead as well as introduce additional rounding errors. Although the overall effect of the unnecessary operations was small, they were removed from the scrutinized parts of code.

Several large matrix variables for storing intermediate results were found unnecessary and replaced with scalar variables, improving cache hit rate. Multiple closely placed independent loops with similar working sets of variables were redesigned into single loops, sharing several operations and intermediate variables; again for better cache hit ratio. A small direct effect was gained but one that is able to multiply through loop vectorization and parallelization.

In the final action, responsible for the biggest performance gain, the domain was decomposed to subdomains that fit in processor cache. Those chunks were then processed in parallel on 4 cores of the processor. The

decomposition was implemented on the level of functions and larger segments of code. The number of subdomains was selected based on individual code segments' working set size. Each code segment performs operations sequentially on subdomains rather than on the whole domain, but still starts and ends with the full domain. Therefore the domain decomposition is not shared among the code segments. Further speedup would be possible if the decomposition were implemented on the level of the main loop with all the code segments sharing the subdomains. However, this would require modifications of code too complex for our current project work. Additional potential of such implementation would be in enabling extended parallelization across the single machine borders – parallelization that could efficiently utilize multiple interconnected computers. It is noteworthy that although aimed at parallel execution, the domain decompositions speeds up the execution even on a single core. This happens because of greatly increased cache utilization and in spite of the overlapping subdomains that cause the doubling of the amount of processing on the overlapping regions.

Second series of optimizations deals with the pre-processing phase, which consumes about one third of the execution time. Most of the pre-processing is written as bash scripts and consists of downloading, unzipping, and other file manipulations. There are also several FORTRAN codes that are used to transform output data of other models and observations to input data. We implemented the following pre-process optimizations:

The FORTRAN codes of pre-processing consist of several independent loops that were parallelized with the same principles as the main code. Also the compiler was set up in the same way as for the main NAPOM code, enabling loop vectorization using SIMD processor instructions.

Several independent actions were performed in series (downloading several files, processing of independent input files, etc.), which we converted to parallel and executed concurrently on three nodes. The concurrent execution had the biggest effect on the reduction of pre-process scripts execution time.

Several dependent actions that were executed serially were converted to execute partially overlapped. For example, file unzipping was now takes place while the zip archive is downloading. Also, as soon as individual files are fully unzipped from the archive they are further processed instead of waiting for previous actions to complete fully before starting the next action.

With the above actions performed, the optimized NAPOM package executed much faster. The optimization / parallelization strategy improved utilization of a single core, of all the available cores within a computer. Instead of a single computer, three computers from the clusters are now used, two of them only during pre-processing. The utilization of the whole cluster could be better, but the achieved additional speedup would hardly justify the substantial additional work required.

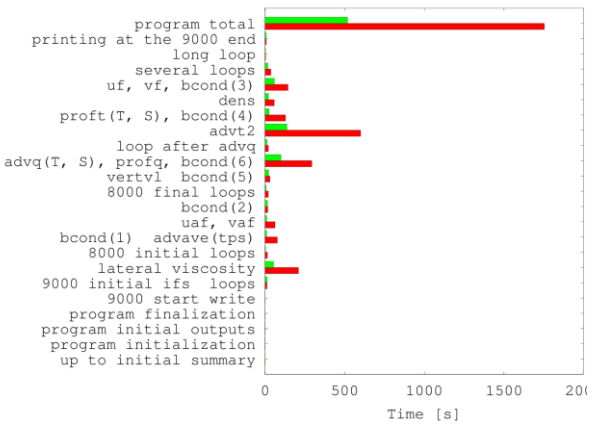


Figure 5. Results of profiling the improved NAPOM simulator (green bars). Results of the original NAPOM (red bars) are added for reference.

IV. RESULTS

We present the speedups resulting from the specified optimization actions. Results are split the same way as optimization actions – to pre-processing and main code.

A. *Pre-process*

The workflow of the optimized pre-processing is shown in 5 and represents a great contrast to the original fully serial workflow. Time measurements over 23 repeats show pre-processing reaches speedup of 6.5. Pre-processing is composed of very limited amount of serial processing and three large and completely independent parts, based on independent sets of files. It was therefore found to be easily parallelizable over three computers, with communication based on remote shell and network file system. OpenMP was put to great extent in parallelizing time demanding FORTRAN programs within a single computer. Finally, multitasking of the operating system and its ability to link outputs and inputs of programs were used to overlap system utilities for downloading, decompressing archives, and executing scripts as soon as new files were extracted. Average speedup of the optimized pre-processing is 5.35; it depends mostly on the time required to download archives with the input files, which is not under our control.

B. *Main code*

Next, the improvements in the main NAPOM code are shown in 4. Most of the demanding operations with high memory usage have been decomposed over y coordinate in order to maximize the cache hit rate. Independent operations with small memory footprint have been parallelized to run concurrently with OpenMP API. Loops with larger memory footprints that shared a set of variables were combined and converted into the smallest possible number of parallel loops. The sequential code has been redesigned into multi-threaded code, which exploits full capabilities of a single computer. All the cores are involved in computations with an optimal reordering of computation regarding the cache architecture. Average speedup of the main code is 3.61.

Average speedup of the whole optimized NAPOM package over the original package, measured on 23 repetitions of daily runs, is 3.85. The use of cluster

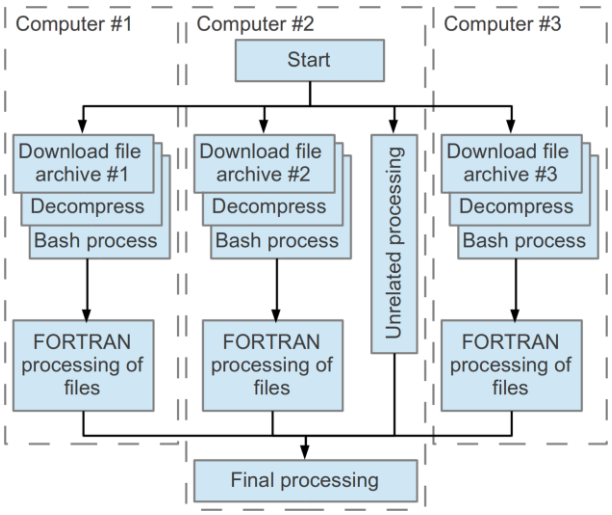


Figure 6. Workflow of the optimized pre-processing. Overlapping boxes represent overlapping of actions in time. All actions executed on the same CPU are marked with broken line.

includes one fully occupied node and two more occupied node for the most of the time of the pre-processing. Total differences between the results of original and optimized NAPOM packages are in the same order of magnitude as the differences acquired in the initial analysis of the NAPOM package. Since results re not binary comparable, the relatively small differences give high confidence that the code has remained functionally identical and no errors were introduced.

V. CONCLUSION

The task of speeding-up NAPOM simulation package was solved using optimizations of the source FORTRAN codes, and the bash scripts of pre-processing. Parallelization was only performed after a careful analysis of the targeted computer system and of the original NAPOM package execution. The main bottleneck was found to be the low main memory throughput, which occurred because of low memory locality, combined with the unoptimized memory access of the main simulation. A speedup of 3.85 was achieved by reorganizing code, increasing its temporal and spatial locality, and only then parallelizing it. Main simulation was only parallelized for the four cores of one computer, while parts of the pre-processing were distributed over three computers and parallelized for four cores per computer. Further parallelizations over the available computers were deemed too demanding while offering only a low additional speedup.

REFERENCES

[1] B. Chapman, G. Jost, and R. van der Pas, “Using OpenMP: Portable Shared Memory Parallel Programming (Scientific and Engineering Computation)”, MIT Press, 2007.

[2] W. Gropp, E. L. Lusk, A. Skjellum, “Using Mpi - 2nd Edition: Portable Parallel Programming with the Message Passing Interface”, MIT Press, 1999.

[3] The Princteon Ocean Model. http://www.aos.princeton.edu/WWWPUBLIC/htdocs.pom/.

[4] M. N. Özisik, Finite Difference Methods in Heat Transfer. Boca Raton: CRC Press; 2000.

[5] V. M. Kamenkovich, D. A. Nechaev, “On the time-splitting scheme used in the Princeton Ocean Model”, J. Comput. Phys., Volume 228 (8), 2009, pp. 2874–2905.

[6] Trobec R., Sterk M., Robic B., “Computational complexity and parallelization of the meshless local Petrov-Galerkin method”, Computers and Structures, 87 (1-2), 2009, pp. 81-90.

[7] G. L. Mellor, “Users guide for a three dimensional, primitive equation, numerical ocean model: Program in Atmospheric and Oceanic Sciences”, Princeton University, 1998.

Faster solution of large, over-determined, dense linear systems

Željko Jeričević* and Ivica Kožar**
* Department of Computer Engineering/Engineering Faculty, Rijeka, Croatia
** Department of Computer Modeling/Civil Engineering Faculty, Rijeka, Croatia
zeljko.jericevic@riteh.hr

Abstract - The solution of linear least squares system requires the solution of over-determined system of equations. For a large dense systems that requires prohibitive number of operations. We developed a novel numerical approach for finding an approximate solution of this problem if the system matrix is of a dense type. The method is based on Fourier or Hartley transform although any unitary, orthogonal transform which concentrates power in a small number of coefficients can be used. This is the strategy borrowed from digital signal processing where pruning off redundant information from spectra or filtering of selected information in frequency domain is the usual practice. For the least squares problem the procedure is to transform the linear system along the column to the frequency domain, generating a transformed system. The least significant portions in the transformed system are deleted as the whole rows, yielding a smaller, pruned system. The pruned system is solved in transform domain, yielding the approximate solution. The quality of approximate solution is compared against full system solution and differences are found to be on the level of numerical noise. Numerical experiments illustrating feasibility of the method and quality of the approximation together with operations count are presented.

I. INTRODUCTION

The least squares method is one of the most widely used techniques in data processing. For linear models it leads to solving over-determined systems of linear equations. The result of this approach is that usually none of the equations in the system are satisfied exactly, but the sum of squared differences between observations and theoretical model is minimized.

While we have very efficient methods for sparse systems and problems with dimensions in millions are routinely solved, the dense problems are more problematic and rarely solved for dimension over 50000.

The paper presents construction of approximate solutions for large, over-determined, dense systems using some ideas from signal processing and the Fourier space representation.

II. THEORY

The Fourier transform is formally done by multiplication with Fourier matrix but in actual computations FFT is used whenever possible. One of the useful properties of Fourier transform matrices is that they have an inverse.

$$b = Ac \tag{1}$$

Applying the Fourier transform on over-determined system of linear equations (1) is done along the columns. By premultiplying with Fourier matrix F, the transform along the column vector y and along the columns of matrix A is performed (2). Before transformation, the elements of vector b and corresponding rows in matrix A should be rearranged in such a way that the transform has compact representation. In other words, the energy in the Fourier transform should be concentrated in as little number of frequencies as feasible.

$$Fb = FAc \tag{2}$$

For the sake of simplicity Hartley transform [1] should be used in the case of real systems to avoid complication of computing with complex numbers. Noting that $Fb=d$ and $FA=D$ our original system (1) is now transformed with elements of vector d and matrix D in frequency domain (3).

$$d = Dc \tag{3}$$

We can see that the transformation along the columns did not affect the solution of the system, noted here as vector c. Formally the over-determined system can be solved using normal equations and matrix inverse. Although that is not recommended for numerical reasons, it will be used here to prove feasibility of frequency selection and modeling of the linear algebra system [2-4].

III. METHOD

After the transformation the system (3) is of the same size as original system but allows pruning the number of rows, down to a square matrix system. The first step in this process is selection of significant frequencies. That can be accomplished by computing the magnitudes for frequencies in vector d and sorting them in decreasing order. With that information, the vector d and matrix D can be row permuted and partitioned in two parts: parts to

be retained d_0, D_0 and parts to be discarded d_1, D_1, as shown in (4).

$$\begin{bmatrix} d_0 \\ d_1 \end{bmatrix} = \begin{bmatrix} D_0 \\ D_1 \end{bmatrix} [c] \tag{4}$$

This would in effect decrease the size (number of rows) of the system and represents the filtering out of non significant frequencies in order to build a smaller model system with less equations then the original system. In this respect our approach is different then Beylkin's [5] whose idea was to increase the sparsity of the matrix by using the wavelet transform. It is important to point out here that due to the properties of Fourier transform, the significant information from all original equations is still preserved. The non-significant frequencies have the least contribution to vector d. The assumption is that solution of model system will be a good approximation of true solution for the original system.

The normal equations are obtained by pre-multiplying with permuted and transposed matrix D.

$$\begin{bmatrix} D_0^T & D_1^T \end{bmatrix} \begin{bmatrix} d_0 \\ d_1 \end{bmatrix} = \begin{bmatrix} D_0^T & D_1^T \end{bmatrix} \begin{bmatrix} D_0 \\ D_1 \end{bmatrix} [c]$$

$$\begin{bmatrix} D_0^T d_0 + D_1^T d_1 \end{bmatrix} = \begin{bmatrix} D_0^T D_0 + D_1^T D_1 \end{bmatrix} [c]$$

$$g = \begin{bmatrix} D_0^T d_0 + D_1^T d_1 \end{bmatrix} \quad G = \begin{bmatrix} D_0^T D_0 + D_1^T D_1 \end{bmatrix}$$

$$g_0 = D_0^T d_0 \quad G_0 = D_0^T D_0 \quad \Rightarrow \quad G_0 = G_0^T$$

$$g_0 = G_0 [c + \delta c] \tag{5}$$

$$[c + \delta c] = G_0^{-1} g_0 \tag{6}$$

Where G_0 is approximation of D^TD and g_0 is approximation of D^Td. based on the most significant frequencies (computed using only the sub-matrix and sub-vector with index zero). We can now solve the model system (5) to get approximate solution of original system (1). The approximate solution denoted $[c+\delta c]$ is expected to be different from true solution c for difference vector δc due to used approximations for d and D. It is important to show how difference vector behaves depending on size of matrix G_0. The iterative formula for improving the solution of linear system (7)[6] and eigen-decomposition of residual matrix R will be used for that purpose.

$$\delta c = G^{-1} \delta g \tag{7}$$

$$G_0^{-1} \approx G^{-1}$$

$$R \equiv I - G_0^{-1} G \Rightarrow G_0^{-1} G = I - R$$

$$G = G_0 (I - R)$$

$$G^{-1} = (I - R)^{-1} G_0^{-1}$$

$$\delta c = (I - R)^{-1} G_0^{-1} \delta g \tag{8}$$

We can see (8) that by using the residual matrix R the solution can be iteratively improved without calculating G^{-1}. The vector δg and properties of matrix R control the corrections. In that respect R should be a contraction mapping: its eigenvalues should all be non-negative, real and smaller then one.

$$(I - R)^{-1} = \left(Q Q^{-1} - Q \Lambda Q^{-1} \right)^{-1} \tag{9}$$

$$= \left[Q (I - \Lambda) Q^{-1} \right]^{-1} = Q (I - \Lambda)^{-1} Q^{-1}$$

$$\lim_{\lambda \to 0} \left[Q \; diag \left(\frac{1}{1 - \lambda_i} \right)_{i=1,\ldots,n} Q^{-1} \right] \to I \tag{10}$$

The eigen-decomposition of residual matrix R (9) leads to insight about validity of approximation and extent of corrections: smaller the eigenvalues are, closer $(I-R)^{-1}$ is to identity matrix as can be seen from the limit (10) and consequently closer G_0^{-1} is to G^{-1}.

The requirement on eigenvalues of residual matrix is naturally fulfilled by discarding the weakest frequencies from (4) The R matrix is in fact the effect of those missing frequencies, i.e. the insignificant frequencies deleted from matrix D.

In actual practice, $(I-R)^{-1}$ is calculated using Taylor order expansion. The convergence of expansion will be faster with smaller eigenvalues of R.

$$(I - R)^{-1} = I + R + R^2 + R^3 + \ldots$$

$$G_n^{-1} \equiv \left(I + R + \ldots + R^n \right) G_0^{-1}$$

$$G_\infty^{-1} \to G^{-1}$$

The computational advantages of solving the model system are significant. In order to illustrate FLOPS savings on systems of different size, the systems with 10 and 50 unknowns and variable number of equations are shown. Fig. 1. is for the Gram-Schmidt method applied to systems with ten unknowns. It shows number of FLOPS for solving original system, Fourier approximation and wavelet approximation. Both approximations reduce rectangular system to the square size.

Fig. 2. is for the Gram-Schmidt method applied to systems with fifty unknowns. It shows number of FLOPS for solving original system, Fourier approximation and wavelet approximation. Both approximations reduce rectangular system to the square size.

The savings are more impressive for systems with more unknowns and are also better for the wavelet transform which is of order n while the Fourier transform is of order $n\ln(n)$ number of operations for computing the transform. The results suggest that reduction of FLOPS for an order of magnitude is possible when using the Gram-Schmidt method.

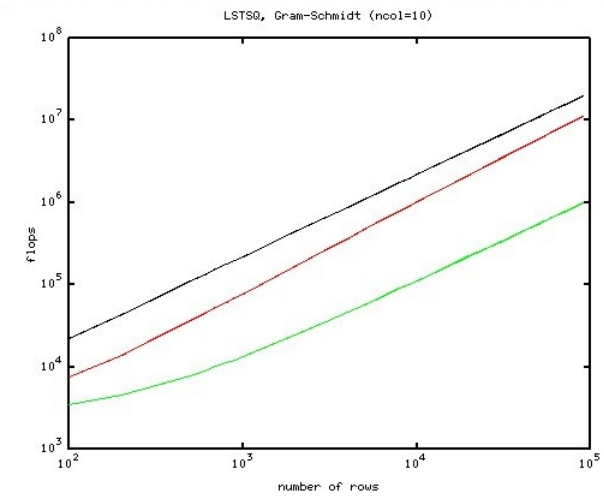


Figure 1. Log-log plot showing FLOPS as function of number of rows. . The number of columns is constant = 10. Upper curve is for exact solution, middle for Fourier and lower for wavelet approximation.

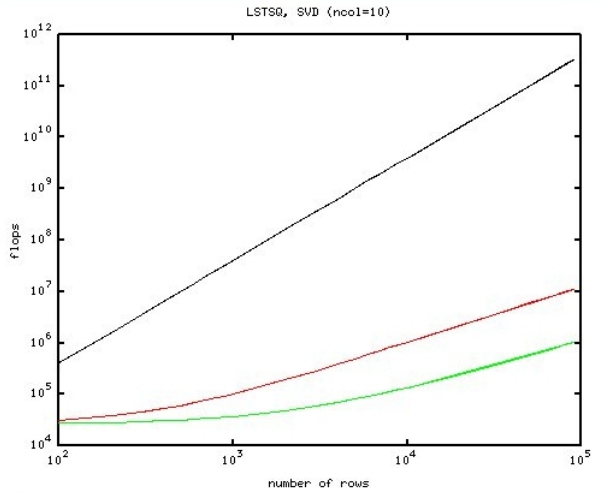


Figure 3. Log-log plot showing FLOPS as function of number of rows. . The number of columns is constant = 10. Upper curve is for exact solution, middle for Fourier and lower for wavelet approximation.

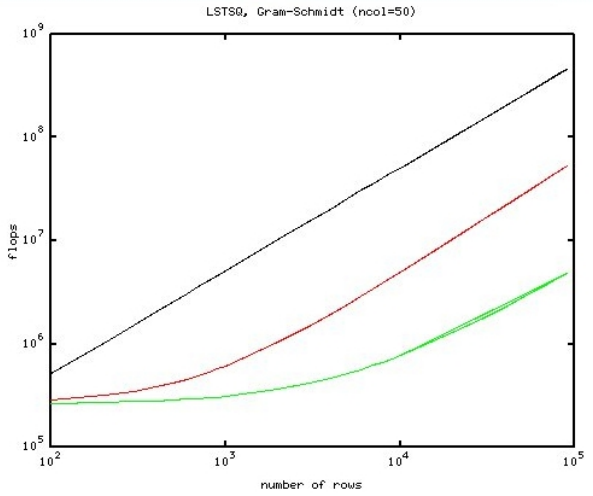


Figure 2. Log-log plot showing FLOPS as function of number of rows. . The number of columns is constant = 50. Upper curve is for exact solution, middle for Fourier and lower for wavelet approximation.

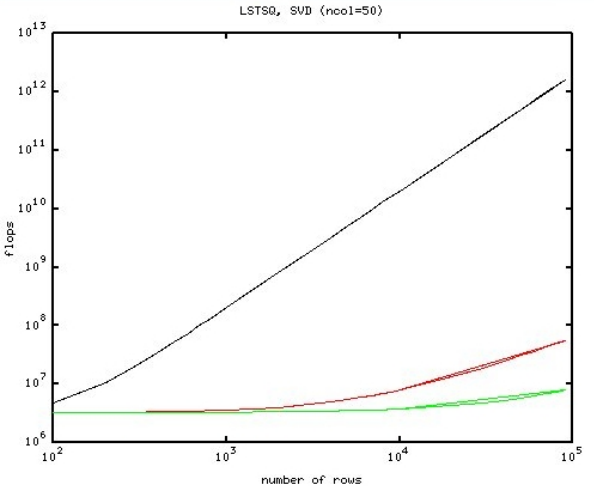


Figure 4. Log-log plot showing FLOPS as function of number of rows. . The number of columns is constant = 50. Upper curve is for exact solution, middle for Fourier and lower for wavelet approximation.

The same counting is done for SVD which is the recommended method for solving the least squares [7] problems. Due the complexity of SVD calculations, the savings are even more impressive. See Fig. 3 and Fig 4. The results suggest that reduction of FLOPS for four orders of magnitude is possible when using the SVD.

One of the numerical experiments is summarized in Table I. Original system was of size 351 by 5 and represents the problem of fitting the polynomial of 4th degree. The polynomial equation is:

$$P(x) = 24 - 14x - 13x^2 + 2x^3 + x^4 = 0$$

In the domain, 351 equidistant points in interval [-4, 3] were generated, and corresponding co-domain points were also generated and used as a vector of observations. For computing the complete system in Fourier space 360 points were used in order to use FFT (original data were padded with zeros).

TABLE I

#Rows	LSTSQ calculated polynomial coefficients				
	C_0	C_1	C_2	C_3	C_4
360	23.9999	-14.0000	-13.0000	2.0000	0.99999
300	23.9999	-13.9999	-13.0000	1.9999	1.00000
250	23.9999	-14.0000	-13.0000	1.9999	0.99999
200	24.0000	-14.0000	-12.9999	1.9999	0.99999
150	24.0000	-13.9999	-13.0000	1.9999	0.99999
100	24.0000	-13.9999	-13.0000	1.9999	1.00000
50	23.9999	-14.0000	-12.9999	1.9999	0.99999
40	23.9999	-14.0000	-13.0000	2.0000	0.99999
30	24.0000	-13.9999	-13.0000	1.9999	0.99999
20	24.0000	-14.0000	-12.9999	1.9999	0.99999
10	23.9999	-13.9999	-12.9999	2.0000	0.99999
5	24.0009	-14.0000	-13.0000	1.9999	1.00000

The numerical experiments for observation vectors with superimposed Gaussian noise were also performed. Noisy data showed the same type of robustness in stability of solution with respect to decreasing number of frequencies (data not shown).

IV. CONLUSION

The approximate method for a fast solution of linear least squares problems has been proposed. The quality of the approximation can be tested and controlled. We developed the equation which shows how the eigenvalues of residual matrix determine quality of solution and extent of corrections to be applied in order to approach the true solution.

REFERENCES

[1] Bracewell, R.N., "The Hartley Transform", Oxford University Press, New York, 1986 p 160.

[2] Jeričević, Ž., "Approximate Solution of Linar Systems" Croatica Chemica Acta, Vol 78 (2005) 601-615

[3] Kožar I, Jeričević Ž and Pecak, T 2009 "Approximate solution of eigenvalue problem for modal analys Conf. of Croatian Society of Mechanics (ICCSM) : Book of Abstracts ed I Smojver and J Sorić (Dubrovnik Croatia) p 133

[4] Jeričević Ž., Kožar I., and Pecak, T. "Modeling of Linear Algebra System with the Fourier Transform", MIPRO (2010) GVS, 231-235

[5] Beylkin, G., Coifman, R., and Rokhlin, V, "Fast Wavelet Transforms and Numerical Algorithms", Com. Pure App Math. Vol. 44 (1991) 141–183

[6] Press, W.H. et all, "Numerical Recipes", 3rd Ed., 2007 Cambridge University Press, Cambridge, pp 1235

[7] Lawson, C.L. and Hanson, R.J. "Solving Least Squares Problems", 1974, Prentice-Hall, Englewood Cliffs, pp 340

Virtual computers and virtual data storage

Alen Šimec, Ognjen Staničić
Tehnical Polytehnic in Zagreb/Vrbik 8, 10000 Zagreb, Croatia
alen@tvz.hr, ognjen.stanici@tvz.hr

Abstract – Virtual data storage represents a new business model which includes various concepts such as virtualization, design of distributed applications and management which enables flexible data access. These methods include using networks of remote servers instead of local servers and personal computers for storing, managing and editing data. Locations containing servers which execute applications and store data are not strictly defined, hence terms "virtual data storage" or "storing data in the cloud" are used. As the need for storing data increased, managing that data became harder as well. Backing up data in large organizations is an inconvenient task. In spite of the increase in power and storage capacity of computers, prices of storing and maintaining data remains high. Various technologies and solutions have developed over time to overcome this problem and in the end they evolved into virtualization of the system for data storage.

I. INTRODUCTION

Virtual data storage represents a new business model which includes various concepts such as virtualization, design of distributed applications and IT management which enables a flexible approach in application scaling and placement. Virtual data storage represents a pay per use model which enables simple and accessible access to computer resource groups via the internet. On the basic level, virtual data storage is a simple means of providing information technology resources as services. Virtual data storage includes using networks of remote servers instead of local servers and personal computers for storing, managing and editing data. Locations containing servers which execute applications and store data are not strictly defined, hence the term "in the cloud" is used. End users access programs via their browsers, whereas the software is stored on remote servers. Virtual data storage enables efficient resource usage, when needed only the required amount of resources is accessed. There is a current trend in the computer industry of developing programs which can be accessed by multiple users without installing anything on their computers. Virtual data storage is also useful if we want to increase the business capacity without investing in new infrastructure, training new staff or licensing new software.

For the end user, virtual data storage brings lots of benefits. Costs of hardware purchasing are eliminated, software licenses have been made obsolete and there is no need for hiring new employees. For these reasons virtual data storage represents a big improvement in IT evolution because it changes the business model. Access to data is unlimited and easily available from any location worldwide.

II.CONCEPT AND DEFINITION

In the 1960s IBM introduces the concept of Virtual machines as means of providing simultaneous, interactive access to their mainframe computers. Every VM was an instance of a physical machine and provided users with an illusion of direct access to the physical machine. It was an elegant and transparent way of sharing resources and processor time of very expensive hardware. Every VM was a completely secure and isolated instance of the core system. Users could execute, develop and test applications without having to worry about crashing the system. Virtualization was used for decreasing hardware supply costs and improving productivity by enabling multi user access on the same computers. With the decrease in hardware prices and the appearance of multi-processor operating systems in the 1970s and 1980s, virtual machines have almost become extinct. The idea of virtualization reemerged in the 1990s with the appearance of a wide spectrum of PC hardware and operating systems. The main task of virtual machines is enabling execution of applications, designed for different hardware and operating systems on a certain machine. This trend is present today.[1]

In the computer world, applications and users perceive a virtual environment the same way they do a real one, even though core mechanisms are formally different. Virtual environments often

[1] http://www.kernelthread.com/publications/virtualization

represent false images of machines or resources which have better or worse abilities than physical machines or resources. A typical computer system uses lots of such technologies. One example is usage of virtual memory in modern operating systems which provides processes with more memory than the actual amount of physical memory the computer provides. It also enables physical memory sharing amongst hundreds of processes. Another example is multitasking where CPU time is divided which achieves assigning each task its own virtual processor.

Virtualization is a way of separating programs and their core components from the hardware which supports them and presenting a logical or virtual view of these resources. This logical view can differ greatly from the physical view. The goal of virtualization is usually one of the following: higher levels of efficiency, scalability, reliability, availability, agility and creating a unique domain of security and management.[2]

III. VIRTUALIZATION USAGE MODELS

The change in business from paper to computers, automation of business processes and the need for communication with clients caused the need for large numbers of computers and servers. The possibility of serving a greater number of systems on one physical server allows organizations to become data centers, which eliminates the costs of building one. Computer architects have to take into consideration various issues while changing from the standard model of distributed applications to a cloud based model. There are three basic virtualization usage models: public, private and hybrid clouds.

Private clouds are implemented for single user usage and they provide the highest data control, security and quality. Large companies with their own data centers are interested in private clouds (also called internal or corporative clouds) which provide additional optimization using the principle of virtualization or distributed computing. That way IT sectors become service providers for other business sections of the company. The company owns the infrastructure and has control over the execution of applications. The cloud can be implemented inside the company data center or in a remote location. This model enables companies control over resource usage, while a third party takes care of the required proficiency for establishing and managing the environment.[3]

Public clouds use internet access with hired resources. The complexity of implemented information and communication solutions is hidden from the user. The majority of management and system surveillance is done using browsers. Public clouds are maintained by independent service providers. Applications from different users are often mixed together on cloud servers, data storage systems and networks. A part of a public cloud can be separated for usage by a single user creating a virtual private data center. The virtual private data center is not limited to distributing images of virtual computers on the public cloud. It also provides users with a better insight in its infrastructure. This way, users can manipulate not only with images of virtual computers, but also with servers, data storage systems, network devices, and network topology. Creating a virtual private data center with all components placed at the same location helps with reducing data locality problems because network bandwidth is large and frequently free when connecting elements within the same location.

Hybrid clouds are combinations of public and private virtualization models. The hybrid model helps with enabling a system available on request, with the service being provided by a third party. Only a part of the programs, services and data are located in the cloud, while the rest of the IT system is located on the company infrastructure. The system structure allows end users undisturbed access to required data, regardless of the physical location on which they are located and the location of the business application which processes them.

IV. VIRTUAL DATA STORAGE SYSTEM

Virtual data storage systems are used to join physical storage from various devices in order to create an appearance of a single data storage source. Storage servers are computer systems which allow general purpose systems safe access via user accounts. The operating system of the storage server is responsible in deciding which general purpose server is allowed access to which data on storage devices. If storage servers are connected to general purpose systems via special purpose networks, this configuration is often called Storage Area Network (SAN). Storage accessed via the network is called Network attached storage (NAS), regardless whether the general purpose system uses SAN or Local Area Network (LAN).[4]

---

[2] Kusnetzky, D. *Virtualization: A Manager's Guide,* O'Reilly Media, Inc., Sebastopol, 2011., pp 1

[3] Wolf, C. Virtualization: From the Desktop to the Enterprise, Verlag New York, 2005., pp. 12

[4] Judith, R.; Davis, Eve, R.; Promotions, w. Data Virtualization: Going Beyond Traditional Data

Storage devices are based on rotating media such as traditional disks, solid state technology such as solid-state disks (SSD) or dynamic memory with direct access (DRAM). Storage can have several forms: direct attached storage (DAS), network attached storage (NAS) or storage area networks (SANs). It can be connected via several protocols: Fibre Channel, Internet SCSI (iSCSI), Fibre Channel over Ethernet or Network File System (NFS). Even though storage virtualization is not a prerequisite for server virtualization, one of the key consequences of storage virtualization is the ability to rely on "Thin provisioning".

Storage Area Networks are usually used in large data centers for enterprise resource planning applications (ERP) and they use expensive "fiber channel" links and switches.

Internet Small Computer System Interface (iSCSI) is a standard used for connecting data storage devices based on the IP protocol. It is a SAN protocol which allows organizations data storage consolidation into data storage arrays while providing hosts with the illusion of locally attached disks. This protocol is usually used while performing virtualization for starting and initializing virtual machines. ISCSI SAN can have efficiency problems which will prevent its usage in the virtualization architecture. Companies should balance Fiber Channel SAN and NAS with the managing complexity of the iSCSI protocol while implementing medium or large data storage systems.

Network File System is the preferred storage virtualization protocol because of its simplicity, scalability and inexpensive implementation and management. From the hardware aspect NFS is a "plug and play" system which uses network transport elements which are present in modern data centers.

V. VIRTUAL DATA STORAGE FUNCTIONALITY

Data storage system virtualization creates an artificial view on the network which hides the physical network from the clients and the server. It manifests as a "layer" between physical and logical data storage processes. The layer is used for standardizing. Any physical solution for storage can be used in a logical solution, without the need for adjustment. Functionalities provided by virtual data storage are:

- Allowing distributed file systems: Remote storage devices are made to look as though they are connected directly to the system. The local system is unaware of their location or the type of storage device being used.
- Creating artificial storage blocks: Multiple storage devices can be connected to create an illusion of a single, much larger storage device
- Creating block arrays for storage: Applications and data can be spread out over a large number of devices and storage servers in order to improve storage efficiency. This functionality can also be used for improving storage reliability. Same data can be stored on various devices or servers. If one of them fails, data can be reconstructed.
- Allowing better storage space management: Storage devices can be segmented into multiple file systems, allowing the storage device better usability.
- Allowing incompatible systems storage device sharing: Mainframe, Windows, Linux and UNIX all use different mechanisms for storing and accessing applications and data. Storage system virtualization allows all of them to share the same storage devices and files contained in them.[5]

VI.STORING DATA "IN THE CLOUD"

Cloud storage is a network storage model in which data is stored into virtual data storage pools. Host companies control large data centers, and people needing to store their data somewhere buy or rent space from them. In the background, data center operators virtualize resources with regards to client's needs. Physical resources can extend over multiple servers. Cloud storage services can be accessed through an API or through a user interface.

The term "cloud" simply refers to going from local services to internet services, from local data storage to a safe and configurable environment, from using applications with limited space (in gigabytes) to applications without an upper limit, from using Microsoft Office to using web-based office. Between 2005 and 2008 network data storage has become cheaper and safer than local storage.

Integration to Achieve Business Agility, 2011

[5] Smoot, S.R.; Tan, N.K. Private Cloud Computing: Consolidation, Virtualization, and Service-Oriented Infrastructure. 2011.

Clouds offer access to cheap hardware and resources for storage using simple APIs and are based on pay-per-use models, hence renting these resources is much cheaper than looking for other solutions. Users are used to storing data remotely in clouds, so they gain popularity with engineers, small and medium sized companies and casual users.

VII.EXAMPLES OF COMMERCIAL CLOUDS

The most famous commercial cloud service providers are Google, Amazon and Microsoft. With Google App Engine users develop their own web applications which are then deployed on the Google infrastructure. Applications can be written for execution in Java, Go or Python environment. Users write applications and deploy them onto the App Engine service using their Google App Engine accounts. The service is free with a limited amount of resources, whereas getting extra resources requires payment. Users can apply for the Google domain, or have their own domain. After deployment Google looks after the application. Their architecture grants resources to applications, looks after forwarding requests to them and gives them time for putting together responses. Google also offers additional services such as Google Cloud Storage which user applications can use. It is estimated that Google App Engine stores 1 million active applications and has around 250.000 active users.

Infrastructure as a service (IaaS[6]) has improved with the development of virtual technologies and the appearance of server renting. The most known IaaS provider is Amazon, and the service is called Elastic Compute Cloud (EC2). Amazon EC2 is a real virtual computer environment which enables using web-service interfaces for running instances with different operating systems, loading them with an adjusted application environment and managing network access. In the App Engine users deploy their own web applications in the cloud, whereas in EC2 users pay for access to virtual computers which are run in the cloud. A user is granted a virtual computer which he can access through the ssh command line in case of Linux operating systems or through the Remote Desktop console in case of Windows operating systems. Prices depend on the strength of the computer and start from a couple of cents per hour using Linux and go to a dollar per hour using Windows with preinstalled additional software such as for example a database management system. Amazon Web Services (which include Amazon EC2) hit revenue of around $500 Million in 2010.

Microsoft Azure is a typical example of a Platform-as-a-service (PaaS[7]) model, because it allows users application development on a remote computer (in the cloud). Windows Azure is a service which allows companies and individuals application deployment in Microsoft data centers, with payment for resources used. This approach absolves companies of infrastructure maintenance and allows them to focus on software development. Windows Azure enables savings in expenses for computer resources since users only pay for the resources they actually used. The platform is extremely easy to use and it is only necessary for a company to choose how much resources they want to reserve for a certain application. Windows Azure is completely scalable and allows simple increase in resource usage if the need arises as with an increased number of users of a web application. Visual Studio 2010[8] is used for developing Windows Azure applications. This way companies and individuals can develop applications on their own computers and deploy them on Windows Azure via integrated tools. The application appears instantly on the platform and users can start using it. Beside Visual Studio 2010, applications can also be developed in the Java Eclipse environment. Other programming languages are also supported such as PHP, Python and Ruby which allows interoperability of the platform. It is estimated that there are around 50 million active user accounts.

Skydrive is a personal cloud which provides users with immediate and protected access to private data from all devices with the possibility of sharing files and folders with others. Instead of using multiple services, SkyDrive users get a single service which allows them to access their data without the need to copy files from cloud to cloud in order to share or to search multiple data locations in order to find their files. Skydrive is used by 17 million active users which share their photographs privately or

[6] IaaS (*Infrastructure as a service*) servers deliver resources on demand, fetching them from large pools built in data centers.

[7] PaaS (*Platform-as-a-service*) facilitates the deployment of applications without the cost and complexity of buying and managing the underlying hardware and software. It provides everything neccessary for the application development nad delivery system to be available via the Internet. PaaS offers tools for application design, testing, implementation and server functionalities. It can also offer application services such as team collaboration, integrated database, security, scalability, warehousing, etc.

[8] A development tool completely adjusted for programming applications and services for the Windows Azure platform.

collaborate on Office documents. Currently Skydrive stores around a million gigabytes of user data and a big growth is expected in the future. These numbers are important because they show there is a tendency for Skydrive to become a service considered valuable by users and used daily.[9]

Cloud computing is definitely on the rise. A study done by Gartner estimates the cloud computing market is worth $150 Billion in 2013. Research also shows that 60% of server workloads will be virtualized by 2014, which compared to 12% in 2008 is a staggering increase. An independent research firm TNS surveyed more than 3500 cloud computing users in eight countries around the world. 93% of respondents said cloud computing had improved their data center efficiency and 80% saw these improvements within six months of moving to the cloud. 82% of respondents said they had saved money on their most recent cloud project. Another study surveyed 3258 companies showing areas of cloud usage, the two largest were accounting (20%) and payroll (20%). The rest were collaboration (17%), file/data storage and backup (15%), business class e-mail (14%) and CRM (14%).[10]

VIII.CONCLUSION

All recent cloud computing forecasts show rapid adoption of cloud computing in enterprises. The cloud computing marketplace will reach $16.7 Billion in revenue by 2013, according to a new report from the 451 Market Monitor, which is a growth from $8.7 Billion in 2010, a compound annual growth rate of 24%. Forrester forecasts that the global market for cloud computing will grow from $40.7 Billion in 2011 to more than $241 Billion in 2020. Cisco predicts that Global cloud IP traffic will increase twelvefold over the next 5 years, accounting for 34% of total data center traffic by 2015. The growth of cloud computing will also have positive consequences on other parts of the IT industry such as mobile services and platforms.

[9] http://cloudcomputingtopics.com/2011/11/5-cloud-computing-statistics-you-may-find-surprising/

[10] https://community.dynamics.com/product/crm/crmnontechnical/b/crmsoftwareblog/archive/2012/01/25/smb-cloud-adoption-study-statistics-show-that-cloud-usage-is-on-the-rise.aspx

IX.REFERENCES

Columbus, L. Forecasting Public Cloud Adoption in the Enterprise, http://www.forbes.com/sites/louiscolumbus/2012/07/02/forecasting-public-cloud-adoption-in-the-enterprise-2. 2013.

Judith, R.; Davis, Eve, R.; Promotions, w. Data Virtualization: Going Beyond Traditional Data Integration to Achieve Business Agility, 2011

Kusnetzky, D. *Virtualization: A Manager's Guide.* O'Reilly Media, Inc., Sebastopol, 2011.

Singh, A.. An Introduction to Virtualization. http://www.kernelthread.com/publications/virtualization. 2013

Smoot, S.R.; Tan, N.K. Private Cloud Computing: Consolidation, Virtualization, and Service-Oriented Infrastructure. 2011.

Toomey, L. 5 Cloud Computing Statistics. http://cloudcomputingtopics.com/2011/11/5-cloud-computing-statistics-you-may-find-surprising. 2013.

Wolf, C. Virtualization: From the Desktop to the Enterprise, Verlag New York, 2005.

AdriaScience Gateway: Application Specific Gateway for Advanced Meteorological Predictions on Croatian Distributed Computing Infrastructures

D. Davidović*, T. Lipić * and K. Skala*
* Ruđer Bošković Institute/Centre for Informatics and Computing, Zagreb, Croatia
E-mail: {ddavid,tlipic,skala}@irb.hr

Abstract - Marine traffic and rapidly growing tourism in the Adriatic Sea region are the main reasons for investigating different weather phenomena and developing the prediction models that depend on distributed computing infrastructures. Weather phenomena like storms and waterspouts are common in the in the Mediterranean Basin and thus in the Adriatic Sea region. Their occurrence is becoming easier to follow in past few years thanks to the advanced numerical weather prediction models and modern communication and computing technologies. This paper presents AdriaScience gateway, the application specific gateway for meteorological applications based on WS-PGRADE/gUSE technology developed within the SCI-BUS project. The main goal of the application specific gateway is to facilitate a universal access to the different Croatian distributed computing infrastructures for Croatian meteorology community.

I. INTRODUCTION

The weather conditions have a great impact on the economy especially on tourism and marine traffic that are rapidly growing in the Adriatic Sea region in the past few years. Severe weather phenomena like storms and waterspouts are common in the in the Mediterranean Basin and thus in the Adriatic Sea region. Storms and waterspouts attract attention not only because of their unique appearance but also because of their ability to produce high risk to the community and also significant property damages. Thus, investigation of this kind of weather phenomena and accurate and forehead prediction of their occurrences is important for Croatian economy and society.

Different techniques and models are used to predict weather conditions and special phenomena that occur in the Adriatic Sea region. These techniques depend on gathering the meteorological data from the observatory stations and weather radars or using advanced meteorological prediction models from atmospheric sciences.

The atmospheric sciences of today are advancing towards predicting very complex phenomena at ever smaller temporal and spatial scales. These phenomena have large impacts on both scientific research and operational weather forecast. One of the principal tools utilized in these efforts are numerical weather prediction models. These numerical models usually demand large execution time and resource allocations. Therefore, large parallel systems are required in order to execute them. One of such numerical prediction models is public domain WRF-ARW model (Weather Research and Forecasting model - Advanced Research WRF) [1]. There are some non-free models like ALADIN [2] model that are mostly used by the national weather and forecasting centers.

As numerical weather prediction models require very large computational and storage resources, most of them are developed to exploit different distributed computing infrastructures (DCI) like Clouds, Grids and High Performance Computing (HPC) systems. Accordingly, many of them have been extensively developed and adopted for different DCIs in the last decades. For example, the WRF-ARW model is developed for the cluster and HPC systems but some research has also been done in porting the model on the Grid: SEE-GRID infrastructure [3] and EGEE/EGI infrastructure [4]. The drawback is that the process of the adaptation for each application is not unique and it requires, for the end-user, a certain level of technical knowledge of each infrastructure in order to be properly used.

Therefore, the ultimate goal of the future development would be to make a transparent and uniform access to different DCIs which will hide from the end user the architecture-specific implementation details. One solution is the development of scientific gateways that will hide the infrastructure API and allow users to execute their applications without even knowing on which infrastructure their application is executed. In this paper we will describe how we developed the AdriaScience application specific gateway that utilizes Croatian distributed computing infrastructures for meteorological applications. The gateway is based on the WS-PGRADE/gUSE [5] generic purpose gateway developed within the EU FP7 SCI-BUS project [6] .

II. RELATED WORK

In the last few years were numerous adaptations of weather prediction models on different grid infrastructures. The development of AdriaScience application specific gateway heavily depends on our previous implementation of the WRF-ARW model on the SEE-GRID infrastructure [7] and WS-PGRADE/gUSE framework for easier development of custom Scientific Gateways for various distributed computing infrastructures.

A. WRF-ARW for SEE-GRID

The WRF-ARW model framework on the grid infrastructure is a collection of the user-specific command line execution scripts that are executed on the front-end. The scripts are responsible for submitting model to the grid Workload Management System (WMS), allocating the grid computing elements (CE) and retrieving data from the grid storage elements to the front-end or predefined user machine. The terrestrial data and initial/boundary conditions data are automatically downloaded from the external servers or copied from the grid storage elements. Upon successful download, the WRF model is started. The grid workflow starts with the execution of the WRP pre-processing system (WPS) on one grid computing element, after which the core application is started and the output data are stored on pre-defined grid storage elements. Users can download model outputs from storage elements using the LCG-tools, DM-Web [8] application service developed within SEEGRID project or using customized scripts for automatic WRF model's data collecting.

Submission of a job is the process that sets all model parameters, describes the job using JDL (Job Description Language), updates the WRF model description files and submits model to the grid. Regarding the job described, WMS allocates a grid CE where the user's job will be executed. The main execution script is responsible for defining and setting almost all parameters needed for the model execution. Based on the user command line parameters the script sets model execution type (scientific or operational), number of processors, and model input files and other specific parameters.

As described above, the process of starting the applications and retrieving the data is complicated for the end user who does not have, at least, basic technical knowledge of the grid systems. The process of preparing and submitting the model to the grid is iterative and require series of steps that have to be initialized by the user. Furthermore, the implementation is very specific to the SEE-GRID technologies and it not directly scalable to other DCIs but need a certain adoptions and changes in the execution scripts.

By using Scientific Gateways it is possible to override the scalability and usability problems of the traditional implementation approaches. Accordingly, we have decided to use WS-PGRADE/gUSE framework for the development of AdriaScience application specific gateway that will enable process of customizing, executing and monitoring WRF model workflows easier.

B. Scientific Gateways

Scientific Gateways are frameworks (or sets of tools) which integrate applications, data and tools to enable running applications on distributed computing infrastructures. They are usually integrated via a portal or a suite of applications that can be further customized to meet the needs of a specific community. The gateways enable entire communities of users associated with a common discipline to use computing and storage resources from different infrastructures (national or private) through a common user-friendly interface. In this way, users can focus on their applications instead of learning and managing the complex underlying infrastructure.

Scientific Gateways are becoming to play an important role in many scientific research areas and are being developed for different national and private infrastructures. It is worth mentioning the Scientific Gateways developed on United States infrastructures: Extreme Science and Engineering Discovery Environment (successor of TeraGrid; XSEDE) infrastructure and National Energy Research Scientific Computing Center (NERSC) infrastructure and European Grid Infrastructure (EGI) Scientific Gateways.

Several Scientific Gateways for meteorological and global climate research communities were also emerged in United States and Europe. In United States, there were developed NASA Earth Science Gateway[9], LEAD portal[10] and two Earth System Grid Gateways: one for XSEDE[11] infrastructure and other for NERSC[12] infrastructure. In Europe for EGI infrastructure exists Web-based EarthServer Science Gateway [13].

The relevance of Science Gateways will further increase with the development of more sophisticated user interfaces and easier access mechanisms for many different distributed computing infrastructures. WS-PGRADE/gUSE is a framework that enables easier development of user-specific gateways for various distributed computing infrastructures.

C. WS-PGRADE/gUSE

WS-PGRADE/gUSE is an open source framework for development of generic workflow-based DCI gateways which is being developed within the SCI-BUS project [6] that allows customization and development of the application-specific gateways according to users' needs. The generic gateway supports submission of the applications on different DCIs including clusters (LSF, PBS), service grids (Globus, gLite, ARC, Unicore), BOINC desktop grids, and clouds (Google App Engine, CloudBroker).

The main part of the gateway is gUSE (grid and cloud User Support Environment). gUSE is a virtualization environment that provides a scalable set of high-level services for managing different DCIs. Each DCI is handled by its specific plugin that can be easily added allowing portal developers to add support for their specific DCI.

WS-PGRADE portal is a web-based user interface for gUse services. It is based on Liferay [14] portal

technologies. WS-PGRADE uses the client APIs of gUSE services to turn user requests into sequences of gUSE specific Web service calls. A generic portal can be easily extended to the application specific portal via the Application Specific Module (ASM) [5]. ASM is an API web interface which provides a set of tools for developing customized portlets.

The main feature of the WS-PGRADE/gUSE portal is the workflow support. The portal incorporates a set of tools and editors that enables creation, management and storage of the application workflows. Furthermore, the portal also provides a support for embedded workflows, i.e. workflows that are part of other workflows. In addition the workflows can be shared among users directly on the portal or via the SHIWA Simulation Platform [15]. The portal also features the simultaneous using of multiple DCI for the same workflow. Therefore, it is possible to off-load parts of the workflows on the computing infrastructures that are the most suitable for the particular parts of the application workflow.

III. ADRIASCIENCE GATEWAY

The purpose of WS-PGRADE/gUSE based AdriaScience Scientific Gateway is to provide easier usage of meteorological applications on Croatian distributed computing infrastructures. Currently, the focus is put on integration of WRF-ARW model which will later enable easier integration of other meteorological applications such as ALADIN or ROMS. The WRF-ARW model is integrated as scientific workflow that can be used through user-friendly portlets enabling robust specific user scenarios.

A. *Supported DCIs*

The AdriaScience gateway supports two national distributed computing infrastructures both operated by the University Computing Centre (SRCE), University of Zagreb.

Croatian National Grid Infrastructure (CRO-NGI) [16] is the largest distributed computing infrastructure in Croatia. It consists of 5 grid sites with more than 1180 computing cores, 36 graphical processing units (GPUs), and 47TB of the storage capacities. CRO-NGI sites contain basic services of Globus Toolkit 5 (GT5) grid middleware system. Accordingly, we use GT5 DCI Bridge Plugin from gUSE for access to this infrastructure.

Isabella computing cluster [17] is the largest Croatian HPC system with 480 CPU cores and 2 TB of the main memory. Also it provides 36TB of the shared storage space. On the Isabella are more than 180 registered user and more than 58 projects. Isabella uses Son of Grid Engine (SGE) job management system. To access Isabella resources we developed gUSE DCI Bride Plugin for SGE.

B. *Scientific workflows*

One of the most important parts of the WS-PGRADE/gUSE portal are scientific workflows. A workflow is defined as a sequence of the connected steps (logical parts) that describe the overall application execution throughout time. A workflow also describes data flow between the individual parts and its dependencies. Our objective is to create workflows that will describe our specific meteorological problems with the emphasis on the WRF-ARW prediction model. The workflows are uniformly describing the data flow and the certain parts of the WRF-ARW execution process.

WRF-ARW model is integrated as a gUSE scientific workflow composed of 7 processes. The composition is structure of directed acrylic graph defined using WS-PGRADE/gUSE Graph Editor tool. Fig. 1 depicts the composition of WRF-ARW workflow. The workflow is described generally because it should be applicable of various DCIs. Each yellow square represent one individual computational part of the WRF-ARW model, green boxes input files/data, grey output data and the lines between them the data flow.

WRF-ARW workflow processes are:

- **IBC-data – data collection**. In this process are downloaded initial/boundary condition data from remote servers and stored on the DCI's storage elements. The path to the condition data location on storage elements is passed as an input to *ungrib* process.
- **geogrid – pre-processor**. Define WRF grid, generate map, elevation and land information. As an input requires the location of the terrestrial data. The recommendation is to execute *geogrid* process on the same DCI computing element as *ungrib* and *metgrid*.

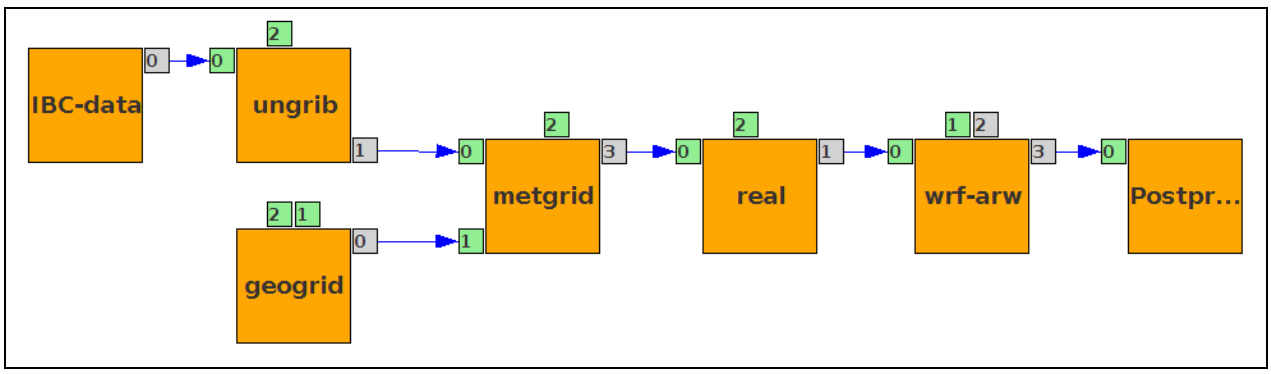


Fig. 1 WRF-ARW WS-PGRADE/gUSE workflow

- **ungrib – pre-processor**. This process unpacks and prepares real-data analyses/forecasts from another (global) models. The inputs are real-data previously stored on the targeted DCI's (CRO-NGI or Isabella) storage system in the *IBC-data* process.
- **metgrid – pre-processor**. This process interpolates the data provided from *ungrib* and *geogrid* processes to the WRF grid.
- **real – real-data initialization**. The process initializes real-date and generates input and boundary files for the core program. It requires inputs from *metgrid* process.
- **wrf-arw – core WRF-ARW program**. This is the main WRF-ARW program. The program requires parallel environment (MPI or OpenMP). It is executed on one or more DCI computing elements depending on the configuration. Requires data from *real* process.
- **post-processing** – This process includes embedded workflows for post-processing. Different workflows can be added here depending on the required post-processing operation. It may include visualization, data extraction or data control and verification. Inputs are data in NetCDF file format generated in *wrf-arw* process.

C. AdriaScience portlets

The main end-user tools of the AdriaScience gateway are the application specific portlets. The portlets are used to make configuration of the application workflows easier and faster. The new AdriaScience portlets will extend the functionality mostly in terms of remote data management and configuration of pre-defined specific application workflows. Also, some existing generic WS-PGRADE/gUSE portlets (e.g. authentication and resource configuration) will be used as well.

The WRF-ARW specific AdriaScience portlets are:

- Portlet for managing data on Global Storage System (GSS -http://code.google.com/p/gss/) and Gfarm. The portlet will enable management of users' data on remote storage systems. This portlet will be integrated in the WS-PGRADE Storage menu.
- Pre-processing portlet for preparing static (terrestrial) and dynamical (boundary condition) data. Using this portlet the users will be able to define the start and the end date of the simulation, upload configuration files, the path on the storage systems (Gfarm or GSS) where the data are stored, and the computing resource. The portlet will be integrated in the End User menu.
- Application specific portlet for each pilot application (e.g.WRF-ARW). The portlets will allow users to define the specific workflow parameters, upload the model input data and upload application configuration files (e.g. namelist.input file for WRF-ARW). The application submission portlets will be integrated in the End User menu.
- Post-processing portlet. This portlet will allow users to define post-processing visualization tools. The functionality of the portlet is to define the path of the files to be processed (as returned by the models and stored on GSS or Gfarm) and the specific parameters for visualization tools.

D. Specific user scenarios

Specific user scenarios describe potential usages of the WRF-ARW model workflows and on the Croatian distributed computing infrastructure. The steps of the use-case scenario are:

- User authentication on the DCI (CRO-NGI or Isabella cluster). The authentication is done in the "Security" portlet by retrieving user certificate.
- Preparation of the terrestrial and initial and boundary condition data. The data are downloaded from local computers or remote servers to the publicly available storage systems (national resource operated by SRCE – Gfarm and GSS)
- Configuration of the workflow. Setting the preferred DCI site, number of processors, WRF-ARW description files (namelist.input and namelist.wps), and defining the remote storage resources for intermediate and final output files.
- Submitting the workflow and monitoring the execution process and intermediate files.
- Remote data management of the intermediate and final output files.
- Post-processing of the model output data including visualization and data management

IV. CONLUSION

The main purpose of AdriaScience Scientific Gateway is to provide easier and intuitive usage of meteorological applications for meteorological and global climate research communities on Croatian distributed computing infrastructures. We focused on integration of WRF-ARW model as workflow and application specific portlets using WS-PGRADE/gUSE framework. While AdriaScience Scientific Gateway facilitates the conduction of robust user scenarios, WS-PGRADE/gUSE framework facilitates the development of such user and application specific workflow-based DCI Scientific Gateways.

For future work we plan extending AdriaScience Scientific Gateway with additional meteorological models such as ALADIN, ROMS or Source-sink model (SSM). This will additionally enable creation of nested workflows that combines two or more models for different complex use cases giving impetus to the formation of larger user community. In addition to extending set of meteorological applications we will also consider ways how to provide

different application specific data management functionalities using application specific portlets.

ACKNOWLEDGMENT

This work is partially funded by the EU FP7 project SCI-BUS.

REFERENCES

[1] W. Wei, C. Bruyère, M. Duda, J. Dudhia, D. Gill, H. C. Lin, J. Michaelakes, S. Rizvi, X. Zhang, "WRF-ARW Version 3 Modeling System User's Guide," Mesoscale&Microscale Meteorology Division, National Center for Atmospheric Research, Boulder, USA, 2009.

[2] ALADIN: High Resolution Numerical Weather Prediction Project, http://www.cnrm.meteo.fr/aladin/, 2013.

[3] D. Davidović, K. Skala, D. Belušić, M. Telišman-Prtenjak, "Grid implementation of the Weather Research and Forecasting model," Earth Science Informatics, vol. 3, pp. 199-208, 2010.

[4] V. Fernández-Quiruelas, L. Fita, J. Fernández, A. S. Cofiño, J.M. Gutiérrez, "WRF4G: simpliying atmospheric modeling experiment design in distributed environments," IBERGRID conference, pp. 327-332, 2011.

[5] P. Kacsuk, Z. Farkas, M. Kozlovszky, G. Hermann, A. Balasko, K. Karoczkai, and I. Marton, "WS-PGRADE/gUSE Generic DCI Gateway Framework for a Large Variety of User Communities," Journal of Grid Computing, vol. 10, pp. 601-630, 2012.

[6] SCI-BUS home page, http://www.sci-bus.eu/, 2013.

[7] D. Davidović, K. Skala, "Implementation of the WRF-ARW prognostic model on the Grid," Proceedings MEET&GVS 33rd International Convention MIPRO 2010, Croatian Society for Information and Communication Technology, Electronics and Microeletronics - MIPRO, vol. 1, pp. 253-258, 2010.

[8] B. Marović, M. Potočnik, D. Okiljević, J. Oklobdžija, "DM-Web – Web access and management of application-specific files in LFC catalogue," Proc. of the SEE-GRID-SCI N. Alameh, M. Bambacus, J. Evans, S. Marley, "NASA's Earth Science Gateway: A Platform for Interoperable Services in Support of the GEOSS Vision," IGARSS 2006.

[9] N. Alameh, M. Bambacus, J. Evans, S. Marley, "NASA's Earth Science Gateway: A Platform for Interoperable Services in Support of the GEOSS Vision," IGARSS 2006.

[10] Christie, M., & Marru, S. (2007). The LEAD Portal: a TeraGrid gateway and application service architecture. Concurrency and Computation: Practice and Experience, 19(6), 767-781.

[11] ESG-NCAR Gateway, http://www.earthsystemgrid.org

[12] ESG - Earth System Grid Climate Gateway and Data-node, http://esg.nersc.gov/

[13] EarthServer Science Gateway, http://earthserver-sg.consorzio-cometa.it/

[14] Liferay portal homepage, http://www.liferay.com/, 2013

[15] SHIWA – Sharing Interoperable Workflows for large-scale scientific simulation on Available DCIs. http://www.shiwa-workflow.eu

[16] Croatian National Grid Infrastructure – CRO-NGI, http://www.cro-ngi.hr, 2012.

[17] Isabella cluster, http://www.srce.unizg.hr/homepage/products-services/computer-resources/cluster-and-grid-technologies/isabella-cluster/, University of Zagreb, University Computing Centre

Handling Web Service Interfaces

J. Matković*, K. Fertalj **
* Elektroprivreda HZ-HB d.d. Mostar, Mostar, BiH
** Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, Hrvatska
jelena.matkovic@ephzhb.ba, kresimir.fertalj@fer.hr

When developing systems based on web service compositions, it is important to define a methodology for handling web service interfaces. Handling web service interfaces comprises publishing, storing and retrieving of interfaces. UDDI is a well-known standard that offers aforementioned possibilities. This paper defines methodology with necessary steps involved in the process of handling web service interfaces and introduces how UDDI deals with it.

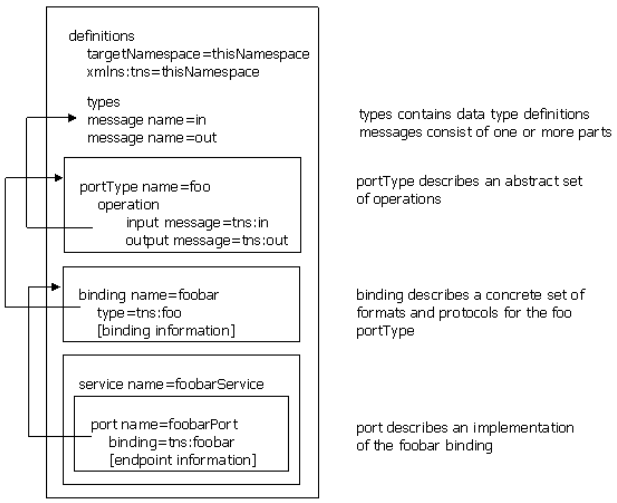
Although web service registries in general do not manage versioning of interfaces, version control process must be considered together with the other processes supported by registries, so a version management and its relation to registries are also considered in the paper.

A final part of the paper introduces models of search engines able to perform search over registries and interface versions based on a given criteria. It is considered how these search engines can be used in conjunction with web service orchestrations.

I. INTRODUCTION

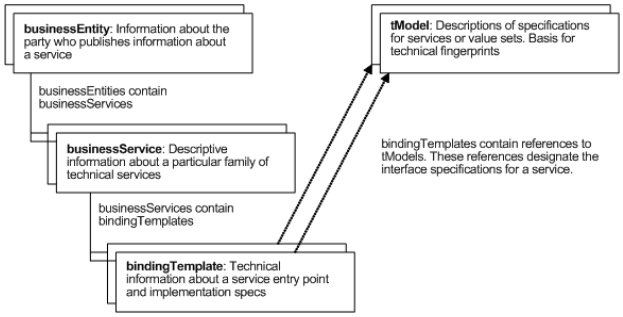
Systems based on web service compositions are built on top of their WSDL [1] interfaces and they depend on them. Compositions do not necessary need to be WS-BPEL [2] orchestrations, but different kinds of interactions between standalone web services. Orchestrations or other clients can be hard coded against WSDL interfaces they access, but more sophisticated way is when clients are hard coded only against abstract part of interfaces (WSDL interface file), and to choose a concrete provider (WSDL implementation file) at runtime. The most sophisticated client would be the one which is not dependent of a WSDL interface at all, and which has some internal logic to automatically perform valid invocations to web services and receiving of responses from them. If the clients are hard coded against concrete WSDL interface, then no repository of WSDL interfaces is needed. In other cases, when clients choose a service provider at runtime or they do not know anything about called web service until invocation time, some kind of a service repository is required since, before invocation takes place, a search over a service repository, based on search criteria, must be performed.

Figure 1. WSDL ELEMENTS AND THEIR INTERRELATIONS [3]



This paper is organized as follows. Section II introduces UDDI [4] since it is the ultimate leader when talking about storing and retrieving of WSDL information. Section III identifies limitations of UDDI model and in order to eliminate them it defines improved proposals that can be used. Section IV inspects possible models for versioning of interfaces and inspects how they can be combined with registries. Section V introduces how to combine search registries and version management with service orchestrations since they do not have direct support in WS-BPEL. Section VI concludes the paper with the analysis of the future work.

Figure 2. UDDI CORE DATA ENTITIES [3]



II. FROM UDDI TO WSDL

UDDI standard defines precise mapping rules from WSDL elements into UDDI entities [3]. Using these rules key elements of WSDL interfaces are described using UDDI entities and stored within UDDI registries. UDDI entities are later composed and transformed back into WSDL elements in order to obtain complete contents of WSDL files, which might be needed for proxy classes to make calls to web services. UDDI entities, used to store information from WSDL files, divide WSDL files into service interface files and service implementation files. A service interface file contains abstract and reusable WSDL elements: *types*, *messages*, *portTypes* and *bindings*, while a service implementation file contains technical elements related to a concrete provider: *service* and *ports*. Multiple service implementation files can share one service interface file using WSDL import mechanism. This would be the case when there are multiple service providers offering the same functionality. Separating one WSDL file into service interface and service implementation files allows multiple service implementation files (referring to the same service interface file) to be stored within UDDI registries. This allows construction of queries able to retrieve all service implementation files that implement specific functionality. UDDI standard defines functions that can be used for writing different kinds of queries but they will be inspected later in the paper after mapping model is introduced. Before introduction of a mapping model, this chapter will introduce WSDL entities since they must be understood first in order to understand mapping model well.

A. WSDL entitites

As already stated, a WSDL file can be divided into: service interface and service implementation files. Service interface file (using aforementioned WSDL elements) is used to describe signatures of exposed operations and bindings associated with these operations. Bindings describe message formats and transport protocols that must/can be used to call and receive responses from operations associated with them. Service interface file does not define where the service is implemented. This information is contained within service implementation file which contains the aforementioned: *service* element and its child *port* elements. Each *port* element is associated with a *binding* element and gives a network address where this *binding* is accessible. Figure 1 graphically depicts interrelations between WSDL elements.

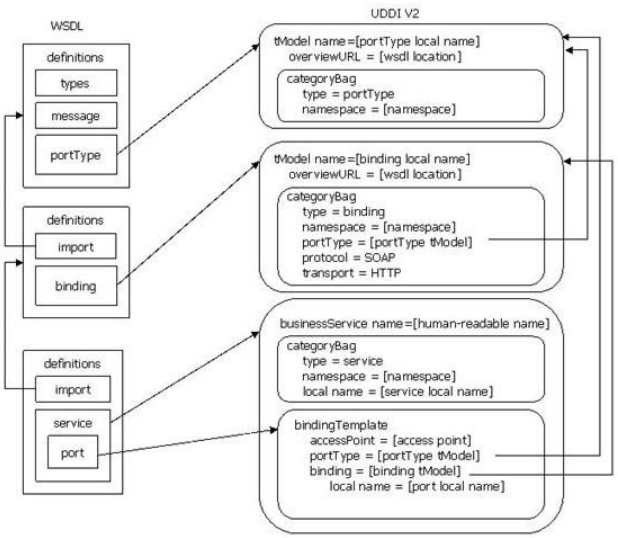
WSDL files should be written in a modular fashion since the import directive allows their separation into multiple files. Following this principle, the first WSDL file can define *types*, *messages* and *portTypes*. The second WSDL file imports the first one and it can define *bindings*. These two files together form WSDL service interface file described previously. The alternative is to write the service interface file entirely as one file, but the first approach allows modularity and reusability. The last file in the process of creating a complete WSDL file is WSDL service implementation file which imports the second file with defined bindings or the whole service implementation file if it is written entirely as one file. Modularity of WSDL files does not affect how those files will be described in UDDI registries but it is implicitly recommended by the mapping model as the best practice.

B. Mapping model

Mapping model defines the rules for transforming WSDL 1.1 entities into UDDI V2 and V3 entities [3]. UDDI *businessEntity* represents a party which is a publisher of a web service and different information about it (name, address etc.) can be specified within this entity. UDDI *businessEntity* can contain multiple *businessService* entities. *businessService* represents logical service and it corresponds to WSDL *service* element. *businessService* element can contain multiple UDDI *bindingTemplate* elements which correspond to WSDL *port* elements. Hence parent-child relationship of *businessService* and *bindingTemplate* entities is used to preserve relation between *service* and *port* WSDL elements in UDDI.

UDDI *tModel* entities are used to represent reusable elements and they are referenced from other UDDI entities that make use of them. Abstract parts of WSDL interfaces (*portTypes* and *bindings*) are represented by

Figure 3. MAPPING MODEL – FROM WSDL TO WS-BPEL [3]



tModel entities and they are referenced when needed. *portType* element and associated *binding* element are mapped to two separate *tModel* elements. In order to preserve relation in UDDI registry that the WSDL *binding* element refers to the WSDL *portType* element, the *binding tModel* in its structure references the *portType tModel.* Similarly in order to preserve relation that the WSDL *port* element references the WSDL *binding* element, the *bindingTemplate* in its structure references the *binding tModel*. Described mapping rules are graphically shown in figure 3.

C. Information within UDDI entitites

UDDI is XML based standard just like WSDL. UDDI elements and APIs are defined in UDDI XML schemas. Its core data entities and existing interrelations between them (Figure 2) have already been mentioned when describing the mapping model. This subchapter gives more detailed explanation of information contained within those introduced entities. As described in the mapping model (Figure 3) not all information from the WSDL file is stored within UDDI registries but only crucial WSDL elements and existing relations between them. Quantity of information stored within UDDI registry is sufficient enough for one user to obtain complete content from the WSDL file which might be necessary for a client to make a call to the web service. Common information of any WSDL element represented by an UDDI entity is: type of the element, its target namespace, its local name and the location of WSDL document where it is defined. Information of the element type enables users to make queries based on the specified type, i.e. "*search all tModels that represent the portType or the binding*". Capturing the local name, namespace, and WSDL location enables users to locate the definition of the represented WSDL element.

D. UDDI query APIs

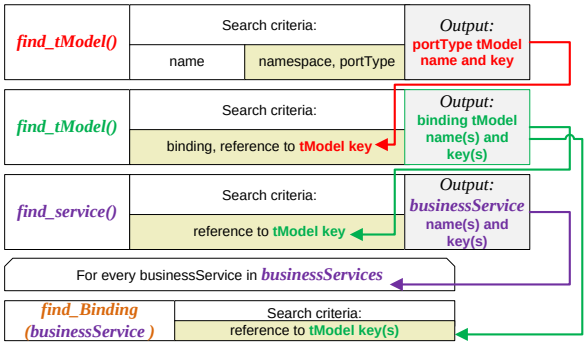
UDDI standard has defined a set of APIs that can be used for issuing queries to UDDI registries. UDDI queries APIs are nothing but defined XML elements that carry information to and from UDDI registries. UDDI client implementations must know how to map application specific data into UDDI query APIs, and to form and send valid SOAP messages that carry APIs, and as well as how to receive and extract information from incoming SOAP messages and to transform them back into the application specific data. This chapter inspects some usual questions issued when searching different elements of WSDL interfaces and it introduces how they can be constructed using UDDI query APIs**.** The task of the chapter is not to semantically analyze all UDDI query APIs (for that refer to [4]).

Query 1: Find portType tModel with a specified name, defined in the specified namespace. This is a first question when an implementation of the specific *portType* wants to be found. In order to find *portType tModels*, *find_tModel* function is used. This function returns names and keys of all *tModels* that satisfy search criteria. Search criteria can be one of the following: a partial or complete name of the *tModel*, its key, or any category information (namespace, type of the element, or any user defined category). In order to find information contained in the *question 1*, search criteria would be: name of the *portType*, category information that *tModel* belongs to the specified namespace and that the type of *tModel* is *portType*. After search is done, we will have the key of the *portType tModel*. This key is very important information that will be used to find other parts of the WSDL interface (desired binding and port). In order to retrieve the complete content of the *portType tModel*, *get_tModelDetail* function is used.

Query 2: Find binding tModel that references specified portType tModel. This is the second question that must be answered after *portType tModel* is found. Since *binding* elements are abstract WSDL elements as well as *portType* elements, they are also described by *tModel* elements. *Find_tModel* function is used for their discovery as well. When this function is used in this second scenario, search criteria would be: category information that the type of *tModel* is *binding* and that it references specified *portType tModel.* If message format or transport protocol is specified in the question, this category information is added to the query as part of the search criteria. In order to obtain the complete content of *binding tModel* element, *get_tModelDetail* function is used.

Query 3: Find all implementations of binding tModel. Implementation of the binding refers to the port that implements specified binding. This is a final part of the search. The first thing that needs to be found is WSDL *service* element represented by *businessService* UDDI element. *find_service* function is used to find *businessService* elements that satisfy search criteria which might be: partial or complete name of the *businessService* element, category information (namespace or some user defined category) or reference to *tModels* (*businessService* references *portType* or *binding tModel* if either one of *bindingTemplate* elements inside of it reference specified *tModels*). *find_service* function returns names and keys of all *businessService*s that satisfy search criteria. Search criteria of *find_service* call in our scenario is reference to the key(s) of the *binding tModel*(s) found in the step 2. After *businessService* elements are found, *find_binding* function is called. This function accepts a key of the *businessService* element as a mandatory input parameter and a search criterion of this function is reference to *tModels*. It returns one or more *bindingTemplate* elements (related to the *businessService* element sent as input parameter) that satisfy search criteria. In our search *find_binding* function is called for every *businessService* element (returned by previous *find_service* function). A search criterion of each call is a reference to the key(s) of the *binding tModel*(s) found in the step 2. Finally, after *find_binding* function retrieves satisfying *bindingTemplate* elements for each service, we will have all ports that implement specific *portType* or *binding*. An alternative to calling *find_binding* function is to call *get_serviceDetail* function for every *businessService*. This function retrieves a complete content of the *businessService* element which could be parsed and appropriate *bindingTemplate* elements could be extracted. In order to better understand described queries, figure presents them graphically.

Figure 4. FINDING PORT THAT IMPLEMENTS SPECIFIC *PORTTYPE*



E. UDDI publish APIs

UDDI standard defines a set of APIs used to publish information from WSDL interfaces into corresponding UDDI entities according to previously explained WSDL-to-UDDI mapping model. The same set of APIs is used to update UDDI information. Publishing to UDDI registries must be transactional, i.e. it must be synchronous and atomic. No partial responses of success should be returned to the client. Every published UDDI entity should have its own unique key which can be user defined or assigned automatically by the registry. As already stated, the same functions are used for saving new as well as for updating existing entities. If the key of the entity passed in the publish function is already an existing key, then update operation will be executed on the specified entity, otherwise new one will be inserted into the registry. The task of the chapter is not to semantically analyze all UDDI publish APIs (for that refer to [4]).

Previous chapter has introduced steps related to retrieving WSDL information in order to make a call to the specified *portType* implemented using the specified *binding*. This chapter introduces steps necessary when publishing one WSDL interface. A publisher application must follow WSDL-to-UDDI mapping model and it must know how to send publish API structures as well as how to receive publish API responses. Of course, requests and responses are sent as SOAP messages in the same way as when the query is issued.

When doing publishing, the first thing that needs to be performed is mapping WSDL *portType* element into *tModel* entity and calling *save_tModel* function. The analogous thing is done to WSDL *binding* element since it is represented as *tModel* element as well. After *portType* and *binding* elements (service interface) are stored, it is time to store service implementation, i.e. WSDL *service* and *ports* elements. *Service* element is mapped into *businessService* element and corresponding port elements are mapped into *bindingTemplate* elements contained within. Since *bindingTemplate* elements reference *businessService*, *binding tModel* and *portType tModel* elements, before their publishing it is necessary to obtain the keys of all referenced elements created previously.

In order to better understand described actions, figure presents them graphically.

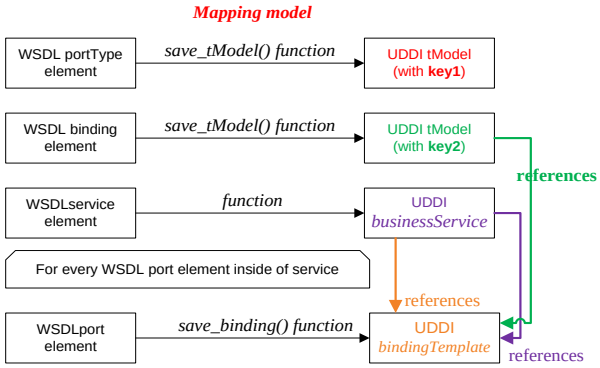
III. ADDITIONS TO UDDI IDEA

UDDI standard has successfully resolved almost all issues related to publishing, updating or querying of WSDL interfaces. WSDL-to-UDDI mapping model is very good because it implies separating of WSDL files into service interface files and service implementation files. This ensures reusing of abstract WSDL definitions stored within UDDI registries. Mapping model is normalized, which implies that no information from WSDL files is stored on more than one place in registries. Referencing is based on the principles similar to the relation of primary and foreign keys found in relational databases. The only case when normalization is not met is when *bindingTemplate* element references *portType tModel* element since it references it transitively already over *binding tModel* element. But it is possible this is performed in this way in order not to slow responses that come from UDDI registries. If *bindingTemplate* element referenced only *binding tModel*, then a search for the *bindingTemplate* that implements specific *portType*, would first require UDDI registry to issue sub-query for finding *binding tModels* that reference specified *portType*. Afterwards *bindingTemplate* elements, that reference *binding tModels* found in sub-query, would be found finally. This is much slower in contrast to the situation when *bindingTemplate* elements reference *portType tModel* elements directly. Then no sub-query for finding *binding tModels* is necessary. So conclusion is that this absence of normalization is met in order to have faster registries because that is more important than fully normalized registries.

The second improvement of UDDI standard relates to versioning issues although this will be discussed in the next chapter more detailed. UDDI standard defines means for putting WSDL files into user-defined categories which can later be exploited as search criteria. Since UDDI has not explicitly resolved issues related to versioning, these categories can be exploited to store information that one WSDL file is a new version of some previous one. To implement this, all UDDI entities related to different versions of the same WSDL file would be stored in one category.

The third improvement of UDDI standard is related to the smallest granularity of WSDL information that can be stored within UDDI registries. According to UDDI it is

Figure 5. PUBLISHING ONE WSDL FILE



WSDL *portType* element. But it is possible and logical to implement that smallest granularity that could be stored within UDDI registries, is WSDL operation. This would allow finding services that implement one specific operation since *portType* element aggregates more than one operation.

The fourth improvement related to UDDI is its relation to other web service standards. UDDI has not defined relation to any standard, except SOAP standard which is used for accessing UDDI registries. However, relation to WS-policy standard must be clearly identified. It must be resolved how WS-policy [5] enriched WSDL interfaces will be stored within UDDI registries and hoe WS-Policy information can be used as a search filter.

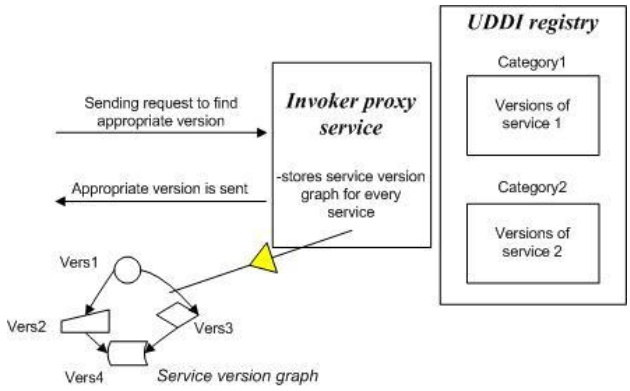
IV. VERSION MANAGEMENT

Web services version management includes activities related to keeping information about different versions of starting web service interface. It also includes care about client adaptation to new versions if it is necessary. As already stated in the previous chapter, support to different versions of the same service is no built-in feature of UDDI registries, so custom versioning mechanisms must be used.

Reference [6] introduces possible versioning mechanism based on "*service versions graphs*". *Invoker proxy service* is put between the client and registries and it stores a "*service versions graph*" for every service in its memory. "*Service versions graph*" is made up of service versions and it shows how they relate to each other. Of course, only transparent functional changes are stored as new versions in a "*service versions graph*", while non-transparent changes, that require adaptation of the client code, must be stored in registries as new services. It is desirable to have as better as adaptable client, but client and its adaptation to service changes is a topic for some other paper.

Search action from a client should go through this *invoker proxy service*. The *invoker proxy service* is initialized with a domain (identifies starting service or a "*service versions graph*"), a selection strategy and a rebinding interval. After initialization, the *invoker proxy service* performs selection over public registries and chooses the most appropriate version based on the selection strategy. The *invoker proxy service* periodically queries the registry according to the rebinding interval, and checks if the previous binding is still accurate. If there is a new version that better matches criteria according to the selection strategy, the *invoker proxy service* discards the current binding. Visual illustration of the described approach is shown in Fig.

Figure 6. APPROACH TO VERSION MANAGEMENT



UDDI category mechanism can be used to map information from the "*service versions graph*" to UDDI registries. Versions that belong to one "*service versions graph*" are put in one category that identifies that graph. Relations between adjacent versions in the graph can also modelled using UDDI category mechanism, so there must be automated algorithm that will form sub-categories that will be used for storing information about order of versions in the graph.

This approach for versioning management is convenient for changes that do not require client adaptation. As already stated in the introduction of the paper, there are three kinds of a bond between clients and accessed web service after the service has been discovered through the use of the registry [7].

The first and the simplest case is when a client is hard-coded against accessed WSDL file. This client is dependent of the service abstract interface file as well as of the service implementation file and no changes at runtime can be made to either of them.

The second case is when a client is dependent of the complete or partial service abstract interface file and the service implementation file is chosen at runtime based on user's search criteria. The level of user's dependence of the service interface file can be decreased using wrappers around invocation frameworks.

Runtime binding is the third and the most sophisticated case. In this case, even the service interface file (the operations signatures) and the protocol are not known at development time. The client can still discover a service based on some search criteria, but with an unknown service interface. In this case, some kind of reflection mechanism must be implemented at the client side that enables the client to dynamically discover the semantics of the service and format of valid requests. This type of service discovery is the most complex and is not used often because it requires very complex client logic to dynamically interpret the semantics of an unknown service interface. As already stated, client's binding to a specific service after the service has been discovered through the use of the registry, is not a concern of registries or versioning management but it will considered together with the development of web service clients and invocation framework wrappers.

V. RELATION WITH SERVICE ORCHESTRATIONS

As already stated, UDDI language has not defined its relation with any web service standard but SOAP standard used for accessing UDDI registries From the perspective of discovery, it is important to include versioning and WS-Policy in UDDI registries in order to enhance capabilities when searching for interfaces. But in order to

gain benefit of registries in the world of business processes, implemented as web service orchestrations, it is important to have a connection between registries and WS-BPEL processes.

In general, there are several approaches for providing web service discovery. Besides registries, there are web service search engines, dedicated discovery web services etc [8]. An overview of search mechanisms has been presented in [9].

Dedicated discovery is a complementary approach that must be used to access service registries from WS-BPEL processes since WS-BPEL standard does not define a direct connection with UDDI standard. The *dedicated discovery service* acts as a client who initiates communication with registries and performs discovery based on a set search criteria. This *dedicated discovery service* is invoked by a WS-BPEL process when a partner discovery is needed. WS-BPEL process can set WSDL *portType*, *binding* or some other abstract WSDL parts that become search criteria upon which the *dedicated discovery service* generates a request to registries. After search is done, the *dedicated discovery service* returns endpoint references (WS-Addressing interpretation of WSDL *port* elements) as a search result allowing the dynamical binding of WS-BPEL processes at runtime.

An emphasis of this paper is on UDDI registries but in general web service registries can be accessed using SOAP protocol and messages defined in some schemas other then UDDI schemas. Search engines also provide means for discovery. An evaluation of several search engines has been presented in [10]. The *dedicated discovery service* can be used as an intermediary between WS-BPEL processes and search engines as well.

Besides dedicated discovery mechanism, another approach that enables integration of WS-BPEL language with registries or search engines is to use extensibility mechanism of WS-BPEL language. An example of such approach is presented in reference [8]. It is based on DPWS standard [11] which includes support for *WS-Discovery* [12], *WS-Transfer* [13] and *WS-MetadataExchange* [14] specifications. WS-BPEL engine must support all of the mentioned specifications in order to run these WS-BPEL discovery extensions. For WS-BPEL engines, that do not support mentioned specifications, the *dedicated discovery services* (performing the same logic) are introduced.

VI. CONCLUSION

This paper gives an overview on UDDI standard with emphasis on WSDL-to-UDDI mapping model and UDDI query APIs. It identifies what queries must be issued against registries in order to obtain an endpoint reference or a complete WSDL interface.

Version management and integration of UDDI with WS-BPEL are analysed in the second part of the paper. A possible mechanism for version management [6] is presented and it is decided that integration of this possible version management mechanism and UDDI registries can be implemented using UDDI category mechanism.

The last chapter of the paper identifies how WS-BPEL processes can access UDDI registries. Since there is no direct support, this must be done by a specialized web service (*dedicated discovery service*) that acts as a client of UDDI registries and performs search based on its internal logic. This web service is called from a WS-BPEL process when there is a need to find appropriate partner web services that will satisfy search criteria. Search criteria are forwarded to the *dedicated discovery service* as an input parameter so it knows what type of services to find in the registries.

A future work will be focused on a few fields in the area of managing service registries. It will be considered how to enrich WSDL files with WS-Policy elements and how to store such files in service registries. In the area of version management, a future work will exploit UDDI category mechanism in order to store service version graphs. *Invoker service* (used for storing versions graphs) will be implemented as well. In the area of integrating service registries with service composition, a future work will be focused on implementing the discovery service. Considering the area of invocation from clients to web services, development of the wrapper, that will allows clients to be more flexible and less dependent of a service interface, will be considered.

VII. LITERATURA

[1] *Web Services Description Language*, (WSDL) 1.1, 15 March 2001.

[2] *Web Services Business Process Execution Language*, OASIS Standard, Version 2.0, 11 April 2007.

[3] *Using WSDL in a UDDI Registry*, UDDI Spec TC, Version 2.0.2

[4] *UDDI*, UDDI Spec Technical Committee Draft, Version 3.0.2, 19 October 2004.

[5] *Web Services Policy*, 1.2 - Framework (WS-Policy), 25 April 2006.

[6] P., Leitner, A., Michlmayr, F., Rosenberg, & S., Dustdar, "End-to-End Versioning Support for Web Services" in Services Computing, SCC '08. IEEE International Conference, Honolulu, Hawaii, USA, 2008., pp. 59-66.

[7] W. Roshen, "Web Services Overview" in SOA-Based Enterprise Integration: A Step-by-Step Guide to Services-Based Application Integration, 1st ed. City of Publisher, The McGraw-Hill Companies, 2009, ch. 7, pp. 129–130.

[8] H., Bohn, F., Golatowski, & D., Timmermann - "Dynamic Device and Service Discovery Extensions for WS-BPEL" In Service Systems and Service Management International Conference, 2008., pp. 1-6.

[9] J. K., Garofalakis et. al. "Contemporary Web Service Discovery Mechanisms" Journal of Web Engineering, vol. 5, no. 3, pp. 265-290, September, 2006.

[10] D., Bachlechner et. al., "Web Service Discovery A Reality Check," DERI Digital Enterprise Research Institute, University of Innsbruck, Innsbruck, Austria, Tech. Rep., January 2006, Available at: http://www.deri.ie/fileadmin/documents/DERI-TR-2006-01-17.pdf.

[11] *Devices Profile for Web Services*, DPWS v1.1, 1 July 2009.

[12] *Web Services Dynamic Discovery*, WS-Discovery v1.1, 1 July 2009.

[13] *Web Services Transfer (WS-Transfer)*, 27 September 2006.

[14] *Web Services Metadata Exchange (WS-MetadataExchange)*, 13 December 2011.

Loosely or Tightly Coupled Affinity for Matrix - Vector Multiplication

Goran Velkoski, Sasko Ristov and Marjan Gusev
Ss. Cyril and Methodius University,
Faculty of Information Sciences and Computer Engineering,
Skopje, Macedonia
Email: velkoski.goran@gmail.com, sashko.ristov@finki.ukim.mk, marjan.gushev@finki.ukim.mk

***Abstract*—Today's CPU cores usually possess private L1 and L2 cache and share L3 cache with other cores of the chip (die). Private or shared cache could have significant impact to the algorithm performance in parallel implementation, i.e. using tightly coupled CPU cores with the same last level L3 cache, or loosely coupled CPU cores with private L3 cache per chip. Private cache increases the overall cache size used during the execution. On the other side, shared cache provide implicit prefetching of the data reducing cache misses if all CPU cores of the chip use the same data. In this paper we analyze the matrix vector multiplication (MVM) algorithm performance represented with speed and speedup. We realize sequential and parallel implementation in multi-chip multi-core multiprocessor in order to determine the CPU affinity that provides the best performance for parallel implementation using the same number of tightly coupled CPU cores and their counterparts - loosely coupled CPU cores. The results show that working on loosely coupled cores with private L3 cache is better than working on tightly cores with shared last level L3 cache in the region where the problem size can be stored in the total L3 cache of loosely coupled CPU cores, but in the same time cannot be placed in only one L3 cache of tightly coupled CPU cores.**

***Index Terms*—Gustafson's law, shared memory multiprocessor, high performance computing**

I. Introduction

The CPU frequency improvement from megahertz to gigahertz saturates since the main CPU vendors refocused their development projects from high-speed (high-frequency) uniprocessors to multicore per chip multiprocessors [1]. The requirements of loosely coupled computing power led to development of different CPU architectures with increased number of cores on a single chip. Multicore multiprocessor possesses at least two CPU cores placed in one chip. Multichip multiprocessor is an architecture where at least two chips (CPUs) are placed in a single motherboard. Other direction of development is manycores, such as GPUs.

Executing some algorithm with more CPUs is limited with maximum linear speedup according to Gustafson's Law [2]. However, scaling is limited with the serial part of the algorithm as defined by the Amdahl's Law [3]. Apart that the most of the real life problems are limited by the Amdahl's Law since there are data dependencies which must be serialized, there are many granular and scalable linear algebra algorithms often used in the real life that can achieve almost linear speedup. The most common examples are MVM and matrix-matrix multiplication. In this paper we focus on the performance of highly and easy parallelized MVM and avoid the matrix matrix multiplication performance drawbacks due to cache set associativity [4].

Cache memory can additionally improve the application execution because it reduces the gap between the main memory and the CPU speed. But several challenges exist if the CPU cores should use private (dedicated per core) or shared cache. Sharing the same cache among several cores can improve the memory access since loading a cache line (block) from the memory into the cache will reduce the other cores access time for some location in the same cache line [5]. On the other side, replacing some cache line for the requirements of a single core will increase the access time for the addresses of that cache line. We use multiprocessor with 4 chips with 4 cores each and utilizing different cores of different chips we want to determine the best CPU affinity (logical association with a given thread) to achieve maximum performance with the same number of CPU cores. We realized two different groups of experiments utilizing the same number of tightly and loosely coupled CPU cores. We have set a hypothesis that using the platform with more cache memory will achieve better performance since MVM is cache intensive algorithm, i.e. each element of the vector is read N times, where N denotes the size of the vector.

The paper is organized as follows. Section II presents the related research that we found in the literature. We present the testing methodology used for the experiments in order to provide reliable results in Section III. The results of the experiments are presented and discussed in Section IV. Finally, we conclude our work in Section V and present our plans for future work in Section VI .

II. Related Work

Soft and hard affinity policies exist in operating system (OS). The former is a policy where the OS assigns the threads to a certain core, while the latter is the policy where the developers assign the CPU affinity. In this paper we use hard affinity.

Several papers analyzed the multicore multiprocessor in order to improve the performance of threads scheduling them among specific cores. You and Zhao [6] exploit the performance of multi-core CPUs using hard affinity in OS in order

to reduce the response time of web servers. Chonka et al. [7] improve the efficiency of security application by restricting it in a specific core of multicore multiprocessor and adding ubiquitous multicore (UM) framework [8], i.e. classifying the applications into several classes and execute them on separate cores. They allow real-time multimedia using the same (UM) framework on multiprocessor [9]. Islam et al. [10] improved the performance running each classifier process in parallel within their dedicated core.

Broquedis et al. [11] even tripled the speedup using the hwloc (affinitiy) with OpenMP which focuses on cache memory while parallelized. Specifying the CPU affinity improves the performance due to data locality and cache memory existence. Ristov and Gusev [12] achieved even superlinear speedup (speedup greater than Gustafson's linear, i.e. greater than the number of processors) using hard affinity. Kosec and Depolli [13] explained the background of superlinearity measuring the cache hit and miss ratio while executed the same algorithm with sequential implementation with one core and parallel implementation with two cores.

CPU affinity is very important in virtual environment. Current definition of weight does not collaborate with CPU affinity and Chen et al. [14] improve Xen's credit scheduler as an affinity-aware one. Li et al. [15] improved the CPU utilization by analyzing and reducing cache misses between virtual and non-virtual environment and thus pining or unpining the virtual CPUs from the real physical CPU. Ristov et al. [16] achieved superlinear speedup in the cloud using environment variable GOMP_CPU_AFFINITY and OpenMP. Although full virtualization environments are slower than native [17], Gusev and Ristov [18] determined that cache intensive algorithms provide better performance if the problem size can fit in private cache of the cores.

III. Testing Methodology

In this section we present the testing methodology used for the experiments in order to provide reliable results. Each test case is tested upon its correctness.

A. Testing Algorithm

We use MVM algorithm defined in (1) as test data.

$$C_{N\cdot 1} = A_{N\cdot N} \cdot B_{N\cdot 1} \tag{1}$$

Each element c_i of result vector $C_{N\cdot 1}$ is calculated as inner product of row i of matrix $A_{N\cdot N}$ and vector $B_{N\cdot 1}$, for each $i = 0, 1, \cdots, N-1$. For simplification, we use squared matrix $A_{N\cdot N}$. Each matrix and vector element is stored as double precision number (8 bytes each).

B. MVM Algorithm Implementations

Two implementations of the MVM algorithm are used, i.e. the sequential and parallel implementation. The former consists of one thread which multiplies the whole matrix $A_{N\cdot N}$ and vector $B_{N\cdot 1}$. The latter uses $P \in \{2, 4\}$ threads. Each thread multiplies the row block matrix $A_{N/P\cdot N}$ and the whole vector $B_{N\cdot 1}$. We test the performance of those matrix and vector sizes which provides the best efficiency, i.e. the multiple of 4.

C. MVM Operations

In this section we analyze the total number of operations that the algorithm realizes in sequential and parallel execution. The single thread in sequential implementation multiplies each matrix row with the vector executing $N \cdot N$ multiplications and $N \cdot N$ additions, or total $2N^2$ floating point operations.

Parallel implementation of the algorithm is realized by P cores such that each core multiplies N/P rows executing $N/P \cdot N$ multiplications and $N/P \cdot N$ additions, or total $2N^2/P$ floating point operations. Therefore, all P cores will execute total $2N^2/P \cdot P$ or the same $2N^2$ as sequential implementation.

D. Testing Environment

The architecture of the multiprocessor consists of four CPUs AMD Opteron 8347 chips, each with four cores. Total sixteen cores are available with maximum usage of sixteen threads for parallel execution. Each core has 64KB its own private L1 (both data and instruction) cache and 512KB private L2 cache. Additionally, four cores of a chip share 2MB L3 cache.

Each parallel implementation has different cache organization for each experiment. Each thread has its own private L1 and L2 caches and shares main memory. The difference is in L3 cache, sometimes it is shared per threads, sometimes it is private per core, and sometimes it is shared but per two cores per chip. The following section explains this organization in more details for each experiment.

E. The Experiments

One experiment is realized for sequential execution and 2 groups of 2 experiments are realized for parallel execution with $P = 2$ and $P = 4$ threads. In each group of experiments we use the same number of threads (2 or 4) but different coupled, i.e. tightly and loosely coupled. From here thereafter, in this paper we will denote the sequential execution with experiment 1, parallel execution on tightly coupled cores with experiment 2.1 and 4.1 for 2 and 4 cores correspondingly, while parallel execution on loosely coupled cores with experiment 2.2 and 4.2 for 2 and 4 cores correspondingly.

Each experiment consists of series of test cases, i.e. MVM execution for different matrix and vector size N, starting from matrix size $N = 8$ to $N = 1800$ with step 4 for the best efficiency in each test case. In order to get reliable results, we repeat each test to run at least 10 seconds without using any optimization and calculate the average response time.

1) Experiment 1 - Sequential Execution: Sequential execution is used to calculate speedup for each experiment. It is realized with one thread on one core C0 of CPU0 as depicted in Fig. 2. C0 core of CPU0 with its private L1 and L2 caches is utilized. L3 Cache of CPU0 is also utilized and L4 main memory.

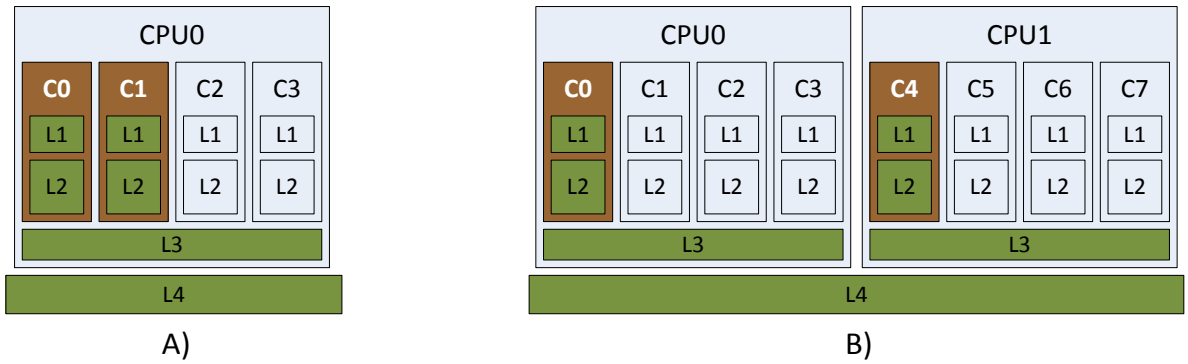


Fig. 1. Utilized cored and caches for experiments 2.1 (A) and 2.2 (B)

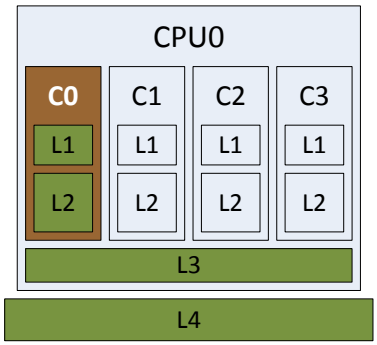


Fig. 2. Utilized cores and caches for experiment 1

2) Experiments with 2 Cores: In this section we present the experiments using total 2 cores, but different coupled.

Experiment 2.1 uses the two tightly coupled cores C0 and C1 of chip CPU0 as depicted in Fig. 1 A). We use 2 CPU cores on the same chip, i.e. L1 and L2 caches are private per core, and L3 Cache and L4 main memory are shared among two cores.

The same number of CPU cores (2) are used in *Experiment 1.2*, but with other CPU affinity, as depicted in Fig. 1 B). This experiment utilizes 2 loosely coupled CPU cores C0 and C4 on separate chips CPU0 and CPU1 correspondingly, i.e. L1 and L2 caches are private per core, the whole L3 cache is private per core and chip, and L4 main memory is shared among two threads.

3) Experiments with 4 Cores: We create two different experiments using 4 CPU cores, also with different CPU affinity. The *Experiment 4.1* uses tightly coupled 4 CPU cores on the same chip, as depicted in Fig. 3 A). L1 and L2 caches are private per core, while L3 Cache and L4 main memory are shared per core and chip.

Experiment 2.2 uses the same number of processors (4) but with different affinity as depicted in Fig. 3 B). This experiment utilizes loosely coupled 4 CPU cores on 4 separate chips CPU0-CPU3. L1 and L2 caches are also private per core, L3 is private per chip and therefore per thread, and L4 cache - main memory is shared among all 4 threads.

F. Test Data

Execution time $T(P)$ is measured for each test case (different matrix size N) in each experiment (different number of processors P and different CPU affinity). Then we calculate the *Speed* $V(P)$ and the *Speedup* $S(P)$ using definitions (2) and (3) [1].

The speed $V(P)$ is calculated by dividing the number of operations and measured execution time using $P \in \{1, 2, 4\}$ threads. We divide also the speed with 10^9 to express it in gigaflops.

$$V(P) = \frac{2 \cdot N^2}{T(P) \cdot 10^9} \tag{2}$$

The speedup $S(P)$ is calculated for parallel execution using $P \in \{2, 4\}$) as a ratio of sequential and parallel execution time by (3).

$$S(P) = \frac{T(1)}{T(P)} \tag{3}$$

G. Test Goals

The test experiments have two goals. The first goal is to model the speed and speedup as a functions of matrix and vector size for MVM algorithm using loosely and tightly coupled cores. The second goal is to determine which experiments provides better speed and speedup.

IV. The Results

In this section we present the achieved speed and speedup of each test case of all 5 experiments defined in Section III-E and compare them for particular group of experiments with parallel implementation, i.e. tightly with loosely coupled counterparts.

A. Speed Analysis

In this section we present the results for the achieved speed in all five experiments. We compare also the speeds of the same group of experiments.

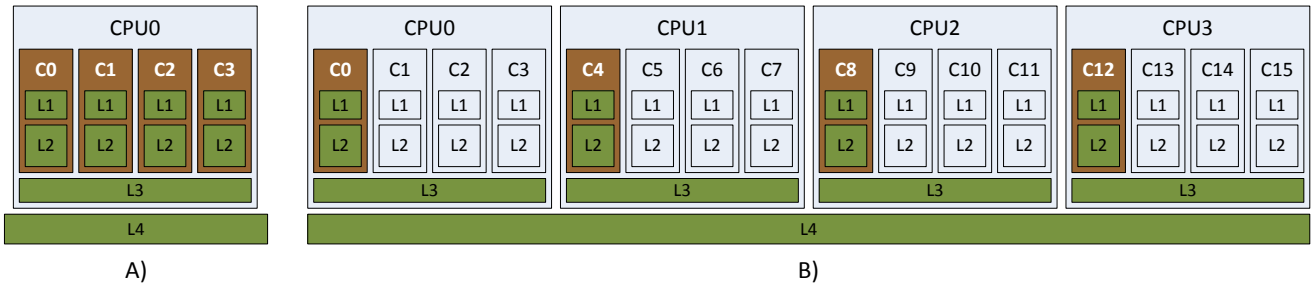


Fig. 3. Utilized cored and caches for experiments 4.1 (A) and 4.2 (B)

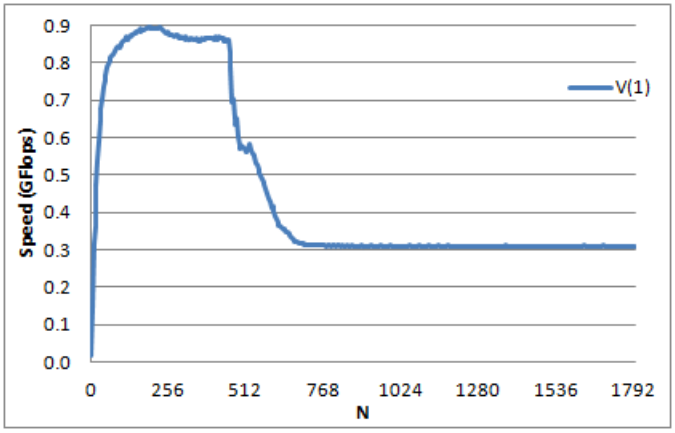


Fig. 4. Achieved speed in experiment 1

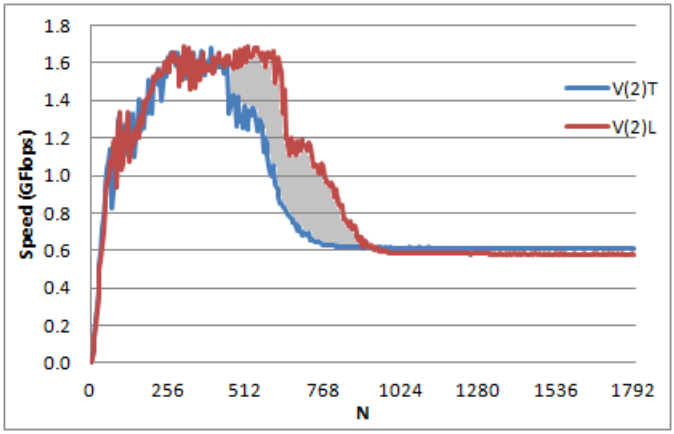


Fig. 5. Achieved speed in experiment 2

1) Speed for Experiment 1: The speed $V(1)$ for sequential execution of MVM algorithm is depicted in Fig. 4. We clearly observe four regions R_1, R_2, R_3 and R_4 going from left to the right with different speed behavior. R_1 region is determined with $N \leq 256$ and the speed increases in this region until it reaches its maximum. Then the speed saturates in the R_2 region for $256 < N \leq 460$. In R_3 region, the speed significantly decrease until $N < 652$ and then saturates to R_4 region.

2) Speed for Experiments 2.1 and 2.2: Fig. 5 depicts the speed $V(2)$ for both tightly and loosely coupled experiments 2.1 and 2.2, correspondingly. We also observe four regions with the same speed behavior as $V(1)$ for both experiments, but with different matrix and vector size ranges.

Table I presents the ranges for the regions for both experiments. Both speeds have similar values for $N \leq 440$, i.e. in the regions R_1 and R_2 for tightly coupled experiment. However, the region R_2 where the speed is saturated in its maximum values is much wider for loosely coupled experiment until $N = 640$. The loosely coupled experiment has greater speed until its R_3 region has finished for $N = 894$. We explain this behavior because loosely coupled cores use doubled L3 cache capacity and smaller number of cache misses are generated. However, the tightly coupled experiment achieves better speed in R_4 region for $N > 894$ due to communication bound in loosely coupled experiment compared to tightly coupled experiment.

TABLE I
SPEED REGIONS FOR EXPERIMENT2 2.1 AND 2.2

Exp.	R_1	R_2	R_3	R_4
2.1	[4,256]	[256,440]	[440,768]	[768, 1800]
2.2	[4,256]	[256,640]	[640,894]	[894, 1800]

We observe that both curves are a little discrepant comparing to the speed $V(1)$.

3) Speed for Experiments 4.1 and 4.2: Similar speed behavior is observed for experiments 4.1 and 4.2 as the experiments with 2 cores, as depicted in Fig. 6. Table II presents the ranges for the regions for different speed behavior for both experiments. Both speeds have the same values until $N = 540$, i.e. in regions R_1 and R_2. Then $V(4)T$ starts decreasing, while the speed $V(4)L$ remains constant until the end of its own R_2 region. Finally, $V(4)T > V(4)L$ in R_4 region due to the same reason explained in previous section.

Also both curves are a little discrepant comparing to the speed $V(1)$.

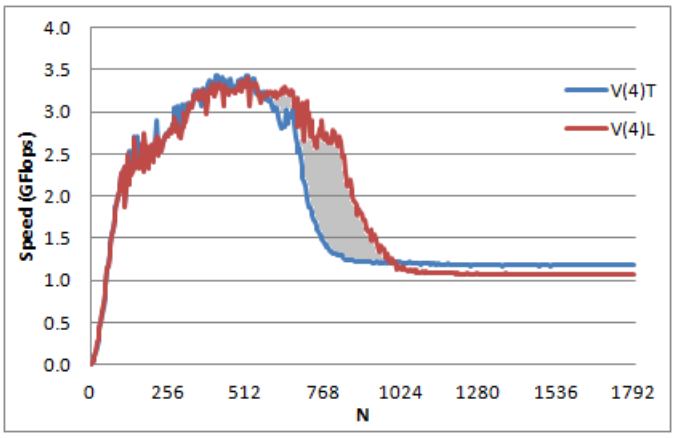


Fig. 6. Achieved speed in experiment 4

TABLE II
SPEED REGIONS FOR EXPERIMENTS 4.1 AND 4.2

Exp.	R_1	R_2	R_3	R_4
4.1	[4,392]	[392,540]	[540,832]	[832, 1800]
4.2	[4,392]	[392,720]	[720,972]	[972, 1800]

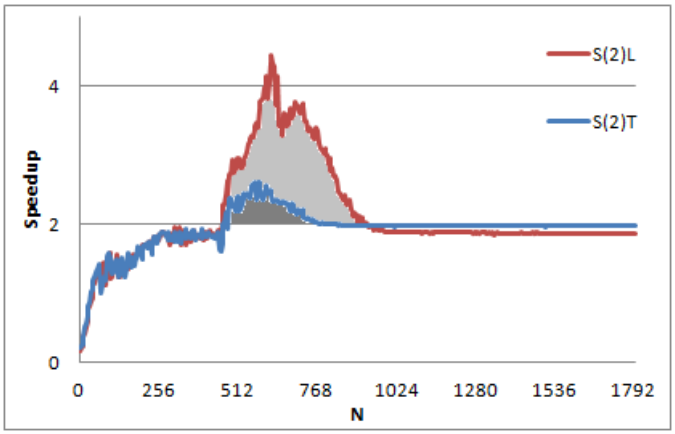


Fig. 7. Speedup regions for experiments 2.1 and 2.2

B. Speedup Analysis

Speedup has different behavior than the speed for both group of experiments. In this section we present the results and compare the speedup of the counterparts in each group of experiments.

1) Speedup for Experiments 2.1 and 2.2: Fig. 7 depicts the achieved speedup for the experiments 2.1 and 2.2, where $S(2)T$ denotes the speedup for tightly coupled 2 cores and $S(2)L$ denotes the speedup for loosely coupled 2 cores. We also observe four regions with different speedup behavior whose ranges are presented in Table III.

Both speedups increase in R_1 and R_2 regions. Then their values start to decrease in R_3 region and saturates in R_4 region. The regions R_2 and R_3 together are determined as superlinear speedup region for both speedups, i.e. the speedup is greater than the number of processors $S(2) > 2$.

TABLE III
SPEEDUP REGIONS FOR EXPERIMENTS 2.1 AND 2.2

Exp.	R_1	R_2	R_3	R_3
2.1	[4,460]	[460,564]	[564,766]	[766,1800]
2.2	[4,460]	[460,620]	[620,904]	[904,1800]

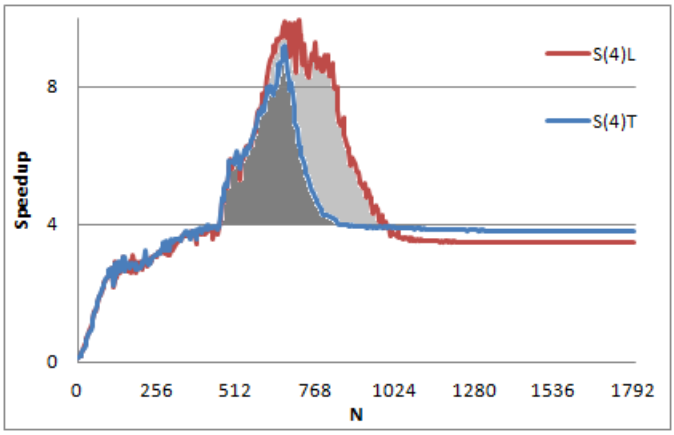


Fig. 8. Speedup regions for experiments 4.1 and 4.2

TABLE IV
SPEEDUP REGIONS FOR EXPERIMENTS 4.1 AND 4.2

Exp.	R_1	R_2	R_3	R_4
2.1	[4,442]	[442,664]	[664,844]	[844,1800]
2.2	[4,442]	[442,664]	[664,844]	[992,1800]

We can conclude that both speedups have similar values in R_1 region. Loosely coupled experiment provides much greater speedup in superlinear R_2 and R_3 regions. However, $S(2)L < S(2)T < 2$ in R_4 region.

2) Speedup for Experiments 4.1 and 4.2: The achieved speedup for the experiments 4.1 and 4.2 is depicted in Fig. 8. We denote the speedup for tightly coupled 4 cores with $S(4)T$ and $S(4)L$ denotes the speedup for loosely coupled 4 cores. We observe the same four different regions for both curves, as presented in Table IV.

We can conclude that both speedups have similar values in R_1 region, but also in almost whole R_2 region. Loosely coupled experiment provides greater speedup in superlinear R_3 region. However, again $S(4)L < S(4)T < 4$ in R_4 region.

V. CONCLUSION

Scaling the MVM algorithm on greater number of processors speedups its execution. However, the achieved speedup does not follows the the number of cores for different CPU affinity, i.e. tightly or loosely coupled CPU cores. The reason lies in different cache capacity and organization, especially the last level cache which impact the most to the algorithm performance. In this paper we analyze the performance of MVM algorithm as an algorithm which is granular and scalable. Our intention is not to improve the algorithm, but to determine if tightly or loosely coupled CPU cores will provide better performance using the same number of cores for

particular matrix and vector size. We used 2 different groups of experiments using 2 and 4 cores on same or different CPU chips (dies).

Three different regions are determined comparing both counterparts with the same number of cores for both experiments with 2 and 4 CPU cores. Both counterparts provides the same performance (speed) when the data can be placed completely in the one L3 cache. Increasing the problem size will degrade the performance of tightly coupled cores since the loosely coupled L3 capacity is doubled. When the problem size reaches the doubled L3 size, the loosely coupled cores provide worse performance than tightly coupled.

Also three different regions exist comparing the speedup for the counterparts with the same number of cores for both experiments with 2 and 4 CPU cores. Both counterparts achieve the same speedup when the data can be placed completely in the one L3 cache. Increasing the problem size provides superlinear speedup for both counterparts, with loosely coupled in front of tightly coupled. The speedups are similar for greater problem size when huge number of last level L3 cache misses are generated and the tightly coupled counterpart achieves better speedup.

We can conclude that working on loosely coupled cores with private L3 cache is better than working on tightly cores with shared last level L3 cache in the region where the problem size can be stored in the greater L3 cache, but in the same time cannot be placed in only one cache. CPU affinity does not affect the performance when the problem size fits in L1 or L2 cache if they are private per core. For greater problem size other algorithms should be used that achieve greater speed.

VI. Future Work

In this paper we analyzed MVM algorithm on single and multi chip multiprocessor. We plan to analyze if other linear algebra algorithms behave the same in the manner of CPU affinity (loosely or tightly coupled), especially matrix-matrix multiplication since all elements are accessed N times and the second matrix B storage pattern in the cache memory is not the same as the first matrix A. Another important research is to analyze the CPU affinity in virtualized cloud environment.

References

[1] J. L. Hennessy and D. A. Patterson, *Computer Architecture, Fifth Edition: A Quantitative Approach*. MA, USA: Elsevier, 2012.

[2] J. L. Gustafson, "Reevaluating Amdahl's law," *Communication of ACM*, vol. 31, no. 5, pp. 532–533, May 1988.

[3] G. M. Amdahl, "Validity of the single-processor approach to achieving large scale computing capabilities," in *AFIPS Conference Proceedings*, vol. 30. AFIPS Press, Reston. Va., Atlantic City, N.J., Apr. 18-20 1967, pp. 483–485.

[4] M. Gusev and S. Ristov, "Performance gains and drawbacks using set associative cache," *Journal of Next Generation Information Technology (JNIT)*, vol. 3, no. 3, pp. 87–98, 31 Aug 2012.

[5] V. Olaru, A. Hangan, and G. Sebestyen-Pal, "Java support packages and benchmarks for multi-core processors," in *Proceedings of the 2011 IEEE International Conference on High Performance Computing and Communications*, ser. HPCC '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 528–535.

[6] G. You and Y. Zhao, "Dynamic requests scheduling model in multi-core web server," in *Grid and Cooperative Computing (GCC), 2010 9th International Conference on*, nov. 2010, pp. 201 –206.

[7] A. Chonka, W. Zhou, K. Knapp, and Y. Xiang, "Protecting information systems from ddos attack using multicore methodology," in *Proceedings of the 2008 IEEE 8th International Conference on Computer and Information Technology Workshops*, ser. CITWORKSHOPS '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 270–275.

[8] A. Chonka, S. K. Chong, W. Zhou, and Y. Xiang, "Multi-core defense system (msds) for protecting computer infrastructure against ddos attacks," in *Proceedings of the 2008 Ninth International Conference on Parallel and Distributed Computing, Applications and Technologies*, ser. PDCAT '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 503–508.

[9] A. Chonka, W. Zhou, L. Ngo, and Y. Xiang, "Ubiquitous multicore (um) methodology for multimedia," *Computer Science and its Applications, International Symposium on*, vol. 0, pp. 131–136, 2008.

[10] R. Islam, W. Zhou, Y. Xiang, and A. N. Mahmood, "Spam filtering for network traffic security on a multi-core environment," *Concurr. Comput. : Pract. Exper.*, vol. 21, no. 10, pp. 1307–1320, Jul. 2009.

[11] F. Broquedis, J. Clet-Ortega, S. Moreaud, N. Furmento, B. Goglin, G. Mercier, S. Thibault, and R. Namyst, "hwloc: A generic framework for managing hardware affinities in hpc applications," in *Proceedings of the 2010 18th Euromicro Conference on Parallel, Distributed and Network-based Processing*, ser. PDP '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 180–186.

[12] S. Ristov and M. Gusev, "Superlinear speedup for matrix multiplication," in *Information Technology Interfaces, Proceedings of the ITI 2012 34th International Conference on*, 2012, pp. 499–504.

[13] G. Kosec and M. Depolli, "Superlinear speedup in openmp parallelization of a local pde solver," in *MIPRO, 2012 Proceedings of the 35th International Convention*, May, pp. 389–394.

[14] H. Chen, H. Jin, and K. Hu, "Affinity-aware proportional share scheduling for virtual machine system," in *Proceedings of the 2010 Ninth International Conference on Grid and Cloud Computing*, ser. GCC '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 75–80.

[15] Z. Li, Y. Bai, H. Zhang, and Y. Ma, "Affinity-aware dynamic pinning scheduling for virtual machines," in *Proceedings of the 2010 IEEE Second International Conference on Cloud Computing Technology and Science*, ser. CLOUDCOM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 242–249.

[16] S. Ristov, M. Gusev, M. Kostoska, and K. Kiroski, "Virtualized environments in cloud can have superlinear speedup," in *Proceedings of the Fifth Balkan Conference in Informatics*, ser. BCI '12. New York, NY, USA: ACM, 2012, pp. 8–13.

[17] K. Ibrahim, S. Hofmeyr, and C. Iancu, "Characterizing the performance of parallel applications on multi-socket virtual machines," in *Cluster, Cloud and Grid Computing (CCGrid), 2011 11th IEEE/ACM International Symposium on*, may 2011, pp. 1 –12.

[18] M. Gusev and S. Ristov, "Matrix multiplication performance analysis in virtualized shared memory multiprocessor," in *MIPRO, 2012 Proceedings of the 35th International Convention, IEEE Conference Publications*, 2012, pp. 264–269.

Computation and Analysis of Sobol Coefficients for Air Pollution Concentrations over the Territory of Bulgaria

Emanouil Atanassov and Sofiya Ivanovska
Department of Grid Technologies and Applications
Institute of Information and Communication Technologies - Bulgarian Academy of Sciences
Acad. G. Bonchev St., Bl.25A, 1113 Sofia, Bulgaria
{emanouil, sofia}@parallel.bas.bg

***Abstract*— One of the main tools for modeling air pollutions over the territory of Bulgaria is the US EPA Models-3 system. The main components of the system are MM5/WRF meteorological preprocessor, SMOKE emission preprocessor and CMAQ chemical transport model. TNO emission inventory is used as emission input. The Models-3 "Integrated Process Rate Analysis" option is applied to discriminate the role of different dynamic and chemical processes for the pollution for all SNAP categories. In this work we evaluate the influence of the different input parameters as concentration in different SNAP categories under constant meteorological conditions to the output concentrations over Bulgaria following methodology of Sobol-Saltelli. In order to obtain reliable estimates of the Sobol coefficients we perform a large of MPI jobs using the clusters in South-Eastern region. Using these coefficients we assess the relative importance of the various input parameters, and their interactions.**

I. Introduction

A. Air pollution problem

Air Quality (AQ) is a key element for the well-being and quality of life of European citizens. Bulgaria also faces AQ problems. It should be noted that, while in Western Europe the photo-oxidant and PM air pollution is at present the major environmental problem, in Bulgaria the classic acidifying pollutants (SO2, NOx), the heavy metals (Hg, Cd, Pb) and the persistent organic pollutants are still a serious problem and so the study of their environmental impact is absolutely necessary. The reduction of the emissions of these compounds is a major task in the environmental policy of the country.

The Common Air Quality Index (CAQI) is designed to present and compare air quality in near-real time on an hourly or daily basis. The CAQI has 5 levels, using a scale from 0 (very low) to >100 (very high) and the matching colours range from light green to dark red. The air pollution is modeled by combining meteorology data and outputs from meteorology models, emissions from neighbouring countries and data about emissions inside the country [7], [8].

The emissions that contribute to air pollution are classified following the following the so-called SNAPs (Selected Nomenclature for Air Pollution). In our studies we denote by x_1 to x_{10} the inputs for the corresponding SNAP category:

x_1	Combustion in energy and transformation industries
x_2	Non-industrial combustion plants
x_3	Combustion in manufacturing industry
x_4	Production processes
x_5	Extraction and distribution of fossil fuels and geothermal energy
x_6	Solvents and other product use
x_7	Road transport
x_8	Other mobile sources and machinery
x_9	Waste treatment and disposal
x_{10}	Agriculture

In order to simulate the impact of the emissions from different SNAP categories on the air pollution of Bulgaria, we use a modification of the US EPA Models-3 system. For the input emission data we use the TNO inventory with $0.25^\circ \times 0.125^\circ$, the inventory contains 8 pollutants: CH4, CO, NH3, NMVOC (VOC), NOx, SOx, PM10 and PM2.5. TNO emissions are distributed over the 10 SNAPs classifying pollution sources according to the processes leading to harmful material release to the atmosphere.The input Meteorological data is from NCAR $1^\circ \times 1^\circ$. The MM5 - the 5th generation PSU/NCAR Meso - meteorological Model MM5 ([5], [6]) is used as a meteorological pre-processor; CMAQ - the Community Multi-scale Air Quality System (Byun et al., 1998, [3]), is the Chemical Transport Model (CTM) of the system, and SMOKE - The Sparse Matrix Operator Kernel Emissions Modelling System [4] - the emission pre-processor of Models-3 system. We note that the Models-3 "Integrated Process Rate Analysis" option is applied to discriminate the role of different dynamic and chemical processes for the pollution. The goal of the present study is to evaluate the influence of the different input parameters as concentration in different SNAP categories under constant meteorological conditions to the output concentrations over Bulgaria in several key points - the cities of Sofia, Goce Delchev, Vraca, Pleven, Rojen, Plovdiv, Burgas and the Konstantin and Elena resort [9], [10].

The model is re-run under constant meteorological conditions and varying the values of the SNAP categories, assuming uniformly random distribution of the correction

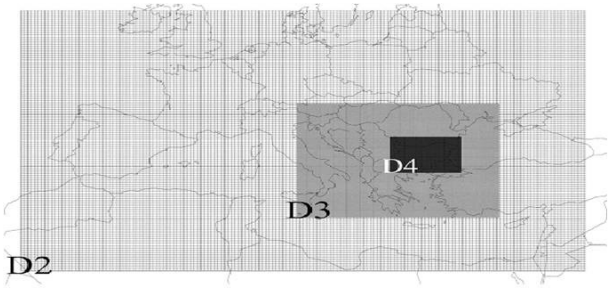


Fig. 1. The three inner computational domains

multiplier, between 0 and 1, for each category. The multi-scale capability of the model system means that we can perform analysis targeted at specific points as outlined above, as well as computing more integrated quantities over larger areas. The Table I below shows the relative scales, reaching 4 levels. The graphic representation of the computational domains and the results that are typically obtained with a single model run can be seen on Figure 1.

	D1	D2	D3	D4
MM5 (km)	81	27	9	3
2 way nesting CMAQ (km)		27	9	3
Grid dimension		166× 115	178 × 151	190 × 140

TABLE I
RELATIVE SCALES

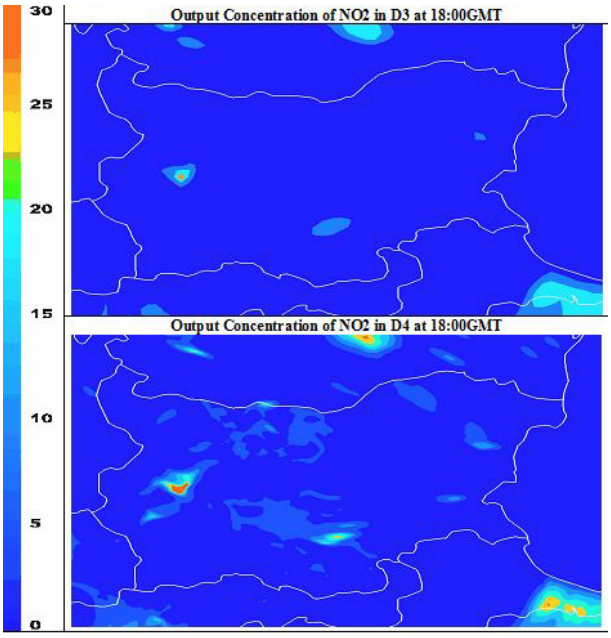


Fig. 2. Concentration of NO2 in D3

B. Sensitivity Analysis Studies

We use in our work the particular methodology of estimating the influence of input parameters to the output parameters of the model, developed initially by Sobol, [12] and later improved by Saltelli, [11].

The mathematical model assumes that an output function

$$\mathbf{u} = f(\mathbf{x}), \text{where} \quad \mathbf{x} = (x_1, x_2, \ldots, x_d) \in U^d \equiv [0,1]^d$$

is a vector of inputs with a joint p.d.f. $p(\mathbf{x}) = p(x_1, \ldots, x_d)$. Total Sensitivity Index of input parameter $x_i, i \in \{1, \ldots, d\}$:

$$S_{x_i}^{tot} = S_i + \sum_{l_1 \neq i} S_{il_1} + \sum_{l_1, l_2 \neq i, l_1 < l_2} S_{il_1 l_2} + \ldots + S_{il_1 \ldots l_{d-1}},$$

where S_i - *the main effect (first-order sensitivity index)* of x_i and $S_{il_1 \ldots l_{j-1}} - j^{\text{th}}$ *order sensitivity index for parameter* x_i $(2 \leq j \leq d)$.

ANalysis Of VAriances (ANOVA) HDMR decomposition of a square integrable function $f(\mathbf{x})$ is obtained as follows:

$$f(\mathbf{x}) = f_0 + \sum_{\nu=1}^{d} \sum_{l_1 < \ldots < l_\nu} f_{l_1 \ldots l_\nu}(x_{l_1}, x_{l_2}, \ldots, x_{l_\nu}),$$

where $f_0 = const$, and

$$\int_0^1 f_{l_1 \ldots l_\nu}(x_{l_1}, x_{l_2}, \ldots, x_{l_\nu}) \mathrm{d}x_{l_k} = 0,$$
$$1 \leq k \leq \nu, \nu = 1, \ldots, d.$$

The functions in the right-hand side are defined in a unique way:

$$f_0 = \int_{U^d} f(\mathbf{x}) \mathrm{d}\mathbf{x},$$

$$f_{l_1}(x_{l_1}) = \int_{U^{d-1}} f(\mathbf{x}) \prod_{k \neq l_1} \mathrm{d}\mathbf{x}_k - f_0, \; l_1 \in \{1, \ldots, d\}$$

$$\int_{U^d} f_{i_1 \ldots i_\mu} f_{j_1 \ldots j_\nu} \; \mathrm{d}\mathbf{x} = 0,$$

$$(i_1, \ldots, i_\mu) \neq (j_1, \ldots, j_\nu), \quad \mu, \nu \in \{1, \ldots, d\}.$$

In our case the input parameters $x_1, \ldots, x_{10}$ are the multipliers, which decrease the emissions in the corresponding SNAP categories, that assumed uniformly distributed in the interval $[0, 1]$. The Sobol' Global Sensitivity Indices are defined as follows:

Definition 1:

$$S_{l_1 \ \ldots \ l_\nu} = \frac{\mathbf{D}_{l_1 \ \ldots \ l_\nu}}{\mathbf{D}}, \quad \nu \in \{1, \ldots, d\},$$

where partial variances

$$\mathbf{D}_{l_1 \ \ldots \ l_\nu} = \int f^2_{l_1 \ \ldots \ l_\nu} \mathrm{d}x_{l_1} \ldots \mathrm{d}x_{l_\nu},$$

total variance

$$\mathbf{D} = \int_{U^d} f^2(\mathbf{x}) \mathrm{d}\mathbf{x} - f_0^2, \quad \mathbf{D} = \sum_{\nu=1}^{d} \sum_{l_1 < \ldots < l_\nu} \mathbf{D}_{l_1 \ldots l_\nu},$$

They satisfy the following properties:

$$S_{l_1 \ \ldots \ l_s} \geq 0, \quad \sum_{s=1}^{d} \sum_{l_1 < \ldots < l_s}^{d} S_{l_1 \ \ldots \ l_s} = 1.$$

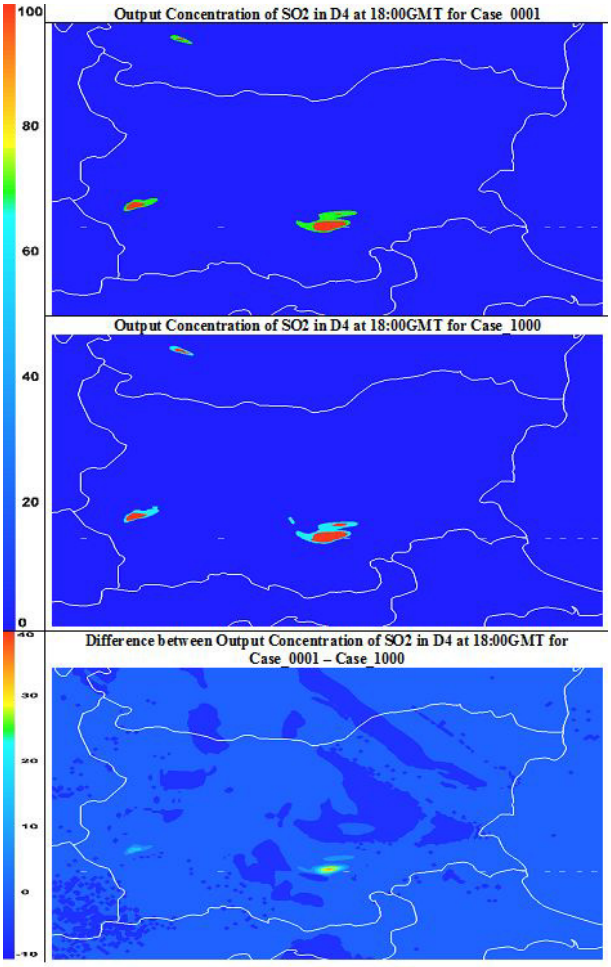


Fig. 3. Concentration of SO2 in D4

In our study we concentrate on the first order coefficients, on one hand, as well as the total coefficients. In this way we can estimate the amount of linearity in the model and the relative importance of the various SNAP inputs to the final outputs of the model. In the next section we present further details about our methodology and the achieved numerical results and the we show our conclusions.

II. Methodology and Numerical Results

The main computational resource in the Bulgarian Grid infrastructure is the HPC Cluster at IICT-BAS. It has HP Cluster Platform Express 7000 enclosures with 36 blades BL 280c with dual Intel Xeon X5560 @ 2.8Ghz (total 576 cores), 24 GB RAM per blade, 8 controlling nodes HP DL 380 G6 with dual Intel X5560 @ 2.8 Ghz, 32 GB RAM. The total amount of storage available in two SAN disk systems is 96 TB, organized in two Lustre filesystems for the shared home and scratch directories.

The multi-scale numerical experiments and the various extensive input data (emissions, boundary conditions, large scale forcing) requires not only computational power but also have substantial temporary and permanent storage requirements. The HPC cluster is interconnected with non-blocking DDR Infiniband via the core Voltaire Grid director 2004 switch, which allows it to obtain more than 92% efficiency on LINPACK. Since the computations for each set of input parameters are independent, we found that the optimal configuration for completing our research was to use all the physical cores in one blade in parallel for one job. That is why such type of computations are suitable for Grid implementation, provided that the target site supports MPI, [1], [2].

For our purposes 1200 output data file were prepared, generated after 1200 data points on 10-dimensional unit cube were generated. The output were interpolated using Radial Base Function [14]. By using such interpolation we prepared the input data necessary for our computations in SimLab [13]. The computations of the Sobol sensitivity indices were carried out using procedure from SimLab and the results are shown in the next two tables. The Table II shows the computed Sobol first order indices, outlining the most influential input parameter for each output. These results are obviously dependent on our particular setting, but it seems that the contribution of the first SNAP is most relevant in half of the categories. Still the first order indexes seem to be rather low compared with the total indexes, which are shown in Table III. This finding underlines the non-linearity of the contribution of the SNAPs to the air pollution measurements. Here we only show the results for Sofia, but there is high degree of similarity for the other points of interest. Obviously, if the same computations are performed for points that are near to important sources of some particular SNAPs, the situation will be qualitatively different.

Output	NO2	O3	CO	SO2	PM2.5	PM10
Maximum first order influence	x_8	x_4	x_4	x_1	x_1	x_1
Value	0.281005	0.029882	0.053433	0.09856	0.079219	0.068339

TABLE II
First order(S_i) sensitivity indices

III. Conclusion

Our experience with these computations shows that they require high amount of computational time in order to achieve sufficiently accurate results about the Sobol' coefficients. That is why decreasing the relative error of the obtained coefficients will require much higher computational capacity, which is currently on the limit of what is totally available around all the Grid resources of South Eastern Europe, which are now in the European Grid Infrastructure. So far we found that the first level coefficients are in most cases relatively small and thus the level of non-linearity in the model is relatively high. Also we found that the variability of the total indices with respect to the various SNAPs are not as high as expected. Nevertheless we found which inputs in

	NO2	O3	CO	SO2	PM2.5	PM10
x_1	0.235723	0.372304	0.354357	0.414128	0.374604	0.353887
x_2	0.223580	0.289342	0.271309	0.262395	0.276314	0.261679
x_3	0.253028	0.298078	0.333387	0.287030	0.280722	0.285029
x_4	0.246810	0.305317	0.324893	0.283457	0.320501	0.336260
x_5	0.217021	0.259109	0.271314	0.237278	0.233550	0.237142
x_6	0.223928	0.305098	0.268494	0.307680	0.243665	0.237864
x_7	0.182537	0.280958	0.234788	0.278122	0.222028	0.212509
x_8	0.474537	0.300624	0.258790	0.269298	0.250679	0.241850
x_9	0.211635	0.306437	0.262087	0.274366	0.245141	0.237476
x_{10}	0.217746	0.311945	0.272089	0.299062	0.236402	0.233778

TABLE III
TOTAL(S_{Ti}) SENSITIVITY INDICES

our case are the most influential in our setting. Although the results shown are for a particular geographic position (the city of Sofia), the situation is similar for the other points, for which we performed the same types of computations. The structure of the tables that we obtained shows that the interactions between the various SNAPs are complex and thus the political decisions that affect levels of emissions and consequently levels of air pollutants are effectively a multi-criterial optimization.

ACKNOWLEDGMENT

This work was supported by the National Science Fund of Bulgaria under Grants DCVP 02/1, DMU 03/61 and the European Commission under EU FP7 project EGI-InSPIRE (contract number RI-261323).

REFERENCES

[1] E. Atanassov, D. Dimitrov and S. Ivanovska, "Efficient Implementation of the Heston Model Using GPGPU", J. of Monte Carlo Methods and Applications, J. of Monte Carlo Methods and Applications, 21-28, 2003.

[2] E. Atanassov, S. Ivanovska and A. Karaivanova, "Tuning the Generation of Sobol Sequence with Owen scrambling", LNCS 2010, Volume 5910, Springer, pp. 459-466, 2010.

[3] D. Byun, J. Ching, "Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System", EPA Report 600/R-99/030, Washington DC, 1999.

[4] CEP, Sparse Matrix Operator Kernel Emission (SMOKE) Modeling System, University of Carolina, Carolina Environmental Programs, Research Triangle Park, North Carolina, 2003.

[5] J. Dudhia, "A non-hydrostatic version of the Penn State/NCAR Mesoscale Model: validation tests and simulation of an Atlantic cyclone and cold front.", Monthly Weather Review, vol. 121, 1493-1513, 1993.

[6] G.A. Grell, J. Dudhia, and D.R. Stauffer, "A description of the Fifth Generation Penn State/NCAR Mesoscale Model (MM5)", NCAR Technical Note, NCAR TN-398-STR, 138 pp., 1994.

[7] K. Ganev, D. Syrakov, G. Gadzhev, M. Prodanova, G. Jordanov, N. Miloshev, A. Todorova, Joint Analysis of Regional Scale Transport and Transformation of Air Pollution from Road and Ship Transport, Large-Scale Scientific Computing, LSSC 2009, LNCS, vol. 5910, pp.180-187, Springer, 2009.

[8] A. Todorova, G. Gadzhev, G. Jordanov, D. Syrakov, K. Ganev, N. Miloshev and M. Prodanova, Numerical Study of Some High PM10 Levels Episodes, Large-Scale Scientific Computing, LSSC 2009, LNCS vol. 5910, pp.223-230, Springer, 2009.

[9] G. Gadzhev, G. Yordanov, K. Ganev, M. Prodanova, D. Syrakov, N. Miloshev, Atmospheric Composition Studies for the Balkan Region, Large-Scale Scientific Computing, LSSC 2010, LNCS, vol. 6046, pp.150157, Springer, 2011.

[10] Georgi K. Gadzhev, Kostadin G. Ganev, Nikolay G. Miloshev, Dimiter E. Syrakov, Maria Prodanova, - Numerical study of the atmospheric composition in Bulgaria Computers and Mathematics with Applications (2012), DOI: 10.1016/j.camwa.2012.07.002, Volume 65, Issue 3, February 2013, pp. 402422.

[11] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola,, "Global Sensitivity Analysis. The Primer", John Wiley & Sons, 2008.

[12] Sobol, I.M.: Global Sensitivity Indices for Nonlinear Mathematical Models and Their Monte Carlo Estimates. Mathematics and Computers in Simulation 55(1-3), 271280, 2001.

[13] SIMLAB, http://simlab.jrc.ec.europa.eu/

[14] http://www.mathworks.com/matlabcentral/fileexchange/10056

Algorithms for computational procedure acceleration for systems differential equations in matlab.

Hr. B. Nenov*, B. Hr. Dimitrov** and A.St. Marinov***
* Technical university - Varna / "Computer Science and Engineering", Varna, Bulgaria
**Technical University of Varna/Department of electrical engineering, Varna, Bulgaria
***Technical University of Varna/Department of electronics and microelectronics, Varna, Bulgaria
ico762001@gmail.com, bdimitrov@processmodeling.org, a.marinov@tu-varna.bg

Abstract - Most engineers use computational environments from the MATLAB family as a favorite tool for solving problems. The combination of being relatively easy to use and its might of calculations make it so attractive. A significant problem in solving more complex tasks is the time needed for obtaining results. Different approaches are used in order to reduce it to an acceptable range of values - distributed systems, multiprocessor systems and others. In this paper we are discussing other useful approaches for solving the above-mentioned problem. The optimization of the MATLAB code in terms of effectively allocating memory, vectorization, parallelism and using a Graphical processor unit for calculations improve the solving process and dramatically decrease processing time. Thus, the need of powerful computing infrastructures such as grid systems or superarchitectures is reduced.

I. INTRODUCTION

The solution of engineering problems related to the implementation of different computational procedures. Most often this is done by specialized software. Complexity and scale of the calculations vary depending on the specifics of the object of the research. The software designed for this purpose is different - from a simple computational application, to software packages with a wide range of tools supporting its own programming language, enabling to solve a very wide range of tasks. That kind of software is Matlab. Matlab is a high-level language and interactive environment for numerical computation, visualization, and programming. Using Matlab, you can analyze data, develop algorithms, and create models and applications. You can use Matlab for a range of applications, including signal processing and communications, image and video processing, control systems, test and measurement, computational finance, and computational biology[6]. Many engineers from many different scientific fields use this tool to solve their problems.

There are two ways for code generating in Matlab. The first one is to receive automatically generated code, from some Matlab tool. The second one is the code to be written by the user[4][6]. In both cases, usually, the result is inefficient code. Unoptimized code is characterized by inefficient use of resources, memory, processor, etc. The result of this is the increasing, sometimes dramatically, of the time for computations. If this increase is within the acceptable limit, the problem is not so serious, but often when much complicated procedures or computational simulations of complex objects computation are solved, the time is going beyond the permissible limits. This is exactly the core problem, which leads to the following consequences:

- making the process of simulation too long;
- impossibility of making the required number of computational procedures to build mathematical sample;
- use of powerful computing architectures, which increases the cost of the research.

In this paper are discussed steps for optimizing the code in order to avoid the above mentioned problems.

II. PROBLEMS RELATED TO EFFICIENCY

A. Analysis of program performance

Matlab "profiler" is standard tool for analysis of performance. It is a good first step for understanding where the problem with performance of the code is. The information from it is helpful to investigating the possible bottlenecks.

a) Problems with memory using

One of the basic problems with the Matlab programing is inefficient use of computer memory. Creating variables in Matlab environment is often overlooked by users. Specificity of Matlab is that each variable is created in the form of a matrix. Even a simple integer variable type is represented as a matrix with one row and one column. For small "*n*" size variables and single use, that is not really a big problem. In other cases, however, especially when there is a multiple use of a variable which is changing its size in loop many times, the result will be significant drop in efficiency and performance of the program.

b) Inefficient structure of source code

Due to its proximity to scripting languages, principles of implementation (m file - Matlab script), Matlab predisposes to linear writing and code execution. There are cases where this is unavoidable, because the used function or functions have linear properties. In many

other cases, this is able to be changed. For operations with large numbers large-dimension arrays or multiple computational procedures the linear method is ineffective and dramatically increases the time for computations.

III. OPTIMIZATION

In computer science, program or software optimization is the process of modifying a software system to make some aspects of its work more efficiently or use fewer resources [1]. In general, a computer program may be optimized so that it executes more rapidly, or is capable of operating with less memory storage or other resources, or draw less power.

After a thorough analysis of the Matlab documentation and many experiments with the set of features and tools which it offers, were synthesized the following optimization steps for writing a Matlab program.

A. Analysis of the task and the possibilities for implementation in Matlab

The purpose of programming is to create a set of instructions that computers use to perform specific operations or to exhibit desired behaviors. There is two ways for Matlab code creation:

- user generated code, which is formalized description of the problem by Matlab language;
- automatic generation of code, using some of the Matlab tools. Most often these tools are modules for simulation of processes.

In the first case, the goal is to find functions and commands in Matlab language through which to solve the task. Because of their diversity, there is a risk of not using the most appropriate functions or the selected combination to be ineffective. In the second case, the automatic generated code has a poor optimization and it is not effective. During the analysis process of the logic of the code, it is necessary to be investigated several key points:

- are there branches in the computational procedure;
- what is the amount of variables;
- what is the size of the used variables and matrices.

Depending on the results obtained in this analysis it is possible to use the following techniques and technologies in Matlab.

B. Parallel Computing Toolbox™

Parallel Computing Toolbox provides several high-level programming constructs that let the user convert the applications to take advantage of computers equipped with multicore processors and GPUs. Constructs such as parallel for-loops (parfor) and special array types for distributed processing and for GPU computing simplify parallel code development by abstracting away the complexity of managing computations and data between the Matlab session and the computing resource that is used.

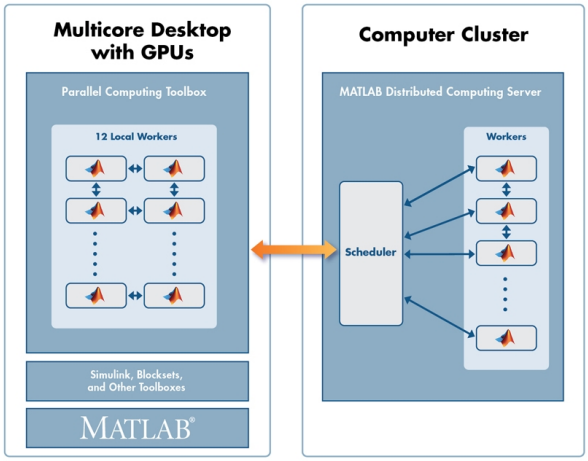


Fig1. Parallel computing with MATLAB [6].

The application can be run on a variety of computing resources without reprogramming it. The parallel constructs function in the same way, regardless of the resource on which the application runs—a multicore desktop (using the toolbox) or on a larger resource such as a computer cluster (using toolbox with Matlab Distributed Computing Server).

Key Features

- ✓ Parallel `for`-loops (`parfor`) for running task-parallel algorithms on multiple processors
- ✓ Support for CUDA-enabled NVIDIA GPUs
- ✓ Ability to run twelve workers locally on a multicore desktop
- ✓ Computer cluster and grid support (with MATLAB Distributed Computing Server)
- ✓ Interactive and batch execution of parallel applications
- ✓ Distributed arrays and single program multiple data (`spmd`) construct for large dataset handling and data-parallel algorithms

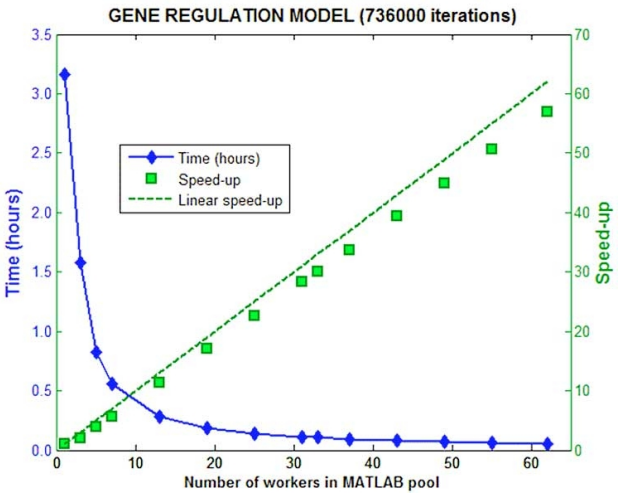


Fig.2 Application Speedup with "Parallel Computing Toolbox"[6].

a) Speeding Up MATLAB Computations with GPUs

Parallel Computing Toolbox provides GPUArray, a special array type with several associated functions that provides computations on CUDA-enabled NVIDIA GPUs directly from Matlab. Functions include "*fft*", element-wise operations, and several linear algebra operations such as "*lu*" and "*mldivide*", also known as the backslash operator (\). The toolbox also provides a mechanism for using the existing CUDA-based GPU kernels directly from Matlab.

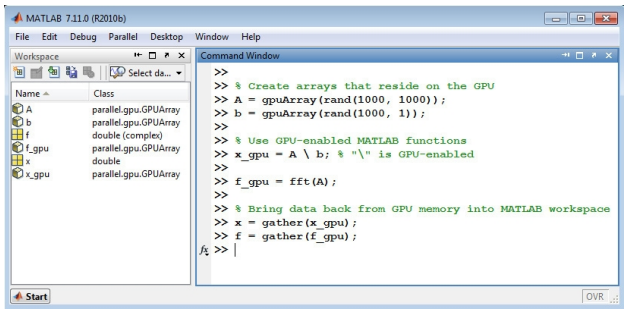


Fig.3 GPU computing with MATLAB. Using GPUArrays and GPU-enabled

b) Best Practices for MATLAB GPU Coding[5]

Profiling Code

The first criteria to determine whether the code is a good candidate for GPU processing is to identify which function is taking up the majority of the time. The Matlab profile tool will help to determine the characteristics of the code. Good indicators include:

- Custom functions that consume the majority of time;
- Custom lines of code that consume the majority of time;
- Use of a parfor loop;
- Custom MEX functions that consume the majority of time;

Bad indicators include:

- Built-in functions that consume the majority of time and do not have a GPUArray corresponding function (e.g. svd);
- High function call rate that is not isolated to an individual function;
- Short execution time;

The next step is to utilize the Matlab debugger to insert breakpoints in the functions/lines of code that use the most processing time to get a sense of what types/size of data is being processed.
Good indicators include:

- Large input/output data (several thousand rows/columns/etc…)
- Simple data types
- Bad indicators include:
- Small data volume
- Complicated data types (i.e. large, multi-level structures)

Test of different types/sizes of inputs will be helpful to understand the range of possible bottlenecks.

Strategies for Using GPU Capability

- **Convert directly to gpuArray built-in function**

The easiest conversion is to replace an existing built-in function with a GPUArray built-in function which is easy to be tested. The existing input array must be turn into a gpuArray and running the code to be seen if the desired function works (i.e., is overloaded) for the gpuArray datatype. If errors appeared, data structures need to be updated to accommodate a gpuArray, or there might be places which must be converted back to a non-gpuArray to perform a built-in operation that does not support the gpuArray type. In some cases, this will provide an immediate speed-up. In other, it will make the code run slower. The fact that the code is running slower is not a bad sign. If the code works, then it is possible that restructuring the code may lead to better performance. Using the Matlab debugger with breakpoints is an excellent tool for accomplishing this task.

- **Find and fix new bottlenecks**

After optimizing code to use GPU functions, it is possible to start scaling up the size of input data which can lead to the discovery of new bottlenecks. It is best to iterate through the process by starting with a fresh profile to establish the new bottleneck and determine whether additional GPU functions can be used. Sometimes after optimizing the code, it becomes obvious that performed calculations are too small. Otherwise, it may not be possible to optimize any further, especially if a loop is used.

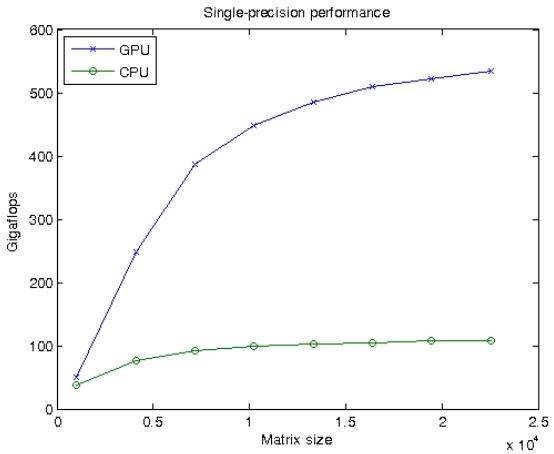


Fig.4 Single-precision performance

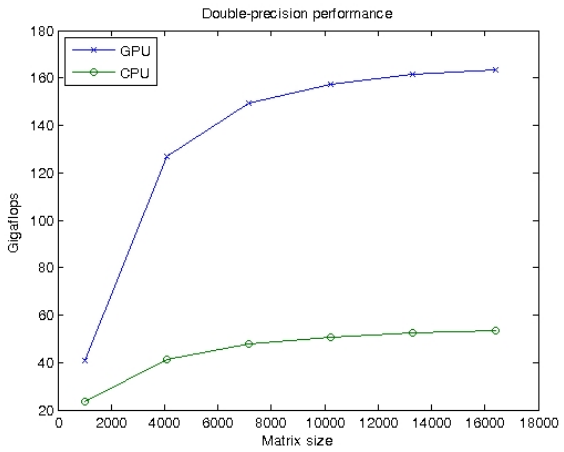


Fig.5 Double-precision performance

C. *Preallocating Arrays[6]*

"For" and "while" loops that incrementally increase the size of a data structure each time through the loop can adversely affect performance and memory use. Repeatedly resizing arrays often requires Matlab to spend extra time looking for larger contiguous blocks of memory and then moving the array into those blocks. Often, to improve the code execution time it must be preallocated to the maximum amount of space required for the array.

The following code displays the amount of time needed to create a scalar variable, x, and then to gradually increase the size of x in a "for" loop.

```
tic
x = 0;
for k = 2:1000000
  x(k) = x(k-1) + 5;
end
toc
Elapsed time is 0.301528 seconds.
```

If there is a preallocation of 1-by-1,000,000 block of memory for x and initialize it to zero, then the code runs much faster because there is no need to repeatedly reallocate memory for the growing data structure.

```
tic
x = zeros(1, 1000000);
for k = 2:1000000
x(k) = x(k-1) + 5;
end
toc
Elapsed time is 0.011938 seconds.
```

Using the appropriate preallocation function for the kind of array:

- zeros for numeric arrays
- cell for character arrays

D. *Preallocating a Nondouble Matrix[6]*

When the preallocation is to hold a matrix of some type other than double, it must be avoided using the method

```
A = int8(zeros(100));
```

This statement preallocates a 100-by-100 matrix of int8, first by creating a full matrix of double values, and then by converts each element to int8. Creating the array as int8 values saves time and memory. For example:

```
A = zeros(100, 'int8');
```

E. *Assigning Variables*

For best performance, the following suggestions would be useful:

– When storing data of a different type, it is advisable to create a new variable. Changing the class or array shape of an existing variable slows Matlab down because it takes extra time to process.

This code changes the type of X from double to char, which has a negative impact on performance:

```
X = 23;
-- some code --
X = 'A';   % X changed from type double to char
 -- some code --
```

– It is not recommended to assign a real value to a variable that already holds a complex value, and vice versa. Assigning a complex number to a variable that already holds a real number impacts the performance of your program.

F. *Using Appropriate Logical Operators[6]*

When performing a logical AND or OR operation, there is a choice of whether to use short-circuit operators.

Table I Logical operators

Operator	***Description***
&, |	Perform logical AND and OR on arrays element by element
&&, ||	Perform logical AND and OR on scalar values with short-circuiting

In "if" and "while" statements, it is more efficient to use the short-circuit operators, && and ||, because these operators often do not evaluate the entire logical expression. For example, whenever the first input argument is not numeric, Matlab evaluates only the first part of this expression.

```
if (isnumeric(varargin{1})) && (ischar(varargin{2}))
```

IV. CALCULATION PROCEDURE FOR SIMULATION OF THERMAL PROCESSES

To verify the possible results from the suggestion steps the following task is chosen - process simulation in chamber resistance furnace. The purpose of the calculation procedure is to research using software the technology mode of the complex: furnace-heated detail. The code is systems of ordinary differential equations (SODE)[3][6]. After applying the optimization steps, the second Matlab file was created. Both files were executed many times in equivalent environment and condition. The average times for calculations is shown in Table II and Fig.6.

Table II Time for computation

Case	*Time (in seconds)*
Average time for computation before optimization	157.62
Average time for computation after optimization	15.07

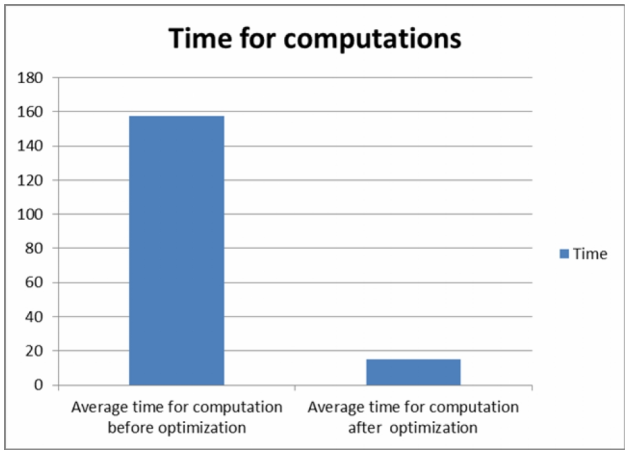


Fig.6 Time difference in computations

It is clear visible, that the optimizations steps make significant difference for the computation time. The difference before and after is 10,5 times.

V. CONLUSION

The software simulations of heat processes are very important and useful approach to understand their nature and specific. Thus obtain information about the process for which there is no prior experimental data. This helps to save electricity, as these facilities can be significant power. For example, the powerful industrial ovens can be in the range from 500 to 800 kW or more. On the other hand, the ongoing processes in these experiments are slow and can take tens to hundreds of hours. Optimization of the code which simulate them and decreasing the time for computations is very helpful. In this way will be possible to be made more experiments and more complex simulations with more precise mathematical models.

ACKNOWLEDGMENT

This paper is developed in the frames of project "Increasing energy efficiency and optimization of electro-technological processes and devices", № МУ03/163 financed by the National Science Fund

REFERENCES

[1] Agner Fog. Optimizing software in C++.An optimization guide for Windows, Linux and Mac platforms Copenhagen University College of Engineering. Copyright © 2004 - 2012. Last updated 2012-02-29.

[2] Мэтьюз Дж. Г., Куртис Д. Ф. Численные методы использование MATLAB. "Вильямс" Москва 2001 год

[3] Розин Л.А.,Методконечныхэлементов.Санкт-Петербургский государственый университет. 2000 год

[4] http://www.ansys.com/

[5] http://www.cac.cornell.edu/matlab

[6] http://www.mathworks.com/

Support for data-intensive computing with CloudMan

Y. Kowsar[1] andE. Afgan[1,2]
[1]Victorian Life Sciences Computation Initiative (VLSCI), University of Melbourne, Melbourne, Australia
[2]Centre for Informatics and Computing (CIR), Ruđer Bošković Institute (RBI), Zagreb, Croatia
ykowsar@student.unimelb.edu.au, enis.afgan@[unimelb.edu.au, irb.hr]

Abstract - Infrastructure-as-a-Service (IaaS) compute infrastructure model has showcased its ability to transform how access to compute resources is realized; it delivered on the notion of Infrastructure-as-Code and enabled a new wave of compute adaptability. In previous work we have been developing CloudMan (usecloudman.org) as a versatile solution for enabling and managing compute clusters in cloud environments via a simple web interface or an API. However, CloudMan only supported batch processing workloads. As the magnitude of the data produced and processed in digital form grows, the need to support big data applications in clusters in the cloud becomes more evident. In this paper, we have extended the batch processing capability of CloudMan and presenteda novel architecture for supporting big data analysis workloads in cluster-in-the-cloud environment.We also implemented the details through CloudMan using established big data platforms.

I. INTRODUCTION

As the information age continues, the rate at which data is produced is continuing its exponential growth; it is fueled by everything from news services and social media activities to sensor networks and research-driven devices. Recent advances in biology and the advent of Next Generation Sequencing (NGS) systems is a prime example of such developments. This data growth has led to what is often referred to as the data deluge and has posed a shift in the research problems at hand[1]. Several years ago, the rate at which, for examplebiological, data was produced was a limiting research factor. Namely, biologists would generate genomic sequences one nucleotide at a time and this allowed the data to be easily stored and analyzed even on a researcher's personal computer.

With the availability of NGS systems those days have been surpassed. Today, it is possible to sequence an entire human genome, consisting of ~3 billion nucleotides, in just a few days, at better quality, and at a fraction of cost compared to just a few years ago. This ability to generate so much data has, however, led to major challenges when dealing with storage and analytics from the data.

Although often primarily described as a challenge and an obstacle, the reality is that the availability of the increasing data volume presents enormous opportunities. The real power of the data will not come just from the sheer volume, but from the ability to analyze it. It is thus vital to provide flexible yet accessible solutions that enable researchers and companies alike to move beyond the data collection and step into the world of data analytics.

Although IaaS is an established approach for providing computing facilities on the cloud, many workloads still execute only in a more structured and traditional cluster computing environment where jobs are handed off to a job manager and possibly executed in parallel.As a step in this direction, we have been developing a cloud resource manager called CloudMan [2]that facilitates creation of a compute platform [3] providing clusters in the cloud environment. CloudMan iscapable of handling a range of workloads, including biological analyses (see [4]). Cloud computing in general allows compute and storage resources to be requested, provisioned, and utilized to handle the necessary scaling of a computational problem at hand. However, those resources are often provisioned as bare virtual machines and disks without application context or coordination. CloudMan helps in this regard by orchestrating all the steps required to provision a functional compute and application environment by establishing a traditional batch queue job manager atop flexible cloud resources.This setup provides a user with an accessible and functional data analysis platform that is rooted in familiar cluster-based concepts but leverages cloud resources instead.

In this paper, we describe recent advances in CloudMan where support for data intensive workloads has been added. Until now, CloudMan provisioned a functional and scalable batch queue processing system through Sun Grid Engine (SGE) compute cluster in the cloud as well as a range of bioinformatics tools as part of CloudBioLinux [5]and Galaxy[6], [7], [8].Extending the CloudMan capabilities,we have now introduced the big-data-platform-on-demand concept by presenting a new architecture through adopting and utilizinga well established big data platform component;Hadoop [9]. In addition we have addedsupport for federated computing concept into CloudManusingHTCondor [10]. These advances empower the next wave of data analysis workloads to be seamlessly executed in the cloud environment and easily integrated with the already existing and more traditional pipelines. Support for these types of workloads also facilitates easier development of Hadoop-based tools (due to the accessibility of the cloud and the functional execution/development environment) and provision groundwork for federation of clusters across multiple clouds and/or data centers. It is worth explicitly stating that CloudMan is not limited to biological

workloads and can just as readily be utilized for any other workload where a compute cluster is necessary.

This paper is organized as follows. In chapter II we provide an overview on the two main components namely, Hadoop and HTCodor, used in the presented architecture and provide details on how CloudMan benefits from this architecture with a brief overview of other existing solutions. In chapter III we present our architecture and its implementation details. In chapter IV test case studies and usage have been provided. Finally, in chapter V the conclusion and future direction have been stated.

II. BACKGROUND

A. Big Data: Hadoop

Big data is a concept introduced as a result of recent explosion of available data in various fields. This also applies to sciences that were traditionally not considered data-intensive and is having a ripple effect on how research is being transformed. One possible approach for tackling big data problems comes in the form of distributed and cloud computing where use of commodity computers enables feasible and accessible solution for a variety of problem domains. These approaches are providing opportunities for researchers and companies alike to store, retrieve and interpret their data withina reasonable amount of time and budget. Hadoop is an example of such a solution; it is a framework for analyzing big data across clusters of computers using theMapReduce programming model. In this model a problem is divided into several independent smaller problemswhere each sub-problem can be solved over a single node of a cluster (*the map phase*). The process of solving the problem continues by collecting and merging the results from each sub-problem (*thereduce phase*) to shape a single output. Programs written in this style can be automatically parallelized and executed on a large cluster of commodity machines[9]. Combining this computational model with the readily available computing infrastructure and making it all accessible will go a long way toward tackling the big data problem.

B. Federated Computing: HTCondor

Complementing the notion of distributed computing is the notion of federated computing. In this model, the aim is to assemble distributed computing resources that are possibly geographically and administratively disparate into a cooperating unit. Joining of such resources makes it possible to achieve higher workload throughput and resource utilization; the aim is to utilize all the known and available resources available over a period of time without assuming high availability of those resources [15]. Joining a shared pool of resources brings other complexities such as enforcing each organization's policy and maintaining service layer agreements. In these scenarios, flexibility is a key for managing complex system.To achieve this goal, HTCondor has introduced twoconcepts: flocking and gliding[16], [17]. In flocking scenarios, the overflow workload will be sent to other, off-site HTCondor resources to keep the runtime promises. Flocking was first introduced by the HTCondor team as a solution to submit the work overflow to other HTCondor remote resources. In this scenario, remote resources managed by a HTCondor component can be acquired when necessary, given they are not currently being utilized. Although flocking can fulfill off-site resource rental, it will not apply when remote resources are not managed by a HTCondor component. To solve this problem, gliding was introduced. The gliding scenario is a dynamic, on-demand deployment of HTCondor's necessary daemons over remote resources that are otherwise not controlled by a HTCondor component. In this solution, use of resources managed by other job manager (such as SGE or PBS) can be achieved. This is accomplished through a Globus Toolkit [18] connection string, which is shared between remote resources. Although traditionally utilized across a dedicated set of resources, it is possible to envision use of federated computing across multiple clouds and/or clouds and dedicated resources. Enabling application execution environments and consequently applications to run in such environments would help minimize vendor lock-in, increase code portability, and create an equilibrium market.

This work was, in part, supported by the Genomics Virtual Laboratory (GVL) grant from the National eResearch Collaboration Tools and Resources (NeCTAR).

C. Beyond the state of the art

Several projects exist in the field of data-intensive and/or federated computing atop cloud resources; the ones most relevant to the work being presented in this paper are presented in Table 1.

TABLE I. LIST AND OVERVIEW OF PROJECTS RELATED TO THE WORK PRESENTED HERE

StarCluster [11]	StarCluster is an open source project created by STAR group at MIT that allows anyone to easily create and manage their own cluster computing environments hosted on Amazon's Elastic Compute Cloud (EC2) without needing to be a cloud expert. The project also provides support for Hadoop based workloads.
Amazon Elastic MapReduce [12]	Amazon's Elastic MapReduce (EMR) is a general-purpose web service that utilizes a hosted Hadoop framework running on Amazon EC2 and Amazon Simple Storage Service (S3) allowing anyone to process data-intensive workloads on Amazon Web Services (AWS) cloud.
ConPaaS [13]	This projects provides an open source system in which resources that belong to different operators are integrated into a single homogeneous federated cloud that users can access seamlessly. ConPaaS is a PaaS layer providing a set of elastic high-level services and runtime environments, including support for MapReduce jobs, Java and PHP applications, as well as as TaskFarming service. ConPasS is compatible with AWS and OpenNebula clouds.
HTCondor [10]	HTCondor is a workload management system for compute-intensive jobs. It provides a job management mechanism that automatically chooses when and where to run the jobs. In addition to being a *traditional* job manager, HTCondor scavenges CPU resources from potentially federated computers and makes those available via a single job queue. Standalone, HTCondor can also submit jobs to Amazon's EC2 resources.
Crossbow [14]	Crossbow is a domain-specific tool that focuses on whole genome resequencing analysis. It combines Bowtie and SoapSNP with Hadoop in a fixed data analysis software pipeline to enable fast data analysis on AWS or a local workstation.

CloudMan brings forward parts of each of the projects listed in Table 1 and yet delivers a solution that caters to users not supported by any of the existing projects. At its core, CloudMan is an accessible cloud manager that establishes functional application execution environments. It operates on a range of cloud providers (namely, AWS, OpenStack, OpenNebula, and Eucalyptus), thus making it the most versatile cloud manager available. Built with CloudBioLinux and Galaxy, it provides ready access to over 100 bioinformatics tools and hundreds of gigabytes of biological reference datasets that can be utilized via Galaxy's web interface, Linux remote desktop, or the command line interface. Nonetheless, CloudMan is not limited to biological workloads and can also be utilized for any workload where a compute cluster is necessary. CloudMan can be installed (via automation; see CloudBioLinux) on any Ubuntu machine image, thus instantly turning that machine image into a functional cloud cluster. Alternatively, prebuilt CloudMan images can be started without any installation via a single web-form (http://biocloudcentral.org). Once launched, CloudMan provides a web interface for managing cloud and cluster resources. Our new architecture extends the already extensive set of provided features and provides support for the new wave of applications and use cases of big data problems.

III. IMPLEMENTATION

The CloudMan platform is currently available for immediate instantiation on the AWS public cloud and is also compatible with private or academic clouds based on OpenStack, OpenNebula, or Eucalyptus middleware (visit biocloudcentral.org). Deploying CloudMan on a private cloud requires creation of a custom machine image, configuration of which is automated via CloudBioLinux configuration scripts (see http://github.com/chapmanb/cloudbiolinux). An example of such a deployment is available on Australia's national cloud, NeCTAR, which is based on OpenStack middleware.

CloudMan itself is written in Python as a standalone web application and is open source licensed under the MIT license. Internally, CloudMan is based on service-oriented architecture; each of the components within CloudMan (e.g., file systems, job managers, applications) are defined as self-contained services. Services implement a standard interface and control all actions related to the given service. The master control thread manages and coordinates all of the running services. Instructions on how to use all of the features of CloudMan are available at (http://usecloudman.org) while the source code is available at (http://bitbucket.org/galaxy/cloudman).

A. Service Implementation: Hadoop

Along with the cloud resource management, resource allocation and management in a distributed system is a task, which requires a lot of effort. Various systems have been implemented to fulfill this need. Among these systems, Sun Grid Engine (SGE) has gained reputation for being a stable, efficient and reliable system. On the other hand, having the necessary infrastructure to ease the development of solutions for big data problems by using the computing power from distributed systems is quite demanding and Hadoop has been introduced as a solution for this task. Integrating Hadoop with SGE (i.e., providing an available and ready to use distributed architecture to the end user), gives us the capability of having the grid engine manage distributed resources while having Hadoop as a platform service available on demand. This approach was pioneered by [19], where Hadoop systems was pre-installed over cluster's nodes while the Sun Grid Engine allocated the necessary resources to its users on the fly. Although this solution is satisfactory for systems where cluster's nodes are pre-configured and Hadoop is the most demanding sub-system used, running Hadoop daemons over a cluster can be seen as waste of resource as it might not be in use over the entire lifespan of given cluster node. Furthermore, the available solution has separated Hadoop jobs from other jobs in the system, which would decrease job throughput by reducing the system's coherency.

We reengineered the available solution to increase system throughput by having Hadoop sub-system be provided by the grid engine on-demand (vs. static). For this purpose, a grid engine integration script will inspect jobs submitted to SGE and whenever a Hadoop job is found, requested computing resources will be acquired from SGE and a Hadoop environment setup dynamically. In other words, before assigning a job to the resources, Hadoop will be installed on the acquired subset of the SGE cluster. The process of inspecting the job is realized in a constant amount of time, $O(1)$, and is handled by a job submission template script, which is made available to the end users. Each time a Hadoop job is detected, the script will read the allocated resources from SGE and dispatch the required binaries to each resource. This process happens in parallel across all the nodes so the time complexity of deploying Hadoop over the allocated nodes is minimum and reduced to the amount of time required to dispatch Hadoop onto a single system. After Hadoop installation is completed, the job is sent to the Hadoop master node to run. Once the job completes, grid engine integration script cleans all the resources by removing any unnecessary files dispatched to each node and killing all the daemons. This process is necessary as CloudMan is providing a hybrid platform to be used by different cluster tools and techniques. Therefore it is desirable to remove the Hadoop's files and daemons from every single node and get the system ready for any other non-Hadoop/Hadoopuses to keep the system layer agreement. On the other hand, as mentioned the process of deploying Hadoop gets only a constant amount of time, around 4 minutes (please see figure 2 for details) which in comparison with the actual job's running time can be ignored; as big data analysis are often computing intensive tasks which take a while to run. For a visual representation of this process please see figure 1.

B. Service implementation: HTCondor

Using private or community resources while having on-demand access to additional processing units from the cloud, to spread the workload when necessary or desired, is the promise that a hybrid IaaS solution offers. In the presented architect we have adopted flocking and gliding concepts from HTCondor (Please refer to section IIB for their definition) as a solution for providing a hybrid IaaS solution through CloudMan. In this approach we have targeted scenarios where multiple clusters with in a cloud

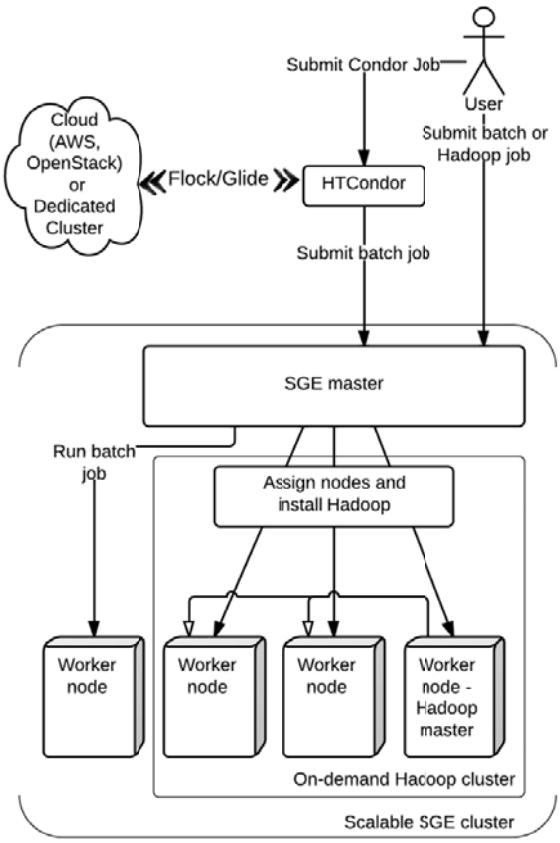


Figure 1: CloudMan architecture approach for supporting big data in a federated computing fashion.

are managed by an organization or in collaboration with other organizations and theses resources can be shared among users to maximize the throughput while minimizing the cost. The throughput can be optimized in the scenarios where one or multiple clusters are free whilstother clusters are heavily loaded. In these cases the jobs will be flocked (especially if the remote free resource is managed by CloudMan) or glidedinto the remote resource to have load balancing system across the resources. The other scenario where our hybrid solution is beneficial is when a huge pool of jobs has to be taken by a limited computing resource. In this case a remote resource can be acquired or rented by the organization for sending the job pool's overflow to the remote resource.Using this feature, CloudMan will keep the SLA promises with the minimum amount of overflow job costs.For this to be happenin the context of CloudMan, it is necessary for the remote organization resource to provide the connectionstring to the CloudMan's local HTCondor. Although access will be granted through the connection file, gliding scenario will not be functional until the resource provider grants enough privilege to the CloudMan user to run necessary daemons over the remote resource. After having been granted sufficient privileges to run the daemons, the local HTCondor can send the job to the remote resource by receiving the remote resource IP or DNS and the path to the Globus file connection string.

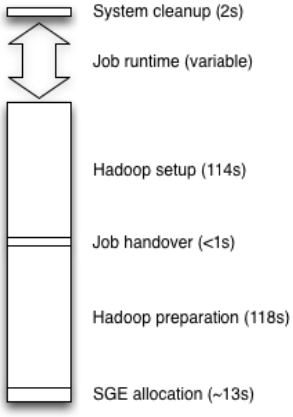


Figure 2. The Hadoop job preparation process through SGE and a sample timeline

Integrating these two concepts from HTCondor into our solution represents a first step in enabling CloudMan as a solution for hybrid IaaS. The actual integration with HTCondor is achieved by pre configuring the HTCondor settings in CloudMan and, whenever CloudMan starts, it sets up the HTCondor master node. As workers are added to the system, CloudMan automatically adds each worker node to the established condor pool, resulting in a single shared pool. To share the created resource pool in a flocking scenario, a negotiation between the two parties (resource provider and resource consumer) is required i.e. they should first come into a level of agreement on how, when and what resource can be used through the CloudMan. After reaching a level of agreement, the resource provider should grant sufficient privileges to CloudMan's HTCondor. This is achieved via CloudMan's DNS and HTCondor's configuration file. Lastly, the provider's DNS or IP should be passed into CloudMan, for which a form has been integrated into CloudMan's web interface. CloudMan will then set the necessary configuration, letting HTCondor negotiate and use remote cluster's resources as required (see Usage and Results section for an example).

IV. USAGE AND RESULTS

Submitting a Hadoop job requires setting two parameters in a template script file. The script is provided by CloudMan and specifies the number of CPU slots the job requires and a path to the job executable. From here submitting this job is equivalent to submitting any other SGE job(Please see figure 3 for a sample job details).

The challenge in this approach was to find a way to get both SGE and Hadoop to work without interfering with each other while keeping the resource utilization well-balanced. This goal is achieved by decoupling SGE from Hadoop where SGE prepares the necessary environment for Hadoop and then hands over the job to Hadoop. From this point on Hadoop handles the job as a self-contained system until the job has done. SGE is then informed of job completion and cleans the environment.

Figure 2 captures the process of creating the Hadoop environment within SGE and also shows sample times for the given process.

```
#! /bin/sh

echo "I'm process id $$ on" `hostname`
python random_lines_two_pass.py "dataset_$2.dat" "output_$2.dat" "3"
date
exit 42
```

(A)

```
executable=python_random_lines.sh
universe=vanilla
arguments=Example.$(Cluster).$(Process) [file_number]
transfer_input_files = dataset_[file_number].dat
transfer_output_files = output_[file_number].dat
output=results.output.$(Process).$(Cluster)
error=results.error.$(Process).$(Cluster)
log=results.log
notification=never
should_transfer_files=YES
when_to_transfer_output = ON_EXIT
queue
```

(B)

```
$ qsub -v HADOOP_HOME=/home/ubuntu/hadoop/hadoop/home\
,JAVA_HOME=/usr <hadoop file script>
```

(C)

Figure 3. HTCondor flocking example. A is the script to be run by HTCondor in this exmaple it is named python_random_lines.sh. B is the HTCondor script which should be submitted to HTCondor named remote_job.submit. Here we submit these scripts multiple times into AWS cluster (From Figure 3) using multiple different input files. It is worth noting that the Exit 42 at the end of python_random_lines.sh will let HTCondor detect when the job is finished. Then afterwards it can transfer the related files back to the local HTCondor.Part C shows how to sumit a Hadoop job into Sun Grid Engine.

Submitting HTCondor jobs in a CloudMan cluster requires a HTCondor job submission script to be created. Once composed, the script is submitted to HTCondor as shown in figure 4.

Although employing HTCondor as an engine for the federated cloud resource management gives us the capability of utilizing the disparate resources, the big data migration challenge remains an issue to be taken care of. Job files can be broadly grouped into two categories: the systematic ones, in which the system should take care of, and the user input data. In the context of CloudMan, a simple but effective solution for handling systematic data has been adopted; namely, any cluster managed by CloudMan will have the same copies of files in their local directories. This includes installed applications, application configuration files, as well as reference data, which is often required by a variety of bioinformatics tools and reaches multiple terabytes. Therefore, jobs requiring access to these data can be used without any data copying from one CloudMan cluster to another. Whilst this is an efficient solution for systematic files, the user input data is a bottleneck in job dispatching. This is because flocking a job into a remote cloud or cluster requires transferring the input data next to the job, which means carrying the required amount of data over the network. It is worth mentioning that although tackling this bottleneck is an interesting question going under research, in this paper, our focus is on introducing the necessary architecture and its implementation details to provide an accessible platform for enabling further research in this field. Currently, CloudMan is relying on HTCondor job scheduler's decision for submitting jobs into the remote resources. We are exploring alternative scheduling approaches to utilize data locality between clusters with the aim of submitting data intensive tasks local to the data while sending compute intensive tasks to the available remote resources.

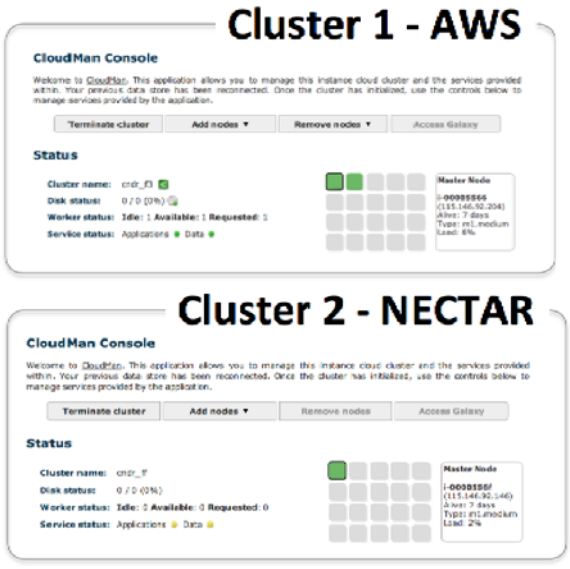


Figure 4. The CloudMan cluster setup for HTCondor flocking scenario. In this picture two CloudMan cluster has been created: one on the Amazon cloud (AWS) and one on the NeCTAR cloud (NeCTAR). The flocking happen from AWS to NeCTAR.

V. CONCLUSION

Infrastructure-as-a-service (IaaS) is the delivery of hardware and its associated software as a service[16]. CloudMan was first introduced to facilitate management of such a service with a focus on easing the process of preparing a basis for providing bioinformatics tools on the cloud. Successful projects, such as Galaxy, the Genomics Virtual Laboratory [20], CloudBioLinux, and SARA Cloud [21], are using or providing cloud services to the end users based on CloudMan.

In this paper we extended the capabilities of CloudMan to accommodate managing cloud resource in a more versatile and federated environment. This was achieved by integrating Hadoop platform into CloudMan and thus providing Hadoop-as-a-Service. With Hadoop's ability to process big data, availability of such a platform enables easier access to a new wave of (bioinformatics) tools for both, development and usage. In addition, we have integrated HTCondor into CloudMan. With HTCondor integrated, it is possible to easily span boundaries of a single compute cluster and extend the capacity of any administratively single resource, regardless of whether it is in the cloud or not. This enables overflowing excess workloads to a different set of computational resources (e.g., local cluster to a cloud or from one cloud provider to a different cloud provider). This also opens a new door towards federated job execution and data-locality based job execution. The architecture and the developed platform discussed here is a step towards further research in the space of big data movement and job scheduling to exploit data locality. This

is due to the fact that by using CloudMan it has become feasible to have a federated cluster computing over the cloud; something that otherwise requiresconsiderable effort on behalf of researchers or users.

Adding these features to CloudMan significantly extends the feature set of CloudMan and confirms its role as an accessible cloud management console and a provider of big-data-platform-on-demand.Based on our knowledge and supported by the relevant works described in Section II, this is the first solution that delivers an efficient approach for providing ready to use big-data-analysis-platform over a federated cluster on the cloud environment; our solution utilizes existing technologies to deliver a novel architecture and functionality to cloud environments, minimizing required investment and yielding an innovative approach.

ACKNOWLEDGMENT

The CloudMan is being developed by the Galaxy (http://wiki.galaxyproject.org/GalaxyTeam) and LSCC teams (http://www.vlsci.org.au/page/lscc).

Funding: This work was, in part, supported by the Genomics Virtual Laboratory (GVL) grant from the National eResearch Collaboration Tools and Resources (NeCTAR).

REFERENCES

[1] "The data deluge," *Nature Cell Biology,* vol. 14, August 2012.

[2] E. Afgan, D. Baker, N. Coraor, B. Chapman, A. Nekrutenko, and J. Taylor, "Galaxy CloudMan: delivering cloud compute clusters," *BMC Bioinformatics,* vol. 11, p. S4, 2010.

[3] E. Afgan, B. Chapman, and J. Taylor, "CloudMan as a platform for tool, data, and analysis distribution," *BMC Bioinformatics,* vol. 13, pp. 315-315, 2012.

[4] E. Afgan, D. Baker, N. Coraor, H. Goto, I. M. Paul, K. D. Makova, A. Nekrutenko, and J. Taylor, "Harnessing cloud computing with Galaxy Cloud," *Nature Biotechnology,* vol. 29, pp. 972-974, 2011.

[5] K. Krampis, T. Booth, B. Chapman, B. Tiwari, M. Bicak, D. Field, and K. E. Nelson, "Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community," *BMC Bioinformatics,* vol. 13, pp. 42-42, 2012.

[6] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biol,* vol. 11, p. R86, 2010.

[7] E. Afgan, J. Goecks, D. Baker, N. Coraor, T. G. Team, A. Nekrutenko, and J. Taylor, "Galaxy - a Gateway to Tools in e-Science," in *Guide to e-Science: Next Generation Scientific Research and Discovery*, K. Yang, Ed., ed: Springer 2011, pp. 145-177.

[8] A. Nekrutenko and J. Taylor, *Next-generation sequencing data interpretation: enhancing reproducibility and accessibility*.

[9] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Communications of the ACM,* vol. 51, pp. 107-113, 2008.

[10] D. Thain, T. Tannenbaum, and M. Livny, "Distributed computing in practice: the Condor experience," *CONCURRENCY AND COMPUTATION,* vol. 17, pp. 323-356, 2005.

[11] C. Ivica, J. T. Riley, and C. Shubert, "StarHPC — Teaching parallel programming within elastic compute cloud," *Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces,* p. 353, 2009.

[12] *Amazon Web Services*. Available: http://aws.amazon.com

[13] G. Pierre and C. Stratan, "ConPaaS: A Platform for Hosting Elastic Cloud Applications," *IEEE Internet Computing,* vol. 16, pp. 88-92, 2012.

[14] B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg, "Searching for SNPs with cloud computing," *Genome Biol,* vol. 10, p. R134, 2009.

[15] *High performance cluster computing. 1. Architectures and systems* vol. 1: Upper Saddle River, NJ [u.a.] : Prentice Hall, 1999, 1999.

[16] S. Bhardwaj, L. Jain, and S. Jain, "Cloud Computing: A Study of Infrastructure as a Service (IAAS)," *International Journal of Engineering and Information Technology,* vol. 2, p. 3, 2010.

[17] D. H. J. Epema, M. Livny, R. vanDantzig, X. Evers, and J. Pruyne, *A worldwide flock of Condors: Load sharing among workstation clusters*.

[18] I. Foster and C. Kesselman, "The Globus Toolkit," in *The Grid: Blueprint for a New Computing Infrastructure*, I. Foster and C. Kesselman, Eds., ed San Francisco, California: Morgan Kaufmann, 1999, pp. 259--278.

[19] S. Krishnan, "myHadoop: Hadoop-on-Demand on Traditional HPC Resources," ed. The UC Cloud Summit 2011: University of California, Los Angeles, 2011.

[20] *Genomics Virtual Lab (GVL)*. Available: https://genome.edu.au

[21] M. d. Hollander, "The Cloud for Biologists," ed. SARA HPC Cloud Day, Amsterdam, 2011.

IaaS type Cloud infrastructure assessment and monitoring

M. Kozlovszky*, M. Törőcsik*, T. Schubert*, V. Póserné*
* Óbuda University/John von Neumann Faculty of Informatics, Budapest, Hungary
kozlovszky.miklos@nik.uni-obuda.hu

Abstract - IaaS type clouds are multilayered complex systems. From end-user perspective, their service quality and reliability depends on heavily the underlying hardware/software infrastructure, the used technologies and resources (e.g.: human resources). To evaluate successfully an IaaS cloud ecosystem, we need to measure it side by side from many perspectives (e.g.: security, QoS, performance, reliability, etc.). Our aim is to build up an exhaustive parameter tree, which can describe the quantitative and qualitative parameter values of a generic IaaS cloud system. The values of these pre-defined parameters are provided by our software measurement framework, which is capable to evaluate the targeted IaaS cloud system automatically. The measured values of the parameter tree enable the potential tenants/end-users and governmental organizations to evaluate and compare different IaaS type cloud systems available on the market.

Keywords

IaaS cloud, infrastructure monitoring, accreditation

I. INTRODUCTION

Cloud infrastructure becomes the primary business environment for all types of enterprises during recent years. Cloud environments offer flexibility, scalability, adaptability and if everything goes well also resilience. Outsourcing mission critical assets of the business always and immediately pops up question: Is this solution secure / reliable / scalable / cost effective? These questions have to be answered right at starting of the transition planning phase, and usually they stay valid after the service and data migration as well.

The complexity of a generic IAAS type cloud infrastructure with hosted multi-tenant, virtualized business environments is definitely high (generic cloud example shown in Figure 1.).

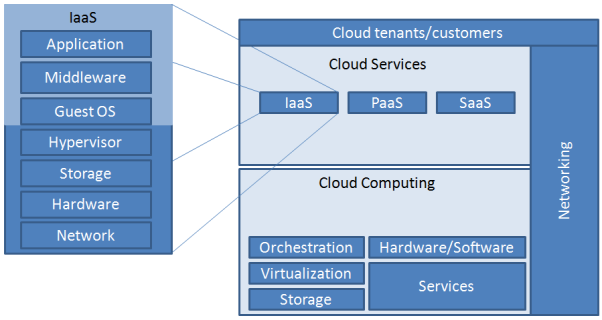


Figure 1. Generic IAAS type cloud infrastructure environment

This paper shows how we have built up our IAAS type cloud service and infrastructure evaluation framework. We describe the used parameter set, which helps us to evaluate (and indirectly compare) available cloud infrastructure instances.

A. Motivation

From literature we can identify four basic cloud deployment models: public, private, community and hybrid. In all these models there is a strong need to know precisely how the services and the underlying infrastructure are performing, how secure is the whole system. In three from the four models (public, community and hybrid cloud environments). A huge information and knowledge gap exists between the service provider and the customers. Key players of the cloud ecosystems should be able to track what is going on in the system at their level. Cloud acceptance is highly correlated directly how transparent and reliable cloud services are offered by the service providers. In multitenant clouds the customer should trust fully in the cloud service provider (how the provider protects its hardware infrastructure, how secure the premises are, how low the oversubscription factor is, how reliable the communication channels are, etc.). Moreover if the customer should be compliant with some business domain specific standards or regulations, the compliance should be validated for the cloud service provider's infrastructure/services.

B. IAAS midleware solutions

Nowadays a few vendor specific and open source middleware solutions are dominating the cloud market.

Commercial cloud middleware solutions are offered by a few big players (Vmware, Cisco, HP, IBM, Amazon) and a vast amount of smaller enterprises.
Major competitors in open source IaaS cloud branch are Eucalyptus[1], Nimbus/Cumulus[2], Open Nebula [3] and Open Stack[4].

C. Cloud monitoring and evaluation

1) Monitoring

Distributed Computing Infrastructures /DCIs/ (such as Clouds, Supercomputers or Grids) are monitored by various reasons and various ways. Widely used monitoring categories are the followings: status monitoring, performance monitoring and security monitoring: Monitoring can be realized both by active and passive - software/hardware - elements. Monitoring software solutions can be native / de facto standard applications (e.g.: SNMP based status monitoring, Linpack based performance monitoring, pattern based security monitoring, NVT/CVE based vulnerability scanning [5, 6, 7], patched/modified software solutions, or self developed monitoring ones. For full-scale cloud monitoring, administrators are keen to combine existing and in-house developed software solutions together.

2) Cloud evaluation

Cyber security is one of the major driving force behind cloud infrastructure and service evaluation. The U.S. government has launched the *Federal Risk and Authorization Management Program* aka. FedRAMP [8], which is basically a risk management program for large outsourced and multi-agency information systems. Its aim is enabling the entire process of assuring cloud instances are appropriate for individual governmental agency applications. Beside that FedRAMP also authorizes and continuously monitors IT services that are used by the federal departments and agencies. There are de facto audit standards of enterprise financial and infrastructure-related internal controls: the Statement on Auditing Standards No. 70 report usually referred to as SAS 70, the SSAE 16 and SOC reports. We need to mention here organizations and groups, which are actively pursuing cloud ecosystem evaluation mainly from security, and data control point of view: the Cloud Security Alliance (CSA) [9], European Network and Information Security Agency (ENISA) [10], the Cloud Computing Interoperability Group, and the Jericho Forum [11].

II. IAAS- TYPE CLOUD EVALUATION

IaaS type cloud infrastructure has a complex structure. To evaluate cloud providers with their IaaS type infrastructure one need to identify a large set of complex evaluation parameters (criteria). All the parameters should fulfill the following basic requirements:

- parameter/criteria should be objective
- parameters should be measurable
- parameter should be relevant from cloud user's and/or service provider's point of view

A. Evaluation categories

We have identified 19 main evaluation parameter groups (shown in Table 1.).

Table 1. Parameter categories

No.	Parameter group	Notes
1	Generic parameters	Generic Cloud Provider parameters (size, facilities, human resources,...)
2	Generic functionalities	Generic Cloud Provider services and functionalities
3	Performance (infrastructure)	Cloud Provider's infrastructure performance
4	Quality (infrastructure)	Cloud Provider's infrastructure quality (homogeneity)
5	Scalability	Cloud Provider's infrastructure scalability
6	Security	Infrastructure and service security as a whole
7	High availability/Disaster recovery	Cloud Provider's infrastructure stability, error proneness
8	Viability	Cloud Provider's business stability, survivability
9	QoS	Cloud Provider's service and infrastructure quality, and their assurance
10	Standards	Cloud Provider's (re)used standards in services and infrastructure
11	Monitoring	Cloud Provider's monitoring capabilities
12	Accounting/Logging	Cloud Provider's Accounting/Logging capabilities
13	Data maintenance	Cloud Provider's data maintenence capabilities
14	User friendlyness	Cloud Provider's service and infrastructure usability
15	Green/Efficiency	Cloud Provider's energy efficiency parameters
16	Transparency	Cloud Provider's business/service transparency
17	Legal and Regulatory Compliance	Cloud Provider's legal and regulatory compliances
18	Service and Infrastructure Assessment	Cloud Provider's existing service and infrastructure assessments
19	Knowledge base and competence	Cloud Provider's competence level in its services and infrastructure

These groups provide us an exhaustive parameter field for Cloud Service Provider assessment. The evaluation parameter set has a tree shape graph structure. Each parameter group contains parameter sub-group (and the sub-groups can contain parameter sub-sub groups as well with an indefinite depth). With this tree-like organization setup we can scale up the parameter field and include additional evaluation parameters easily without to influence other parts of the parameter tree. As a total of 450 parameter records have been identified so far inside the parameter tree.

B. Security evaluation parameter sub-groups

We have identified the following main evaluation parameter sub-groups within the security parameter category (shown in Table 2.).

Table 2. Security assessment parameter sub-groups

No.	Parameter sub-.groups	Notes
1	IT Human resources	IT personnel (size, etc.)
1.1	Dedicated security personnel	IT security personnel (certified, non-certified, etc.)
2	Incident management	Incident management parameters
2.1	IT security incident management	IT incident response team parameters
2.2	Business incident management	Business incident response team parameters
3	Risk management	Risk management parameters
3.1	Asset enumeration	Asset enumeration parameters
3.2	Risk enumeration	Risk enumeration parameters
3.3	Business impact evaluation	Business impact parameters
3.4	Reevaluation	Reevaluation parameters (periodical, duration, etc.)
4	Vulnerability management	Vulnerability management parameters
4.1	Infrastructure vulnerabilities	Infrastructure vulnerability parameters (stats, etc.)
4.2	Application Vulnerabilities	Application Vulnerabilities parameters (stats, etc.)
4.3	Service Vulnerabilities	Service Vulnerabilities parameters (stats, etc.)
4.4	Accessed vulnerability repositories	Vulnerability repository parameters
4.5	Automatic vulnerability alarming	Automatic vulnerability alarming parameters
4.6	Automatic vulnerability monitoring	Automatic vulnerability monitoring parameters
4.7	Automatic vulnerability elimination	Automatic vulnerability elimination parameters
5	Identity management	Identity management parameters
6	Patch management	Patch management parameters
7	Authentication	Authentication parameters
8	Authorisation	Authorisation parameters
9	Accounting	Accounting parameters
10	Infection management	Infection management parameters
11	Encryption	Encryption parameters
12	Access management	Access management parameters
12.1	Physical Access Controll	Pysical access controll and prevention parameters
12.2	Digital Access Controll	Digital access controll and prevention parameters
13	Isolation	Isolation parameters
14	Forensic capabilities	Forensic parameters
15	Change management	Change management parameters

III. ACCREDITATION

A. *Accreditation procedure*

For the accreditation procedure we have defined an easy-to-follow guided process. The workflow of the developed accreditation procedure is shown in Figure 3. The IaaS Cloud accreditation process is starting by the request of the Cloud Service Provider (owner of the infrastructure) or 3rd party (e.g.: governmental) organization. The main elements of the accreditation procedure are the following:

1. Negotiation phase: In this phase the preliminary accreditation service requirements (focus of the accreditation, special assessment targets), financial, and legal aspects have to be identified.

2. Data collection phase: In this phase data is collected focusing on the existing cloud infrastructure, existing services, and existing procedures of the Cloud Service Provider. Data collection is realized by the auditors (manual data collection) and by measurement applications, probes, tests (IT supported data collection). The evaluation parameter set is pre-defined.

3. Data processing phase: In this phase each of the collected data sets are filtered, processed and converted to a single value, which is stored as the evaluation parameter result.

4. Parameter evaluation phase: In this phase each parameter of the whole evaluation parameter set is processed one-by-one by pre-defined evaluation rules.

5. Accreditation phase: In his phase the evaluated parameter set is multiplied by the parameter impact matrix, and according to the predefined decision rule the cloud provider receives the official accreditation result.

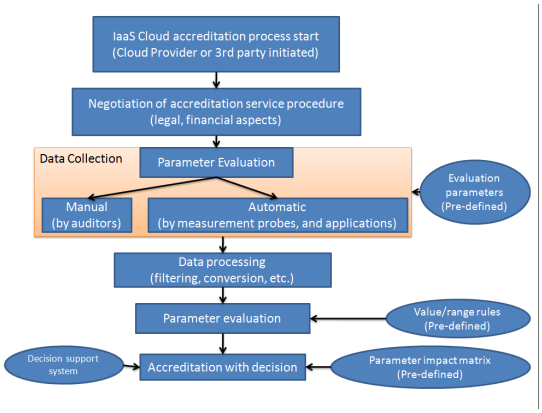


Figure 2. IaaS cloud accreditation procedure

IV. DATA COLLECTION

Definitions each of the parameter values are realized by manual or automatic data collection. Manual data collection is maintained by the auditor person during the precise, on-site checking process The IT supported data collection is realized by a multimodal measurement framework. The internal architecture of the Data Collection framework enables easy integration of various measurement tools (shown in Figure 2.). As an example in the followings we detail Vulnerability Management measurements. With our measurement framework we are able to measure various vulnerability parameters of a certain DCI. We are using a large set of security measurement applications in a plug-in like manner. We are combining some in-house developed applications (such as GSSVA [12] or AVAT [5]), with open source solutions, such as OpenVAS - Open Vulnerability Assessment System [13] which based on Nessus, Nexpose [14]], or commercial solutions as example Qualys's Vulnerability Management software [15,16].

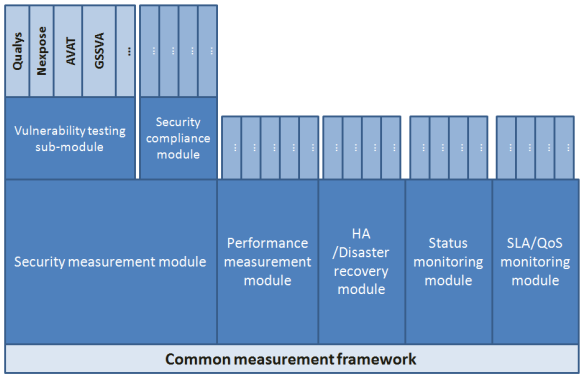


Figure 3. Schematic overview about the measurement framework used during the data collection phase

V. CONCLUSIONS

We have defined and built up an exhaustive tree-like graph of cloud evaluation parameter sets (our set contains more than 450 evaluation parameters). We showed the core parameter categories and as an example one of the sub-parameter sets (the security parameter group) to detail how the whole tree is organized. We have defined a generic core service assessment workflow and have identified its main tasks and additionally the relations between the tasks. We have used the defined parameter tree and the evaluation workflow in a cloud infrastructure related R&D project focusing on cloud providers. With the whole parameter tree we are able to realize IaaS type cloud service and infrastructure assessment. As an output we are able to use the defined processes and evaluation parameters to initiate a cloud accreditation service both for cloud service providers and cloud customers.

VI. ACKNOWLEDGMENTS

The authors would like to thank TÁMOP-4.2.1.B-11/2/KMR-2011-0001 – "Critical infrastructure defense research" for its financial support. This work makes use some of the results produced by the KMR_12-2012-0055-Cloud accreditations service. Both projects have been supported by the European Union.

REFERENCES

[1] http://www.eucalyptus.com , 2013 February

[2] Cumulus: An Open Source Storage Cloud for Science, Bresnahan, J., LaBissoniere, D., Freeman, T., Keahey, K. ScienceCloud 2011, San Jose, CA. June 2011.

[3] B. Sotomayor, R. S. Montero, I. M. Llorente, and I. Foster. Virtual Infrastructure Management in Private and Hybrid Clouds. IEEE Internet Computing, vol. 13, no. 5, pp. 14-22, Sep. 2009.

[4] http://www.openstack.org , 2013 February

[5] S. Acs, M. Kozlovszky, and P. Kotcauer; Advanced Vulnerability Assessment Tool for Distributed Systems, HP-SEE User Forum 2012

[6] Martin, Robert A.: Managing Vulnerabilities in Networked Systems. IEEE Computer Society COMPUTER Magazine, pp. 32-38, http://cve.mitre.org/, 2001.

[7] Mell, P., Scarfone, K., Romanosky S.: A complete guide to the common vulnerability scoring system, version 2.0. Forum of Incident Response and Security Teams, June 2007

[8] The Federal Risk and Authorization Management Program (FedRAMP); www.fedramp.gov ; February 2013

[9] Cloud Security Alliance – CSA; https://cloudsecurityalliance.org/ ;February 2013

[10] European Network and Information Security Agency – ENISA; http://www.enisa.europa.eu/ ; February 2013

[11] The Opengroup Jericho Forum; http://www.opengroup.org/getinvolved/forums/jericho ;February 2013

[12] Acs, S., Kozlovszky, M., and Balaton, Z.: Automation of security analysis for service grid systems: PARENG2009, The First International Conference on Parallel, Distributed and Grid Computing for Engineering, Pcs, Hungary,2009.

[13] The OpenVAS website, http://www.openvas.org , 2013 February

[14] http://awards.scmagazine.com/winners/2012/130 , 2013. November

[15] https://www.qualys.com , 2013February

[16] http://www.qualys.com/research/knowledge/, 2013. February

Energy Efficiency in Large-Scale Distributed Computing Systems

R. Trobec*, M. Depolli*, K. Skala** and T. Lipić**
*Jožef Stefan Institute/Department of Communication Systems, Ljubljana, Slovenia
**Ruđer Bošković Institute/Centre for Informatics and Computing, Zagreb, Croatia
{roman.trobec, matjaz.depolli}@ijs.si
{skala, tomislav.lipic }@irb.hr

Abstract – The ever-increasing energy consumption in large-scale distributed computing systems such as clusters, grids and clouds raises social, technical, economical, and environmental concerns. Therefore, designing novel energy-efficient approaches, to reduce energy consumption, at all levels of distributed system architecture is of great importance for the whole society. However, the essential step towards the introduction of energy-efficiency in large-scale distributed systems is to measure the power consumption, accurately, reliably, and continually in each component of the system. This paper briefly surveys the current approaches for measuring and profiling power consumption in large scale distributed systems. Furthermore, the practical case study of a real-time power measurement in multi-core computing system, as a basic building block of a distributed computing system, is presented.

I. INTRODUCTION

Large-scale distributed computing systems are composed of a large number of computing and storage resources connected through a network. Large-scale distributing computing systems include clusters, grids and clouds. Cluster computing systems can be classified as high availability clusters and load balancing clusters. In high availability clusters a small number of redundant nodes is used in order to eliminate single points of failures, while in load balancing clusters a larger number of nodes is used with focus on providing an environment for High Performance Computing (HPC). Cluster computing system is embedded both in grid computing systems and cloud computing systems [1]. However, compared to cluster computing systems, grid computing systems are more loosely coupled, heterogeneous and geographically dispersed. Cloud computing systems, using the virtualization technology, can provide scalable, reliable and cost-efficient virtual clusters [2][3]. The analysis of their applicability for HPC applications is one of interesting novel research topics [4][5][6].

As such large-scale distributed computing systems grow in size, adding more and more computing nodes and storage resources, their energy consumption is exponentially increasing [7]. In computer systems, energy (in joule), is the electricity resource that can power the hardware components to do computation and the dissipate rate of energy is power (in watts, joules per second). The ever-increasing energy consumption of large-scale distributed computing systems causes higher operation costs (e.g. electricity bills) and has negative environmental impacts (e.g. carbon dioxide emissions). Therefore, while designing such systems the focus is shifted from performance improvements to the energy efficiency.

Energy can be reduced at different levels of distributed architecture: individually on each node or hardware component-level, at the middleware level, at the networking level and at the application level [8][9][10]. A processor is component that consumes dominant amount of energy on each node. Thus, most of techniques are designed to reduce energy consumption of the processor [11][12], but also energy consumption reduction techniques have been proposed for other devices such as the phase-change memory and solid-state disk driver. Suboptimal use of systems resources (e.g. over-provisioning) is the primary source of energy inefficiency at middleware layer [8][9][13]. In over-provisioned system components usually consume excessive power when they are relatively idle or underutilized. Energy-proportional computing [14] introduces the concept of consuming energy proportional to resource utilization so that an idle or underutilized component consumes less energy. For improving energy efficiency at middleware layer, energy-aware resource and workload management algorithms are utilized [13]. Increased network traffic in many bandwidth sensitive network applications results also in increased number of networking devices such as switches and routers and therefore in increased energy consumption. Thus different approaches for new energy aware network protocols and network infrastructure are developed for improving energy efficiency at network layer [15][16]. The energy consumption heavily depends on the characteristic of application running. Some applications are computationally intensive, other are data intensive, while other are hybrid of both. Thus energy-aware programing models, regarding various types of workloads and architectures, are needed in order to develop energy efficient applications [17][18].

All this shows that energy can be saved at all levels of large-scale distributed system architecture. However, the essential step towards the introduction of energy-efficiency in large-scale distributed systems is accurate, reliable and continuous energy consumption measurement

of each component in the system. In this paper, a brief survey of the current approaches for measuring and profiling power consumption in large scale distributed systems is outlined. Furthermore, the case study of a real-time power measurement in a cluster computing system is presented to demonstrate a practical implementation of described approaches.

II. ENERGY MEASUREMENT AND PROFILING APPROACHES

Main energy measuring and profiling approaches can be classified as hardware based approaches, software based approaches and hybrid of both. In following subsections the basic characteristics of hardware-based and software-based approaches are briefly outlined based on two recent extensive survey papers [19][20].

A. Hardware-based Energy Measurement

Hardware-based energy measurement approaches are based on different instruments that measure the current voltage or the energy of different computation, communication or storage segments such as CPU racks of data centers, servers or motherboards to compute the energy consumption of the measured segment. These instruments can be implemented as meters, special hardware devices usually embedded in hardware platforms or as power sensors put on the hardware.

An energy measurement with meters is a direct and a straight-forward approach. A digital multimeter is one example of a meter that is commonly used for such an energy measurement. Digital multimeters can collect the voltage or the energy samples of the measured segment in pre-specified time intervals and send the measurements through a data link connected to the data collection system. Another meter used for the energy measurement is a clamp meter that, compared to a digital multimeter, has a larger measurement range and can be used to measure the energy of segments with much higher currents. The previously mentioned meters are mainly used to measure the DC power since they are connected between the power supply and the measured segment. On the other hand, there are other meters (e.g. "Watts UP" [21]) that can be also used to measure the AC power. However such meters only measure the system level power, because only power supplies are powered by AC.

In their simplicity, meters cannot be used when a higher control of the measurement process is required. In those cases more complex devices must be used - the specially designed ones. In 2000 Viredaz et al. designed a platform called Itsy to measure the energy consumption of mobile devices [22]. Another example of a complex device used to measure the energy is a single board on computer called PLEB [23]. PLEB is design with a set of current sensors on-board while the micro-controller of this device is integrated with an analogue-to-digital converter to read the sensors. Unlike Itsy, which measures only the energy consumption of mobile devices, PLEB can be used to measure the energy of processors, memories, flash drivers and I/O devices.

Finally, approach of sensors integration into hardware is mainly used by high-performance servers. Sensor can be either in-built or external. On-chip energy or temperature sensors provide high accurate solutions for monitoring energy although currently these solutions are still too expensive. Intel has provided Running Average Power Limit (RAPL) energy sensors in recent architectures such as SandyBridge for obtaining the energy consumption of CPU [24]. On the other hand, external sensors that output energy, temperature, humidity or water hazards of servers, are less expensive and also less accurate. For instance, in [25] authors used thermal camera as external device for the remote temperature monitoring of a computer cluster at the blade enclosure-level.

B. Software-based Energy Measurement

The disadvantages of hardware-based energy measurement techniques are seen in the fact that they depend on expensive hardware sensors and considerable knowledge of hardware design. Considering the need of expensive and cumbersome setup, they are also not well suited for online monitoring and energy-aware algorithms (e.g. energy-aware scheduling algorithms). In contrary, software based approaches can be used to supply more fine-grained online energy information, with reduced accuracy, useful for augmenting energy-aware algorithms. Furthermore, compared to hardware-based approaches, software-based approaches are more flexible because they can be applied to different platforms without changing the hardware.

A software-based approach usually builds power/ energy models to estimate the energy consumption at instruction level, program block level, process level, hardware component level or full system level. These models are built based on selection and optimization of power/energy indicators/descriptors of a software or a hardware component of interests. Regarding of power/energy indicators type, software based approaches can be classified as system profile-based approaches or hardware performance counter (PMC) based approaches.

System profile or system events are a set of performance statistical information supplied by the operating system or special software. These events can describe the current state of hardware, software and operating system. Joulemeter [26] tracks the energy consumption of a virtual machine using power/energy models for three system hardware components (i.e. CPU, memory and disk). The CPU energy model uses CPU utilization, the memory energy model uses the number of last level cache misses, while the disk energy model uses the bytes of data written to and read from disk.

Hardware performance counters are a group of special registers that store the counts of hardware-related activities within computer systems related to system hardware components. In [27] PMC-based modeling approaches are separated in two groups: top-down approaches and bottom-up approaches. The top-down approaches do not depend on modeled architecture thus enabling a fast and an easy deployment [28]. On the other hand, the bottom-up approaches depend on underlying architecture and produce more informative, responsive and accurate power/energy models than the top-down

approaches. They enable a breakdown of the energy consumption per components of the architecture but are more complex to deploy than the top-down approaches [29].

III. Use Case: Real Time Power Measurement Of Cluster Computing System

The use case demonstrates a real time power monitoring of a muti-core computer, which is a main building block of computing clusters.

A. Benchmark program

We use a real benchmark application written in C++ that implements the parallel maximum clique algorithm [30], which is essentially an advanced recursive program for searching a tree. The algorithm takes an input graph and returns its maximum fully connected sub-graph. The parallelization of the algorithm employs the multi-threading techniques, which are supported in most of the programming languages without extra libraries or other kind of software support. This makes the algorithm portable to other languages and operating systems. It can run on most modern multi-core computer architectures.

All treads are asynchronously searching their branches in parallel using the global variable about the size of the current clique. The algorithm produces a significant speedup, which is near-linear for some instances of input graphs. We can trade with the number of cores and various system clock frequencies to achieve the minimal energy consumption.

B. Experimental Computing System

We evaluated the performance of the parallel maximum clique algorithm in terms of execution times and corresponding power consumption. All results are shown as the average from 15 experiments. The experimental computing system was built of a dual CPU 2.30 GHz Intel Xeon E5-2630, each of CPUs with six physical cores and runs under server version of Ubuntu 12.04.

C. Real Time Power Monitoring

Latest architectural designs of processors for desktop computers and workstation provide Performance Monitoring Units (PMUs), a hardware support for counting micro-architectural events of a processor core that are otherwise not explicitly visible to the outside world. In addition, latest versions of PMUs can collect data form the parts of the processor that are not parts of the core, e.g., a memory controller, and can even measure the power consumption of the processor and the main dynamic memory.

The CPU performance measurements of our use case are accessed with the Intel Performance Counter Monitor (PCM) [31], which is a low level library that enables access to the PMU of the Intel processors with Nehalem micro-architecture or newer. The PMU statistics can be measured at any point of the application program and/or during the whole execution. Although some statistics, such as cache hit rates, can be measured on the level of individual core, the power consumption is limited to individual processors and to the whole main memory. Detailed measurements of the power consumption are thus possible on application level, provided the analyzed application does not share processors with other applications.

PCM library requires installation by the super-user and also requires super-user privileges for measuring power consumption. Besides the option of linking the library with the target application, a command-line tool "*pcm-power.x*" is provided, which enables the power consumption measurements of custom applications. We use this tool for all of our measurements.

D. Experimental Results

We ran the parallel maximum clique algorithm on a different number of cores and measured the execution time and the power consumption. The results are shown in Figure 1 and Figure 2, respectively. As expected, the execution time in Figure 1 drops approximately proportional to the number of cores and also proportional to the system clock frequency. The power consumption on Figure 2 rises with the number of used cores and with the system clock frequency.

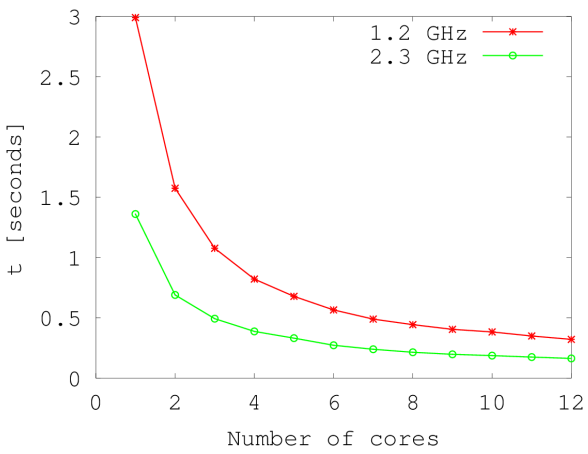


Figure 1. Execution time as a function of the number of cores

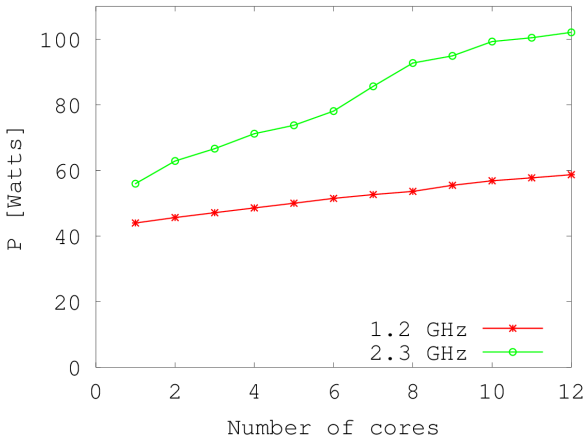


Figure 2. Power consumption as a function of the number of cores.

From Figure 2 we also see that the power consumption of an unloaded system is significant, approximately 75 % and 50 % of the power consumption of a fully loaded system under 1.2 GHz and 2.3 GHz, respectively.

The idle power consumption is partially due to the constant refreshing of the main dynamic memory (DRAM). The PCM measurements show that the idle consumption of DRAM is 2.1 W and is the same for 1.2 GHz, and 2.3 GHz, while it reaches 7.5 to 10 W when used by at least one core.

With no active core the power consumption of the experimental computing system is 7 W, but it increases to 40 W, if at least a single core is active. We assume that this is partially because of the processor logic that is always active (e.g. memory controller and cache), and partially due to the fact that computing cores are not shut down when running in the idle state but rather perform some simple task, such as an endless loop. This prevents a more efficient optimization of the power consumption in real applications.

The energy consumed by our benchmark application equals the execution time multiplied by the power consumption and is plotted in Figure 3. We can see that it decreases with a similar rate as the execution time. Therefore, the most optimal solution, regarding the energy and the execution time, is on all available cores and with the highest possible frequency.

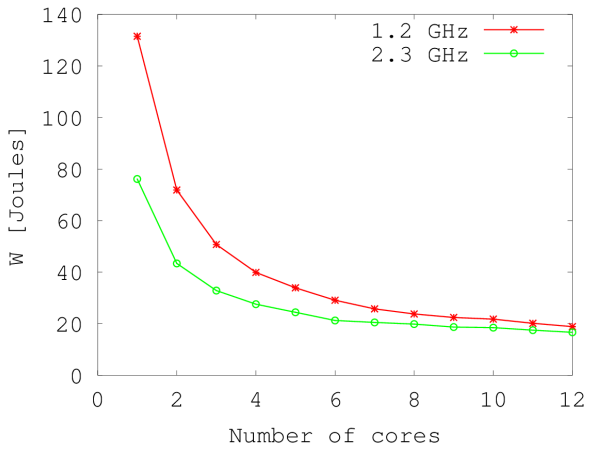


Figure 3. Energy consumption as a function of number of cores.

IV. Conclusion

This paper provides a brief overview on different approaches for measuring and profiling power consumption in large scale distributed systems. A simple practical use case study of a real-time power measurement in a multi-core computing system is presented. The energy consumption of a parallel custom algorithm is measured, by the PCM library, as a function of the number of cores and system clock frequency. The obtained measurements indicate that the most optimal solution, regarding the energy consumption, is to run the parallel algorithm on all available cores and with the highest possible frequency.

This unexpected result can be explained in the following way. The idle power consumption is significant, implying that the whole application should be executed in the shortest possible time. This can be achieved by all cores at the highest possible system clock frequency, because the speedup of the benchmark application is nearly ideal. In our use case, the contribution to a smaller energy consumption is much greater with a shorter execution time than with a smaller number of computing cores. Each additional core used by the application increases the power consumption by about 3% and decreases the execution time by about 50%. Therefore, the decrease in execution time provided by a core has much higher impact on the overall energy consumption as the power consumption of the fully loaded core.

However, if the efficiency of the tested parallel algorithm is not so high, in other words, if the speedup is not ideal, then the situation could change significantly. To optimize the energy consumption more aggressively, we should be able to shout down the computing cores, or at least put them in a sleep state with a significantly lower idle power consumption. To apply the optimization of the energy consumption on general applications, a profiling of their speedup is needed before the final run. These are the topics of our current research and future work.

Acknowledgment

The authors acknowledge the support of the bi-lateral scientific research project “Optimization of energy consumption in distributed computing systems”, founded by the Ministry of Education, Science, Culture and Sport of the Republic of Slovenia and the Ministry of Science, Education and Sports of the Republic of Croatia. The project is in the framework of the joint Slovenian-Croatian cooperation in science and technology between Jožef Stefan Institute (JSI) and Ruđer Bošković Institute (RBI).

References

[1] G. L. Valentini, W. Lassonde, S. U. Khan, N. Min-Allah, S. A. Madani, J. Li, L. Zhang, L. Wang, N. Ghani, J. Kolodziej, and et al., “An overview of energy efficiency techniques in cluster computing systems,” Cluster Computing, 2011, pp. 1-13.

[2] “CloudMan”, http://usecloudman.org/.

[3] “StarCluster.”, http://web.mit.edu/stardev/cluster/.

[4] C. Vecchiola, S. Pandey, and R. Buyya, “High-performance cloud computing: a view of scientific applications,” 10th International Symposium on Pervasive Systems, Algorithms, and Networks (ISPAN 2009), 14-16 Dec. 2009, Piscataway, NJ, USA, 2009, pp. 4–16.

[5] S. P. Ahuja, and S. Mani, "The State of High Performance Computing in the Cloud," Journal of Emerging Trends in Computing and Information Sciences, vol. 3, February 2012, pp. 262 – 266.

[6] P. Mehrotra, J. Djomehri, S. Heistand, R. Hood, H. Jin, A. Lazanoff, S. Saini, and R. Biswas, “Performance evaluation of Amazon EC2 for NASA HPC applications,” Proceedings of the 3rd workshop on Scientific Cloud Computing Date, New York, USA, 2012, pp. 41–50.

[7] A.-C. Orgerie, L. Lefevre, J.-P. Gelas, “Demystifying energy consumption in Grids and Clouds," International Green Computing Conference, 15-18 Aug. 2010, pp. 335-342

[8] A. Beloglazov, R. Buyya, Y. C. Lee, and A. Zomaya, “A taxonomy and survey of energy-efficient data centers and cloud computing systems,” Advances in Computers vol. 82, no. 2, 2011, pp. 47-111.

[9] C. Cai, L. Wang, S. U. Khan, and J. Tao. "Energy-Aware High Performance Computing: A Taxonomy Study," Conference on IEEE 17th International Parallel and Distributed Systems (ICPADS), 2011, pp. 953-958.

[10] J. Shuja, S. A. Madani, K. Bilal, K. Hayat, S. U. Khan, and S. Sarwar, "Energy-efficient data centers," Computing, vol. 94, no. 12, 2012, pp. 973-994.

[11] Y. Liu, and Z. Hong, "A survey of the research on power management techniques for high performance system,." Software: Practice and Experience vol. 40, no. 11 , 2010, pp. 943-964.

[12] N. B. Rizvandi, and A. Y. Zomaya, "A Primarily Survey on Energy Efficiency in Cloud and Distributed Computing Systems," arXiv preprint arXiv:1210.4690, 2012.

[13] A. Y. Zomaya, and C. L. Young, "Energy-efficient distributed computing," Awareness Magazine, January 2012.

[14] L. A. Barroso and U. Hölzle. "The Case for Energy-Proportional Computing," Computer , 2007, pp. 33–37.

[15] K. Bilal, Kashif, S. U. Khan, Sajjad A. Madani, K. Hayat, M. I. Khan, N. Min-Allah, J. Kolodziej, L. Wang, S. Zeadally, and D. Chen. "A survey on Green communications using Adaptive Link Rate," Cluster Computing, 2012, pp. 1-15.

[16] A. P.Bianzino, C. Chaudet, D. Rossi, and J-L. Rougier. "A survey of green networking research," Communications Surveys & Tutorials, IEEE vol. 14, no. 1, 2012, pp. 3-20.

[17] C. Zhang, K. Huang, X. Cui, Y. Chen, "Power-aware Programming with GPU Accelerators," IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 21-25 May 2012, pp. 2443-2449.

[18] C. Lively, X. Wu, V. Taylor, S. Moore, H.-C. Chang, C.-Yi Su, and K. Cameron, "Power-aware predictive models of hybrid (MPI/OpenMP) scientific applications on multicore systems," Computer Science-Research and Development, vol. 27, no. 4, 2012, pp. 245-253.

[19] S. Benedict, "Energy-aware performance analysis methodologies for HPC architectures—An exploratory study", Journal of Network and Computer Applications, vol. 35, issue 6, November 2012.

[20] H. Chen, and S. Weisong, "Power Measuring and Profiling: State-of-the-Art".

[21] Watts up, https://www.wattsupmeters.com.

[22] M. A. Viredaz, D. A. Wallach, and D. A. Wallach, "Power evaluation of itsy version 2.3", Technical report, Compaq, Western Research Laboratory, 2000.

[23] D. C. Snowdon, S. M. Petters, and G. Heiser, "Power measurement as the basis for power management," In Operat. System Platforms for embedded systemsIn, 2005.

[24] Intel In-built Sensors. Running average power limit for xeon processors, http://www.intel.com/xeon

[25] D. Kolarić, T. Lipić, I. Grubišić, L. Gjenero, K. Skala, "Application of Infrared Thermal Imaging in Blade System Temperature Monitoring," in Proceedings of the 53rd International Symposium on Electronics in Marine, ELMAR, Zadar, 2011, pp. 309-312

[26] A. Kansal, F. Zhao, J. Liu, N. Kothari, and A. A. Bhattacharya, "Virtual machine power metering and provisioning," In Proceedings of the 1st ACM symposium on Cloud computing, pp. 39-50, ACM, 2010.

[27] R. Bertran, M. Gonzàlez, X. Martorell., N. Navarro, and E. Ayguadé, "Counter-Based Power Modeling Methods: Top-Down vs. Bottom-Up," The Computer Journal, vol. 56 no. 2, pp. 198-213.

[28] W. L. Bircher, and L. K. John, "Complete system power estimation using processor performance events," IEEE Transactions on Computers, vol. 61, no. 4, 2012, pp. 563-577.

[29] R. Bertran, M. Gonzalez Tallada, X. Martorell, N. Navarro, E. Ayguade, "A Systematic Methodology to Generate Decomposable and Responsive Power Models for CMPs," IEEE Transactions on Computers", vol. PP, no. 99, pp. 1-1 4.

[30] J. Konc, D. Janezic, "An improved branch and bound algorithm for the maximum clique problem," MATCH Commun. Math. Comput. Chem, 2007, no. 58, pp. 569-590.

[31] R. Dementiev, T. Willhalm, O. Bruggeman, P. Fay, P. Ungerer, A. Ott, P. Lu, J. Harris, and P. Kerly, "Intel R performance counter monitor – a better way to measure cpu utilization," 2012, http://software.intel.com/en-us/articles/intel-performance-counter-monitor.

3D Visual Analytics for Quality Control in Engineering

M. Klemenčić[1] and K. Skala[2]

[1] University of Zagreb, Faculty of Graphic Arts, (PhD study) Getaldićeva 2, 10000 Zagreb, Croatia
[2] Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia

Abstract

This article will focus on 3D visual analytics for quality control in engineering. The **world today is a busy place with a rapidly increasing amount of data to be dealt with every day. Problems are encountered when most of the data are stored without filtering and refinement for later use. Industry has raised the demands for the high technological performance of final products, such as short production time, low manufacturing costs and overall product quality. Digitizing in the real-world has various application domains, and is of vital importance when it comes to methods involving industrial quality assurance. The process of the rapid development of products depends on new technologies, such as 3D scanning, 3D printing and prototyping. Detailed digitizing can provide more product information, making it easier to locate the causes of inaccuracies and optimize production. These possibilities help us check and improve tools and gadgets, control the form of prototypes and test series in production optimization, quality assurance of serial production etc. Quality and rapid digitizing using these systems makes copying easy and thus accelerates serial production, which is why these procedures are used by numerous companies. This paper explores and analyzes existing technology and devices used in industrial quality assurance, and provides a brief review of current options and application opportunities.**

I. Introduction

Every branch of industry or business and basically every personal activity nowadays generate vast amounts of data. In most cases, raw data has no value in itself. Therefore, information needs to be extracted to prevent it from becoming lost among data that may be:

- irrelevant to the current task at hand,
- processed in an inappropriate way or
- presented in an inappropriate way.

Due to this data overload, time and money are wasted. Scientific and industrial opportunities are lost because we still lack the ability to deal with the enormous data volume properly. Reduction of development times, optimization of production processes and higher product quality have become mandatory for the automotive, aerospace and consumer goods industries. Methods and models need to be identified in order to turn data into reliable and provable knowledge. Fully-automated search, filter and analysis only work reliably for well-defined and well-understood problems. The vision of visual analytics is to turn this information overload into an opportunity. Information visualization has changed our view of databases. The goal of visual analytics is to make the processing of data and information transparent for analytic discourse. Thus, visual analytics will foster the constructive evaluation and ultimately the improvement of our knowledge and decisions. Therefore, there are many software and devices on the market that provide important product information, inspection and quality control during production development, i.e., control of the overall process, including design, manufacture and maintenance.

II. Definition of Visual Analytics

Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces. People use visual analytics tools and techniques to synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data; detect the expected and discover the unexpected; provide timely, defensible, and understandable assessments; and communicate assessment effectively for action[1].

Its goal is to provide people with tools and techniques to enable them to:

- synthesize information,
- detect the expected and discover the unexpected
- provide understandable assessments and
- communicate assessments for action effectively

Visual analytics is more than just visualization. Its challenge is to identify the best automated algorithm for the analysis of tasks at hand, identify its limits, which can - not be further automated, and then develop a tightly integrated solution with the appropriate visualization and interaction techniques.

III. Scientific Visualization

In scientific visualization, the data entities to be visualized are usually 3D geometries, and can be understood as scalar, vectorial or tensorial fields with explicit references to time and space. Often, these 3D fields are visualized by isosurfaces or semitransparent point clouds. While current research has focused mainly on efficiency on the efficiency of the visualization techniques to enable interactive exploration, more and more methods for the automatic derivation of relevant visualization parameters have been the focus of research.

IV. Optical 3D Measurement Methods

As previously mentioned, the latest technological development is the optical 3D measuring technique [2, 3]. It is being used for 3D coordinate measuring, quality control, and component and material testing. 3D scanners have evolved through several generations. The first devices for digitizing use touch. Although very precise, the problems were complicated use and a very slow measurement process. Moreover, objects of extremely low hardness were almost immeasurable due to deformations that would occur in contact with the surface.

The next generation of 3D devices were 3D scanners, which evolved in several stages:

- First generation with one projector and one camera
- Second generation with two cameras and one projector
- Third generation with encoded light and two axes
- Fourth generation with four cameras and self-calibration

A. Atos Triple Scan: 3D Surface Measurement

Modern scanners consist of two 8 Mpix cameras and a projector that projects colored light. The Atos Triple Scan [2] from the German manufacturer GOM is an example of such scanner technology. The optical fringe projection of the ATOS Triple Scan captures the entire 3D geometry component in a dense point cloud. The projector's role is to help locate the same object point in both cameras by the projection of a fringe pattern (Fig. 1). This method results in three individual viewing perspectives of the object (3in1 sensor). This is also combined with "Blue Light Technology", which makes the sensor more independent from ambient light. Consequently, nowadays it is no longer necessary to update the modified 3D tools surface in the CAD model by morphing or reconstruction. This is made possible by the CAM system, which supports direct milling on STL datasets. This also makes much faster copying and rebuilding possible.

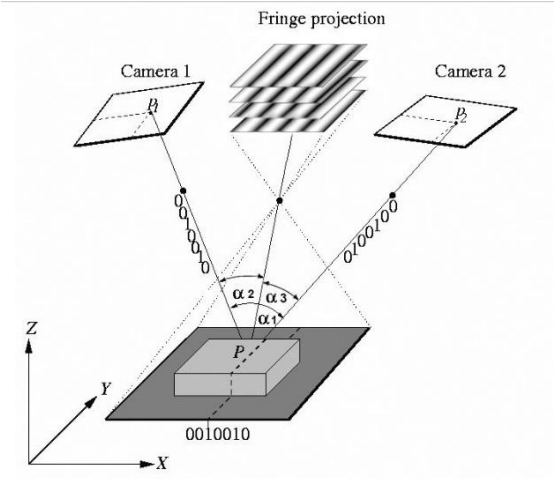


Fig.1. Fringe projection and triangulation

A case study of a tool try-out and wear analysis has shown how to improve maintenance and repair planning, including preventing the production of scrap and analyzing the wear on the tool by comparing the digitized 3D surface and the normal CAD using measuring geometry.

During manufacturing, die have to be modified during try-out. Try-out is used to correlate the complex interaction among the material, tool and press machine. Because of modifications after try-out, final geometries differ from the simulated tool shape and the CAD data model, due to forces such as pressure, friction, tight tolerances and vibration that occur during the stamping operation. These unwanted changes lead to improperly manufactured components that cannot be used in subsequent production process. Nowadays, 3D digitizing is made possible by CAM systems that even support direct milling on STL datasets. These methods also make copying and rebuilding of broken dies in a much shorter time frame possible (Fig. 2).

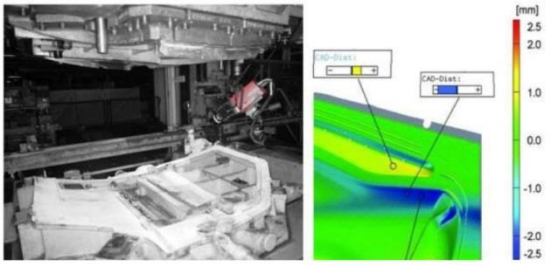


Fig. 2. Measurement of a tool inside a production facility (left) and wear analysis (right)

B. ARGUS: Sheet Metal Forming Analysis

When a final component is being manufactured, it is deformed by powerful forces inside complex tools, which can lead to deformations that are near or beyond the material's limits. By measuring the strain distribution, the ARGUS strain measurement system [2] can provide additional information on how the material is actually deformed, in addition to the CAD model that can check the deviation of the final geometry. The Argus system consists of a handheld camera and notebook, as well as a set of bars with encoded measurement targets. It makes it easy to compare deformation results and the FEM analysis. The measuring process is based on the photogrammetric bundle block adjustment and digital image correlation. Such analysis requires the preparation of the flat original blank with markers that deform together with the metal sheet. The dot pattern is usually made by laser marking, etching or printing. After formation, this pattern is photographed by the user from various directions together with a set of scale bars and encoded measurement targets (Fig. 3). The visualization of a comparison of major strain distribution is then obtained by direct comparison with the numerically obtained strains. The visual representation of the deviations is easy to understand and simplifies the location of areas where improvements need to be made.

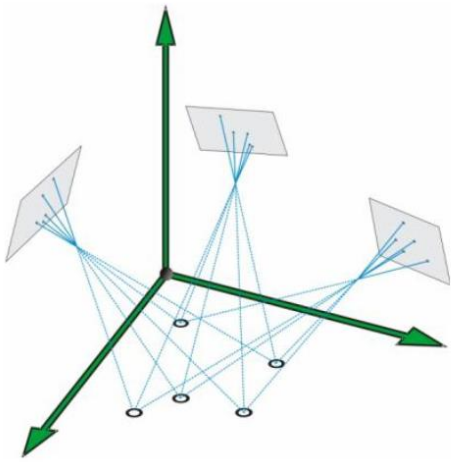


Fig. 3. Photogrammetric triangulation and bundle adjustment

A case study of the verification of numerical forming simulation has shown that ARGUS software makes it easy to compare deformation results and the FEM analysis. It also makes extensive analysis of the complete products of individual components possible under the influence of aerodynamic, thermal or mechanical loads. The necessary inputs for these simulations are the material parameters and a large number of boundary conditions (Fig. 4). Numerical trunk lid simulation results were imported into ARGUS, where they were aligned and registered into the common coordinate system. Visualization of the comparison of major strain distribution on a trunk lid is then obtained by direct comparison with the numerically obtained strains. This easy to understand visual representation of deviation makes it simple to locate the areas where improvements of

the numerical analysis, boundary conditions or tool geometry have to be made.

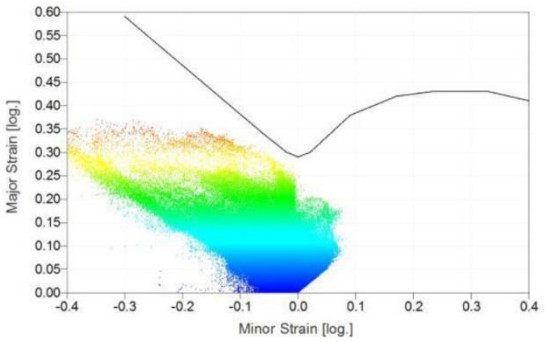


Fig. 4. Forming limit diagram, including strain measuring results and FLC

V. Application challenges

Industrial design is a multidisciplinary activity that involves both creativity and engineering knowledge. Its purpose is to combine applied art and applied science in order to improve the aesthetics, ergonomics, functionality and usability of a product, but it may also be used to improve the product's marketability and production. The objective is to study both the function and form, as well as the connection among the product, the user and the environment. Industrial designers often use various design methodologies in their creative processes. Some commonly used processes include user research, sketching, comparative product research, model making, prototyping and testing. 3D software and CAD are often used to move from concept to production, or to build a prototype and then use industrial scanning to test it for interior defects and also generate a CAD model.

As mentioned previously, there are many great 3D measuring techniques to deliver three-dimensional measurements in order to optimize engineering processes and improve manufacturing workflows of products. Nevertheless, there are still some application challenges in this domain. If visual analysis were applied in the process of 3D modeling, regardless of the type of software, there is a possibility that the product manufacturing process could be minimized even further. Software on the market provides comprehensive 3D modeling, animation, rendering and composition solutions. Frequently used by video game developers, TV commercial studios and architectural visualization studios, they simulate some of the real world such as shadows, draping of fabric, colors, material etc. However, they are unable to predict mechanical properties such as deformation, bending, torsion and deflection. When implemented, visual analysis of these strains would help designers understand the properties of materials and the overall engineering process. This would reduce the gap between designers and engineers, help designers understand the production challenges more easily from the position of engineers , and provide better quality for initial design, reducing the need for subsequent repairs in production and thereby reduce cost. Visual analysis applied in 3D modeling software aims to combine the strengths of computers and humans into an interactive process. In this way visual analytics can help accelerate the development time for products, materials and tools by providing additional information for the effective and efficient improvement of a product. Visual representation should reveal material structural properties data for clear perception and improved understanding by the software user, in this case the industrial designer.

VI. Conclusion

This paper presents a brief review of some software and devices on the market that provide important product information, inspection and quality control. The problems addressed by visual analytics show that all the sciences and many industries rely on the ability to identify methods and models, which can turn data into reliable and provable knowledge. Modern 3D measuring devices can provide this kind of visual analysis, allowing the user to understand and resolve a problem quickly and easily. However, there are still some areas where 3D visual analytics can be used as an effective tool. An option for its application in domain of industrial design software is proposed , although further research is needed.

REFERENCES

[1] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, G. Melançon, "Visual Analytics: Definition, Process, and Challenges", Information visualization, vol. 4950, pp. 154- 175, 2008.

[2] N. Drvar, M. Gomerčić, M. Horvat, "Application of the Optical 3D Measurement Methods in Sheet Metal Processing", Proceedings of the 8th International

Conference on Industrial Tools and Material Processing Technologies ICIT&MPT 2011, 2011.

[3] N. Drvar, "Comparison of the methods for shape and deformation determination of the mechanical structures"[online] Available: http://www.fsb.unizg.hr/library/repository.php

VISUALIZATION SYSTEMS

Remote Visualization of 3D Graphical Models

Dorian Gorgan, Oana Capatana
Computer Science Department, Technical University of Cluj-Napoca, Cluj-Napoca, Romania
dorian.gorgan@cs.utcluj.ro, capatana.oana@gmail.com

Abstract - The complexity of graphical applications has increased considerably in the past years that make them impossible to be executed on computers with limited resources. Such an application domains are Earth Science, scientific simulation and visualization, medicine, physics, and virtual reality. Remote visualization could be a solution for this problem, which relies on rendering graphical applications on remote systems with specialized resources and fast access to huge data models. This paper presents a solution which combines several components that communicate with each other to achieve an efficient and reliable remote visualization system. The design and chosen tools allow the system to deal with constraints such as network bandwidth, latency of the network, and service oriented functionality. The paper explores and analysis the main issues related with GPU graphics clusters.

I. INTRODUCTION

Graphical 3D models have evolved considerably in the past years, in fields like medical image analysis, game industry, architecture or study of natural environment. These models need both large data storage and high computing power of the graphical processing unit. In many cases, one graphical processing unit, even if with very high computing power, may not be enough for rendering complex 3D models. Personal computers, usually, have limited resources. Therefore, a solution must be found that allows visualization on these types of machines.

The solution to this problem is remote visualization. This solution implies a system, with several high computing GPUs, which performs the rendering and sends the result back to the user. Such a system needs to overcome some challenges such as network bandwidth, latency, services allowing the usage of rendering components. One of the main issues is the limited network bandwidth. For instance, to visualize a video of 1280x1024 pixel resolution by 25 frames per second, with 24 bits for each pixel, the network bandwidth should be at least 98Mbps. The latency refers both the graphical processing units and the network latency. Moreover, the services supporting the remote visualization need to ensure the resource allocation, accessibility to allocated resources, and an acceptable level of the usability of these resources.

The research aims to develop and experiment a solution for remote visualization that overcomes the aforementioned challenges. The proposed solution, presented in this paper, is a system which performs the rendering on a graphics cluster. Therefore, the rendering being done remotely, the user is able to visualize complex graphical scenes even if the local Graphical Processing Unit (GPU) does not have high performance capabilities. Web services will be used to make available the functionality of the cluster. Encoding the output of the rendering component into a video will help overcoming network bandwidth limitation. The video will be sent to a streaming server, and the user will be able to watch it using a movie player. The users, generally programmers that create graphical applications, will be able to upload their projects onto the cluster, compile and run them using a user friendly web client.

The rest of the paper is organized as follows. Section II reports related works on remote visualization. Section III presents the functionality and the architecture of the proposed remote visualization system. Section IV describes the practical experiments highlighting the issues and comparing the solutions. Section V concludes on the theoretical and experimental results.

II. RELATED WORKS

Over the past years, several remote visualization systems have been designed, that attempt to solve the problem of massive data sets and limited graphical resources. These systems are generally based on parallel algorithms for rendering. Cedilnik et al. describe in [1] the ParaView framework, used for data analysis and visualization. The ParaView project analyzes very large datasets using distributed computing resources. The system architecture is based on the client/server model. The server is separated into two main components: a data server and a render server. This implies the separation between data processing nodes and rendering nodes. ParaView uses Visualization Toolkit (VTK) as data processor and rendering engine.

Stegmaier et al. present in [2] a solution for hardware-accelerated remote visualization that combines the concept of dynamic linking and the characteristics of X Window System. Dynamic linking relies on that of the most graphical applications use functionality provided by static or shared libraries. For shared libraries only the link to the actual functionality is included in the application. Therefore, loading modules, searching libraries, resolving external references and allocating storage for data structures are performed at run time. The X Window System is a network-transparent graphical windowing system based on the client/server model. The server runs on the local machine and provides its services to the display. The requests to the server are from client applications running locally or remotely. The client and server communicate via X protocol. The main idea of the

solution presented in [2] is that OpenGL rendering requests are redirected to the hardware where the program is running. After completing the rendering, the image is read from a frame buffer and sent to the client machine over the network. The rendering process takes advantage of the advanced graphics features and hardware acceleration of the remote system. The main drawback of this solution is that reading images from the frame buffer and sending them over the network represent expensive operations. Thus, the maximum achievable frame rates are significantly reduced.

Grid services based remote visualization of large medical datasets is presented in [3] by Ahmed et al. Krishnan et al. analyze in [4] the remote visualization and browsing of medical volumes of compressed data.

Streaming is one solution for real time visualization. Unterhinninghofen et al. show in [5] that by using powerful servers through the Mediframe framework the views are rendered and transmitted as video-stream to the mobile devices in real-time. Lamberti and Sanna present in [6] a cluster of PCs, equipped with accelerated graphics cards managed by the Chromium software, which supports remote visualization sessions based on MPEG video streaming involving complex 3D models. The visualization is allowed on smart phones, personal digital assistants (PDAs), and tablet PCs at a frame rate of 30 fps.

Bacu et al. present in [7] a few experiments on the solutions supported by the Chromium [8] graphics cluster to provide fast distributed processing and remote visualization. The performance is evaluated for various cluster configurations and spatial data models. The research concerns as well with accessing the cluster based processing by Web applications through Grid and Web services. Vickery et al. describe in [9] a Web-browser based solution for remote visualization, where two image compression algorithms are proposed for performance enhancement. The paper analyses the way in which the data encryption could affect the final results, concluding that the encryption does not add a significant amount of overhead.

Different techniques are used to reach performance by studying solutions for latency. Hutanu et al. uses in [10] an architecture that hides the high latency of the wide area networks, and shows that an architecture that combines bulk and pipeline processing is the best solution for high-throughput remote data access. Noubissi et al. propose in [11] the remote viewing of large triangular meshes by partitioning the original mesh on the server. The approach generates several fragments of the original mesh at different levels of detail that can be supported by the supposed smaller Transfer Control Protocol (TCP). Finally the client visualizes more accurately for the area of interest, and less accurately the areas out of context.

III. REMOTE VISUALIZATION SYSTEM

A. *Basic Functionality*

The remote visualization system aims to provide the following features:

1. Projects management. The user should be able to:

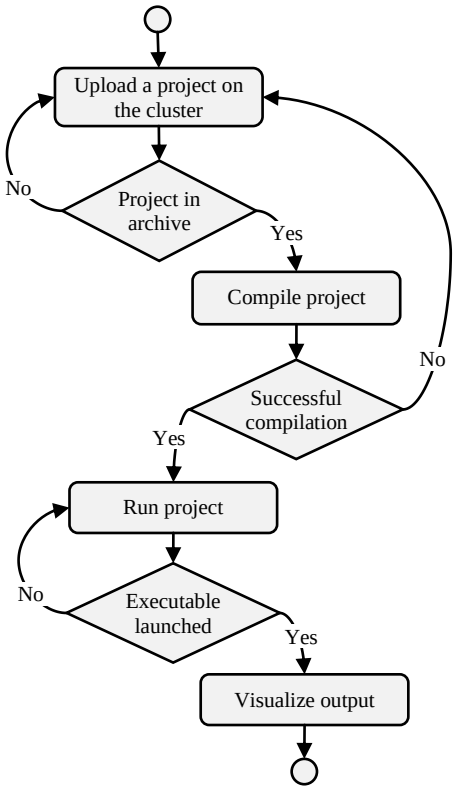


Figure 1. Scenario through which the user uploads, compiles, executes, and visualizes an output video for an OpenGL scene of objects.

   a. Upload projects on the graphics cluster.

   b. View all projects from the cluster, owned by the user.

   c. Choose one of the projects that he owns in order to perform a specification such as compilation and running.

   d. Compile one of the projects, by specifying the source files, the output that will be obtained after compilation and the libraries needed for compilation.

2. Video generation. The user needs to do on the client application:

   a. Choose one of the projects and run them, specifying parameters for the generated video: resolution, frame rate and duration.

   b. View the video from a player using the URL of the stream.

3. Image generation. The user performs on the client application:

   a. Choose one of the projects and request for one or several frames with the scene of objects. The parameters that can be specified are: (i) resolution, for one frame; and (ii) resolution, number of frames, and the interval between frames - for more than one frame.

b. Visualize the frames with parameters previously specified.

By these features the system supports different user scenarios such as the followings:

Scenario 1: OpenGL programmer accesses the system by an easy to use interface, which allows him to upload projects on the graphics cluster, compile them, and view the video output of the rendering process (Figure 1). The user is also interested in specifying parameters for the output video.

Scenario 2: OpenGL programmer chooses one of the projects, and enters the name of the executable and the parameters for generating images: resolution, number of frames to be displayed, and the interval between frames. He runs the already compiled project and view several frames of the scene.

B. System Architecture

The proposed solution is based on the client/server architecture. The modules on each of the components are the following (Figure 2):

1. Server component:
   - Rendering module: it represents the core of the system. For this module a graphics cluster will be used on which the user can run OpenGL applications.
   - Web Services: expose the cluster functionality, and have as well the role of synchronization between the modules on the server.
   - Video encoder: converts the output of the renderer into a video and sends it to the video streamer.
   - Video streamer: receives a video file and streams it back to the user.
   - Image generator: converts the output of the renderer into JPEG files, for them to be displayed to the user.
2. Client component:
   - Web Client: it is used for accessing the Web Services and provides an easy to use interface.
   - Media player: used to play the video coming from the streaming server.

The system modules use a set of protocols to communicate each other. For making requests to the REST Web services [12], HTTP protocol will be used. The operations used, associated with this protocol, are GET and PUT. GET operations are needed for requests such as viewing the projects on the cluster owned by a user. PUT operations support requests such as launching an executable file with parameters, starting the encoder or uploading a project onto the cluster.

Protocols such as real-time transfer protocol (RTP) or real-time streaming protocol (RTSP) are used for streaming media. They allow the transfer of data in real time. The video files are broken in small pieces and send to the user's location in a specific order. RTP/RTSP is the protocol we use for streaming the video. The video that needs to be streamed to the user represents the output of a video encoder and the input of a streaming server.

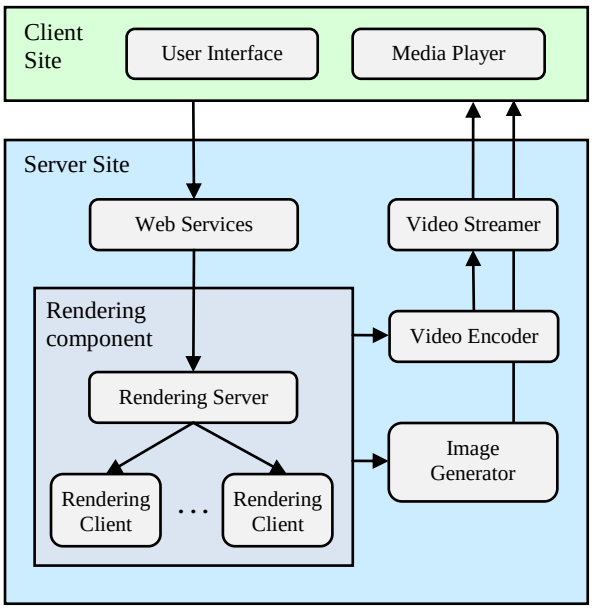


Figure 2. Remote visualization system architecture.

Secure Shell (SSH) network protocol will be used to access remotely the nodes of the cluster and execute commands such as uploading projects or launching executable files.

IV. EXPERIMENTS

A. Functional Evaluation

The functional evaluations aims to prove the proposed solution can be implemented by interoperability of different technological platforms. The main scenario that has to be supported is Scenario 1 (Section III.A): upload a project and visualize an output video.

The experimental steps are the following. The user is asked to provide the credentials. The credentials must match the ones of the cluster account. After the user logged in, he is redirected to the main page of the application.

In order to upload a project, the user navigates to the projects management page. The Figure 3a represents the part of the page that provides the upload functionality. The user can upload a .tar file containing a directory with the source code of his project. The system notifies the user if the project has been successfully uploaded.

In order to compile and run the project, the user navigates back to the main page. The compilation section is presented in Figure 3b. The user chooses one of the projects, enters the name of the source files and of the executable file, and selects the libraries referenced in the source files.

In the console, the user is notified if the project was compiled successfully. Next, the user can visualize the

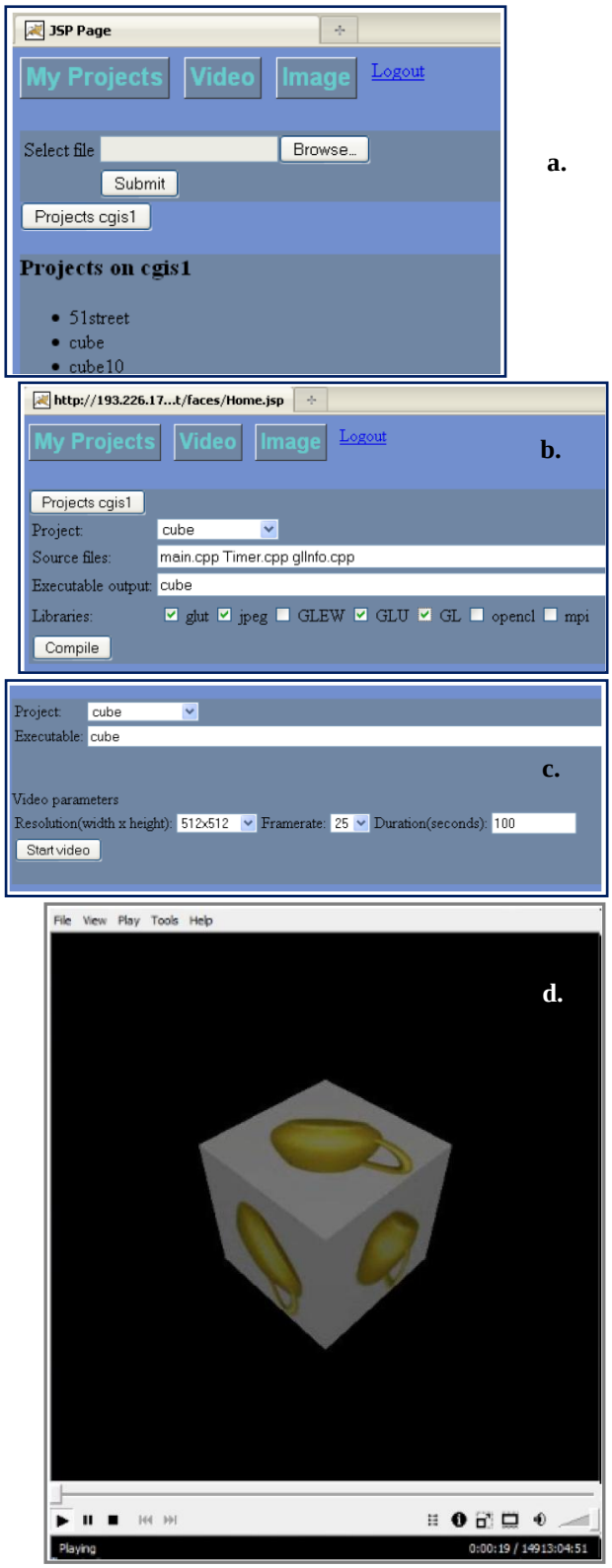


Figure 3. (a) Upload projects on the cluster section. (b) Compilation. (c) Start video generation. (d) Playing the video with MPlayer.

output of rendering application. In Figure 3c it is presented another section from the main page, used for the video generation process. Again, the user chooses the project and enters the name of the executable file. Also, he must enter the parameters for the video: resolution, frame rate and duration.

In case the executable is found in the project folder, the rendering application is started and also the video encoder. The output of the rendering application is a stream of images, which is encoded in a video file. The encoder sends the video to the streaming server.

The system notifies the user about the URL of the video, and the video can be played in a media player using the URL. The output of the rendering application can be viewed as the frames are being generated (Figure 3d).

B. Performance Evaluation

1) Drawing off-screen

Usually, OpenGL is used to render into a window, which is displayed on the screen. Since we do not want to display the output, but to send it to the user as a movie, off-screen rendering will be considered. Off-screen rendering means that the output is rendered into an image buffer. Advantages for using off-screen rendering:

- The rendering process is faster. When drawing onto a window, there exists an overhead associated with OpenGL drawing context switching. Off-screen rendering does not need to perform context switching.
- Images can be generated at higher resolutions than display supports.

OpenGL has an extension providing an interface for creating additional, non-displayable Frame Buffer Objects (FBO) [13]. This type of frame buffer is called application-created frame buffer. Therefore, an OpenGL application can redirect the rendering output to the frame buffer object.

The benefits offered by frame buffer objects are the following:

- They are window-system independent.
- They are easy to set up and save memory.
- They are associated with a single OpenGL context.

A collection of rendering destinations is associated with frame buffer objects: color, depth and stencil buffer. In case an image of a texture object is attached to a frame buffer, it is said that OpenGL performs render to texture. In case an image of a render buffer object is attached to a frame buffer, it is said that OpenGL performs off-screen rendering. The relationship between frame buffer, texture object and render buffer object is highlighted in the

TABLE I. TIME MEASUREMENT FOR RENDERING WITH AND WITHOUT FBO

Number of frames	Time - Render to FBO (sec)	Time - Render to window (sec)
1000	1.739	5.022
3000	4.933	14.709
5000	8.209	24.496
10000	16.228	48.656
15000	24.185	73.049

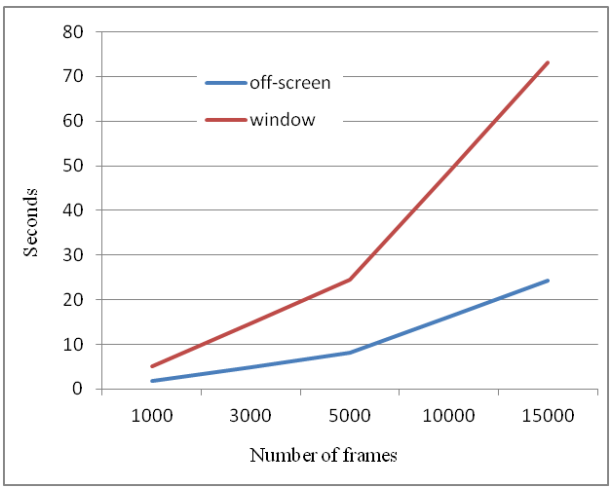


Figure 4. Comparison between rendering to FBO and to window.

tutorial of Song Ho Ahn [13]. A frame buffer object contains several color attachment points, one depth attachment point and one stencil attachment point.

The following tests are used to compare window rendering with off-screen. For this test, was built an OpenGL application, which supports both rendering types. The resolution of the window and the frame buffer was chosen to be 512x512. The tests are based on measuring the time needed for the application to render a specified number of frames.

By analyzing the results from Table I, it can be stated that off-screen rendering reaches a much better performance than window rendering. Figure 4 shows a comparison of the two rendering methods.

2) Video frame rate

The next tests analyze the performance of the system taking into account the frame rate of the video. For obtaining the maximum frame rate, two factors need to be determined: the rate at which the rendering component generates the stream of images and the rate at which the encoder creates the output video.

The experimental results depend on the resolution of the video, as presented in Table II. It can be observed that the number of frames per second decreases significantly as the performance increases. This is due to the fact that the time for saving image files to disk increases as the resolution gets larger.

FFmpeg [14] is a fast audio and video converter that has been used in our experiments. It reads from an arbitrary number of input files and writes to an arbitrary number of output files. FFmpeg allows great flexibility, each input and output file contains any number of streams of different types (video, audio, subtitle, etc). Some of the options that can be specified for the output video file are frame rate, bit rate, number of video frames to record, frame size, pixel format.

The input to a FFmpeg instance can also be a stream of images, that will be converted by the encoder into a video file.

TABLE II. COMPUTATION FOR GENERATING THE STREAM OF IMAGES, ENCODING THE IMAGES INTO VIDEO, AND MAXIMUM FINAL VIDEO RATE.

Resolution (pixels)	Generating stream of images (FPS)	Encoding the images into video (FPS)	Maximum frame rate for the video (FPS)
320x240	176.5	424	170
512x512	62.9	135	60
640x480	55.5	117	50
800x600	36.1	109	33
1024x768	23.3	66	20

FFmpeg supports many video, image and audio codecs: MPEG-1, MPEG-2, MPEG-4, H.264, JPEG, MP3, HE-AAC, etc. Also, there are many video formats supported as input or output, such as AVI, FLV, MP4, etc.

For determining the encoding speed of FFmpeg, the encoding process was tested on the outputs from the previous tests. No frame rate was specified for reading the input. Table II presents the obtained results.

By analyzing the results, it can be observed that, no matter the resolution, the rate at which the encoder creates the video is higher than the rate at which the rendering component generates the stream of images. The rate at which the encoder reads the image sequence will be set a bit smaller than the one of the rendering component to ensure that the images are generated before encoding them in the video.

Another experiment concerns with the maximum frame rates for the video. We use in our experiments the FFserver [15] as a streaming server for video and audio files. It supports several live feeds coming from different FFmpeg instances. When started, FFserver runs in daemon mode. FFserver receives prerecorded files or FFM streams from an FFmpeg instance and streams them over HTTP/RTP/RTSP. Input streams are called feeds. The details about the streams could concern resolution, frame rate, bit rate, format, name of the video, port on which the server will listen for feeds, etc. The measurements are

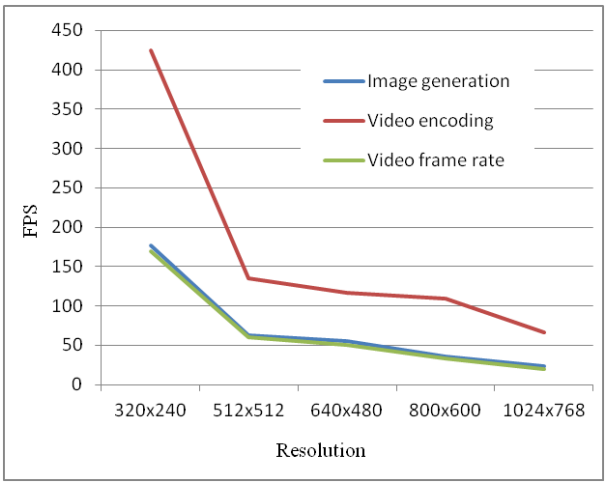


Figure 5. Comparison between computation of generating the stream of images, encoding the images into video, and maximum final video rate.

presented in Table II.

The first conclusion that can be drawn after performing these tests is that the frame rate of the video depends only on the rate of generating images by the rendering component (Figure 5). The second conclusion, regarding performance, is that for medium resolution values the frame rate is very good. For larger values, though, it decreases significantly, but it is still enough to give the user the illusion of movement.

V. CONCLUSIONS

This paper presents a solution for remote visualization of graphical 3D models. The solution combines several technologies and several programming languages to create the components that work together to achieve the functionality desired by the user. The developed system meets the requirements of providing the functionality necessary for uploading and running OpenGL applications. The system allows the user to visualize the output of the rendering component as the frames are being rendered. In matter of performance, the frame rate of the output video has very good values for medium resolution sizes.

One of the drawbacks of the system is the video generation process has a slow start, meaning that the user needs to wait for a few seconds after he requested for the video, to be able to play it from a media player. Another disadvantage of the system would be that in case a high resolution is desired, and also a high value of the frame rate, this would not be possible. It has been shown by experiments that the maximum frame rate for the video at a resolution of 1024x768 is approximately 20 frames per second. Another drawback is the fact that in case the user wants to modify the source files of a project, he needs to do that on his local machine and afterwards, upload the project again.

There are several future directions that need to be followed: (a) Firstly, regarding performance, a solution should be found such that visualization is supported also at high resolutions; (b) User interaction should support the control of viewing parameters, not just watching. The user needs to modify the position of objects in the scene and the camera's parameters; (c) Management of shared resources would allow users to view projects uploaded on the cluster that are made public by their owners. This would also support team work between developers; (d) Source editor would be useful in cases when the users want to modify the sources of the project and recompile; (e) Adding individual source files to projects the user does not go through all the steps implied by the general uploading process.

ACKNOWLEDGMENT

The research has been carried out on the graphics cluster in the Computer Graphics and Interactive Systems Laboratory of the Technical University of Cluj-Napoca.

REFERENCES

[1] A. Cedilnik, B. Geveci, K. Moreland, J. Ahrens, and J. Favre, "Remote large data visualization in the paraview framework", Eurographics Symposium on Parallel Graphics and Visualization, pp. 163-170, 2006.

[2] S. Stegmaier, M. Magallon, and T. Ertl, "A generic solution for hardware-accelerated remote visualization", in Proceedings of the symposium on Data Visualisation 2002 (VISSYM '02). Eurographics Association, Aire-la-Ville, Switzerland, 87-ff, 2002.

[3] A.A. Ahmed, M.S.A. Latiff, K. Abu Bakar, Z.A. Rajion, "Visualization Pipeline for Medical Datasets on Grid Computing Environment", International Conference on Computational Science and its Applications, ICCSA 2007., Aug., pp. 567 - 576. 2007

[4] K. Krishnan, M.W. Marcellin, A. Bilgin, M.S. Nadar, "Efficient transmission of compressed data for remote volume visualization", IEEE Transactions on Medical Imaging, pp. 1189 - 1199, 2006.

[5] R. Unterhinninghofen, F. Giesel, R. Dillmann, "MEDIFRAME - Remote volume rendering visualization framework", IEEE International Conference of Engineering in Medicine and Biology Society,EMBC (2011), pp. 2368-2371, 2011.

[6] F. Lamberti, A. Sanna, "A streaming-based solution for remote visualization of 3D graphics on mobile devices", IEEE Transactions on Visualization and Computer Graphics, pp. 247-260, 2007.

[7] V. Bacu, L. Muresan, D. Gorgan, "Cluster based modeling and visualization of virtual geographical space". Workshop on Grid Computing Applications Development (GridCAD) at the SYNASC Symposium, IEEE Computer Press, pp. 416-421, 2008.

[8] Chromium graphics cluster, http://chromium.sourceforge.net/

[9] R.J. Vickery, J. Martin, J.E. Fowler, R. Moorhead, Y. Dandass, T. Atkison, A. Cedilnik, P. Adams, J. Clarke, "Web-based high performance remote visualization". DoD High Performance Computing Modernization Program Users Group Conference, pp. 364-369, 2007.

[10] A. Hutanu, G. Allen, T. Kosar, "High-performance remote data access for remote visualization". 11th IEEE/ACM International Conference on Grid Computing (GRID 2010), pp. 121-128, 2010.

[11] J.-H. Noubissi, C. Guillet, J.-L. Martinez, "A predictive approach for a real-time remote visualization of large meshes". Second International Conference on Advanced Computing & Communication Technologies (ACCT), pp. 282 - 288, 2012.

[12] Oracle, "The Java EE 6 Tutorial. Chapter 20, Building RESTful Web Services with JAX-RS" [Online]. Available: http://docs.oracle.com/javaee/6/tutorial/doc/giepu.html.

[13] H. A. Song, "OpenGL Frame Buffer Object (FBO)" [Online]. Available: http://www.songho.ca/opengl/gl_fbo.html.

[14] FFmpeg audio and video converter, "ffmpeg documentation." [Online]. Available: http://ffmpeg.org/ffmpeg.html

[15] FFserver streaming server for video and audio, "FFserver documentation." [Online]. Available: http://ffmpeg.org/ffserver.html

Fast network communities visualization on massively parallel GPU architecture

M. Mržek* and B. Jerman Blažič *
* Jožef Stefan Institute, Ljubljana, Slovenia
maks@e5.ijs.si, borka@e5.ijs.si

Abstract - Modeling phenomena with networks has a wide application in many disciplines including biology, economics, sociology, and computer science. In network analysis modularity is an important measure for automatically extracting communities of closely connected nodes. Another important aspect of the network analysis is network visualization. Different techniques for network layout generation exist and the force-driven layout is one of the most popular ones. However, generating force-driven layouts of large networks is both time consuming and can produce a layout where distinct communities of nodes are not separated, but rather remain untangled. Such layouts are harder to be visually inspected by an end-user. In this paper, we propose a GPU-based implementation of a force-driven algorithm for layout generation. By exploiting the massively parallel architecture of modern GPUs we reduce the computational time by orders of magnitude compared with the CPU-based implementation. Secondly, we implement a multi-layer force-driven method for network layout generation where communities are less entangled. Again, by exploiting the GPU we obtain significant speed-up of computation over the CPU implementations. Our results imply that GPUs can speed up significantly the computations in network analysis and thus larger networks can be analyzed in real-time.

I. INTRODUCTION

Networks are a mathematical representation of a set of objects, called nodes, where some pairs of nodes have relations between them, called edges. Examples of networks include but are not limited to road networks [8] - the intersections are the nodes and the roads between them are edges, social networks [9] – Facebook users are nodes and if two of them are friends an edge between them exists, and protein interaction networks [10] – proteins are connected if they can bind together.

Methods developed for network analysis enable deep insight into the properties of the network. Fundamental network operations like breadth-first search, depth-first search, shortest paths between nodes, etc., occur frequently in this domain. Other popular methods include the ranking of the nodes, discovering cliques, and the identification of connected components of the network. Another technique for gaining more insight into the network that is gaining attention in recent years is the community detection.

Subsets of nodes that are internally more densely connected compared to their complement in the set of all nodes are often interpreted as a community [6]. The detection of those tightly connected groups of nodes is a well-studied problem [16] and modularity is a common measure for community quality [2]. Modularity measures the quality of the partition of the set of nodes into communities with respect to the nodes degree distribution of the network. Many heuristics exist for community detection and in [4] Blondel provides a fast multi-level approach with good quality results.

Visualizing large networks with thousands of nodes or more with layouts that expose the network's structure in detail is an on-going challenge [11,14]. Many layout techniques produce an unpleasing result with unrelated nodes drawn closely together. Such a layout may not reveal any underlying structure thus the produced visualization is dissatisfying. Particularly communities of nodes may not be separated from one another, but rather lay intertwined in the resulting layout.

Force-directed network layouts [1] are a physics inspired technique for obtaining a layout where neighboring nodes are drawn together, but non-neighboring nodes are forced apart. The layout is obtained by applying attracting forces on the edges and repelling forces in between the nodes, starting with an arbitrary, usually random, layout. At each step, those forces move the nodes around, converging towards a lower energy minimum which is a chosen criterion. The main two problems of the force-directed methods are the tendency to converge to local minima and their computational expense.

A good initial layout can help the force-directed layout algorithm to converge faster to better minima and the multi-level layout method is known technique for fast and high quality layout computation [5]. The method starts with the original network that is coarsened into a series of networks, each being simplification of the previous. The coarsest network is then laid out and the resulting layout is expanded back until a layout on the original network is achieved. The resulting layout is often computed in less time and has a lower energy thus can converge to better energy minima.

To address the computational expense we seek solutions in the field of GPGPU. Since the stall of the sequential computing speed [17], multi-core or parallel computing is gaining even more attraction and for that reason the Graphical Processor Unit (GPU) is one of the pace setters in this field. The reasons for utilization of GPUs for General-purpose computing on graphics processing units (GPGPU) include the high computational power, the low cost, and their ubiquity in modern desktop computers s. Today's GPUs are consisted

of from a couple of hundreds to a couple of thousands of cores that enable parallel calculations to amateurs and professionals alike. The list of applications that has gained a computational speedup by utilizing GPUs is growing monthly [18]. In the field of network analysis authors have shown that despite the data access irregularity of common approaches in network analysis, GPUs can as well speed up the computations [12-13][19-22][28].

In this paper we present a massively parallel GPU-based implementation of a multi-layer force-directed layout method that exploits the community detection heuristic's multi-level approach for a multi-layer method as described by Dürr and Brandenburg in [7]. The resulting layout can expose the community structure of the network in an effective way. Further, we compare the GPU and CPU implementations to highlight the speedup gain when using GPUs for computing parallel tasks. The paper is organized as follows: in Section II the used methodology is explained in detail. In Section III the results are presented, and the paper ends with conclusions in Section IV.

II. II METHODOLOGY

A. Force-directed layout

The force-directed layout method for drawing networks by means of an attractive and a repulsive force was first described by Fruchterman and Reingold in [1]. Inspired by forces in physics the method is based on assumption that different forces act between the nodes of a graph. Those forces move the nodes around and after a number of iterations they produce a layout where nodes settle at (local) energy optimum, since the sum of the forces on the nodes is near zero. In our approach the following forces are applied: a repelling Coulomb-like force F_R acting between all nodes, an attracting Hook-like force F_A acting between connected nodes, and a drag force that slows down the movement of the nodes.

First, we define a constant *k*, or the optimal distance between vertices as

$$k = 1/sqrt(n), \quad (1)$$

where *sqrt(n)* is the square root of the number of nodes *n*.

The repelling force F_R is defined to act between any pair of nodes, but its size is inversely proportional to the square of the distance between the nodes. This force was inspired by the Coulomb's force between electrons and is similar to the gravitational force. Many variations for the exact expression of the force exist [1] and the most frequently used variant is as follows:

$$F_R = k*k/dist, \quad (2)$$

where *dist* denotes the Euclidian distance between two nodes. The repellent force in our case is inversely proportional to the distance between nodes.

It must be noticed here that the attracting force F_A acts only between neighboring nodes. The idea comes from the fact that the connected nodes are connected with a spring that follows Hook's law. Our spring force is proportional to the square of the distance between neighboring nodes and is given by the following equation:

$$F_A = - dist*dist/k. \quad (3)$$

The forces result in a movement which is slowed down at each iteration by a cooling scheme that is common in simulated annealing. The cooling scheme helps to decrease the node movement allowing the algorithm to converge. This is achieved by multiplying the movement vector with a scalar *t* called the temperature. The initial temperature is $t < 1$ and at each step of the iteration it is diminished. Some schemes even utilize temperature increase in some cases in order the algorithm to escape the local minima.

In force-directed layouts at each iteration the energy of the layout is defined as the Euclidian norm of the movement that is being forced on the nodes. The measure of energy can be used to evaluate the quality of the force-directed layout, since lower energy layouts imply less conflict in node placement and are thus preferable.

B. Community structure in networks

Networks may contain subsets of nodes that are more interconnected than expected by chance and such group of nodes is defined as a community. This can be qualitatively expressed with the measure of modularity [2]

$$Q = 1/2m \ *\sum(A_{ij} - k_i*k_j / 2/m)\delta(C_i,C_j), \quad (4)$$

where *m* is the number of edges, the sum runs over all the nodes $i \neq j$, k_i is the number of neighbors of node *i*, A_{ij} is *1* if there exists an edge between *i* and *j* and *0* otherwise, and $\delta(C_i,C_j)$ is *1* if *i* and *j* are in the same community and *0* otherwise.

The value of the modularity lies in the range $[-1/2, 1)$ and is positive if the number of edges within groups exceeds the number of edges expected by chance. The maximization of the modularity is interpreted as better community structure in the network. However, the problem of modularity maximization is proven to be NP-hard [3]. So for larger networks the only computationally feasible solution for finding the maximization of the modularity is the use of heuristic algorithm.

An effective heuristic method for modularity maximization was proposed by Blondel in [4]. They have shown that their algorithm produces good quality community decomposition for larger networks and bring high modularity values in reasonable time. The algorithm produces a series of coarsened networks until it reaches a final level in which the modularity cannot be increased any further.

The algorithm starts with the given network in which every node forms a community on its own. In the next

steps the nodes change their community to a neighbor's community and if that increases the modularity then the change remains, otherwise the node remains in his previous community. This is repeated, until no increase in modularity can be obtained. At this point, all nodes in the same community are put together into a meta-node, resulting in a network of first level communities, connected together by meta-edges if nodes in two different communities contain connected nodes that are connected. While the nodes in the original network might not be weighted (meaning that they are weighted equally usually with 1), the meta-nodes gain a weight equal to the sum of all the node weights plus twice the sum of the edge weights in the community. The weight of the meta-edge between two communities is the sum of the weights of edges connecting them in the original network. The new network layout is elaborated again with the same approach, until no increase of modularity is observed. The communities detected at the last level are expanded back onto the original network are the final result which maximizes the modularity.

A side product of this algorithm is the multi-level network coarsening, where smaller communities are coarsened together. The resulting structure is a dendogram of nodes grouped into smaller communities and those communities are grouped into larger ones. This structure is then exploited in the multi-level approach for network visualization.

C. Multi-level layout approach

The multi-level approach that was first published in [5] utilizes a coarsening mapping in order to reduce the network to a smaller one that can be laid out faster and then its layout later expanded to the original network. This method is practical since it provides a good initial layout for drawing large networks which can be computed quicker than iterating a random initial layout with a force-directed method.

The method has three distinct steps: network coarsening, laying out the coarsest network, and expanding the layout from the coarsest to the original network. First, a coarsening mapping of a network is required. A fast and efficient mapping that is gradual and uniform is preferred as suggested by Walshaw in [5]. In our case, the dendogram from the modularity maximization heuristic described in section II.B provides the coarsening mapping of nodes from one level to the next by producing series of networks $G_0 = G, G_1, \ldots, G_{L-1}$, where G is the original network, G_{L-1} is the coarsest network and L is the level of coarsening.

In the next step, the coarsest network is laid out with any arbitrary method. Our approach begins with a circular layout that is fed to the force-directed algorithm for a number of iterations. Finally, the layout at the coarsest level is expanded onto the previous one by interpolation. At this point, the nodes are actually meta-nodes representing communities for the nodes at a lower level. The nodes belonging to the same community are laid near the meta-node position. In our approach, the position of the nodes that are in the same community are the same as the position of the meta-node that represents that particular community and a number of force-directed iterations is applied to separate the nodes. This step is repeated until the original network is laid out.

D. GPGPU with CUDA

General-purpose computing with graphics processing units (GPGPU) is used for applications traditionally handled by the CPUs. At the hardware level the used Nvidia GPU is a collection of multiprocessors or warps, each containing a number of processors or cores and a small amount of fast shared memory, but all are also able to access slower GPU global device memory. For example, our Nvidia(R) Tesla(R) M2090 has 16 warps with 32 cores each, for a total amount of 512 cores that can compute in parallel. The amount of shared memory is 48 kBytes and there are 5375 MBytes of available global or device memory. The main bottleneck of the GPU is the massive access to global memory which is relatively slow, thus the number of global memory read-write operations has to be minimal.

CUDA (Computer Unified Device Architecture) [23] is a development framework for writing GPU programs, without the need of extensive GPU architecture knowledge. CUDA is implemented as a C extension and it assumes that the GPU is a SIMD (Single Instruction Multiple Data) machine that can process an extensive number of threads in parallel, thus allowing us to simplify the implementation of a parallel algorithm on the GPU.

In CUDA a kernel program is executed on a specific number of threads. We design the kernel in order to achieve parallel computation of a particular task. Each thread runs the same kernel, but with different indices. Each thread has its own identifier index that ranges from 0 to the number of threads-1. With the utilization of those indices we achieve parallelism as each thread can address a particular part of the global memory in order to get the data required for computation.

The threads are sorted in a grid of blocks. All the blocks contain an exact user predefined number of threads, organized in a 1D, 2D or 3D array. There is a limit to the number of threads in a single block, for example on our GPU card that was used for the experiment we can run a maximum of 1024 threads in a block. As the threads are organized into blocks, so are the blocks organized into a grid. The grid is a 1D, 2D or 3D division of the total number of total blocks. A limit to the number of blocks exists, but is much higher; in our case 65.536 blocks can be called. With this system of grids of blocks of threads we can divide the workload into the desired parallel plan.

E. Network analysis on the GPU

Since the dawn of GPGPU computing around the year 2006, network analysis on GPUs has not received much attention. The main ideas of network analysis on GPUs are described in [19] and [20] including some basic network algorithms like breadth-first search and the all pairs shortest paths. The most recent breadth-first search GPU technique is described in [21]. An evaluation of the benefits of using GPUs for network analysis is given in [22] and in [28].

Even less work was published in the area of generating network layouts on the GPU. In [13] the authors propose a fast GPU multipole network layout method, while in [12] a multi-level layout implementation on the GPU is presented.

The GPU operates best in the field of operations that are common in linear algebra, rather than with objects and pointers, so a proper network representation has to be selected. The most common computationally-friendly approach to represent networks is with their adjacency matrix, where nodes are labeled from *0* to *n-1*, where *n* is the number of nodes, and if a connection between nodes *i* and *j* exist, than the *ij*-th element of the adjacency matrix equals to 1.

However, it must be noticed that real world networks are sparse. The number of edges *m* is usually of the order of *n* rather to the number of all possible edges in-between the nodes, which is *n(n-1)*. So in a matrix representation of a sparse graph a lot of memory space is wasted.

A sparse representation of networks that is suitable for computation is presented by Harish in [20]. The representation of a network is achieved with two arrays. The first array *neigh* is of the size *n* and the *i*-th element of this array is a pointer to the second array *adjac* and its pointing to the first neighbor of the *i+1*-th node. The *adjac* array is of the size *m* and hold the adjacency lists for each node. With this representation the list of all neighbors of the *i*-th node are stored in *adjac* from the index *neigh[i-1]* to *neigh[i]-1*. This representation of the network takes only $n+m$ space, compared to the n^2 that is required by the matrix representation.

III. RESULTS

The computational results were achieved with a computer that has an Intel(R) Xeon(R) E5-2640 @ 2.50GHz CPU and Nvidia(R) Tesla(R) M2090 GPU. The implementation was written in Python for the CPU part and the computations rely heavily on the Python numpy [24] package. The GPU part is written in C++ CUDA and pyCuda [25] is utilized in order to achieve interoperability between Python and CUDA.

In order to verify the computational speedup of the GPU over the CPU implementation when computing

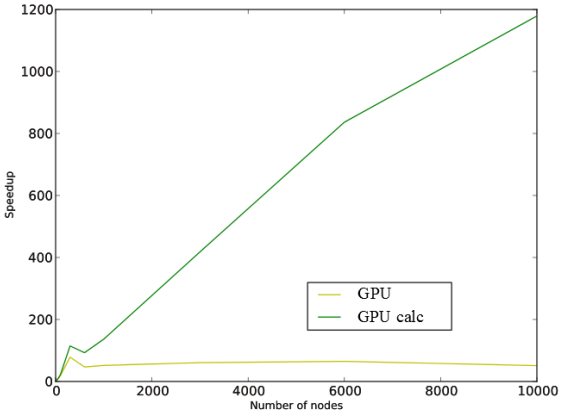


Figure 2. GPU speedup over CPU in the force-directed approach

force-directed layouts was performed, 100 iterations of the force-directed layout algorithm on a set of increasing size networks were first calculated. The class of networks that has a power law degree distribution of nodes, a high clustering coefficient and can be randomly generated was selected [26] as this class of networks has some properties that resemble to real world social networks. The network's size increases from 10 nodes to 10.000 nodes and for each network the layout is computed on the CPU and on the GPU and the employed time is recorded. For the GPU the memory allocation time was recorded separately. Each presented measurement with particular parameters is a mean of 10 consequent measurements.

From the results presented in Fig. 1 it can be seen that the computational time of the force-directed layout is quickly increasing as the network size increases. At a network size of 10.000 nodes the CPU-based implementation takes around 10 minutes to finish, making it highly unpractical. On the other hand, the GPU-based implementation at the same size of the network can produce the same result in about 12 seconds. The third result presented is just the GPU calculation time, without the preparation and the memory transactions to and from the GPU. The total GPU calculation time at the largest network size was around 5 seconds.

The resulting speedups of the GPU implementation over the CPU implementation are plotted in Fig. 2. The

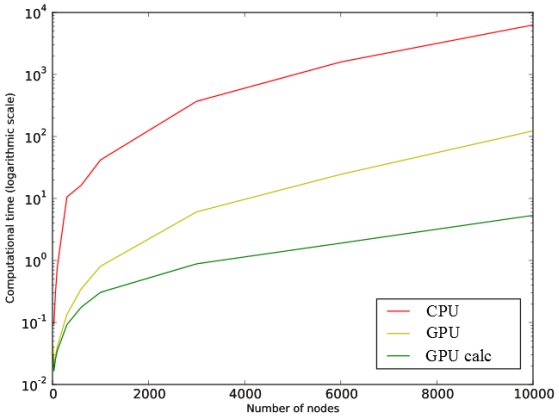


Figure 1. CPU and GPU force-directed 100 iterations computational time at different network sizes

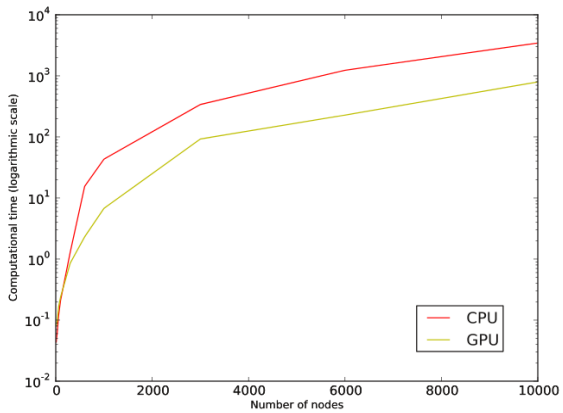


Figure 3. CPU and GPU multi-level computational time at different network sizes

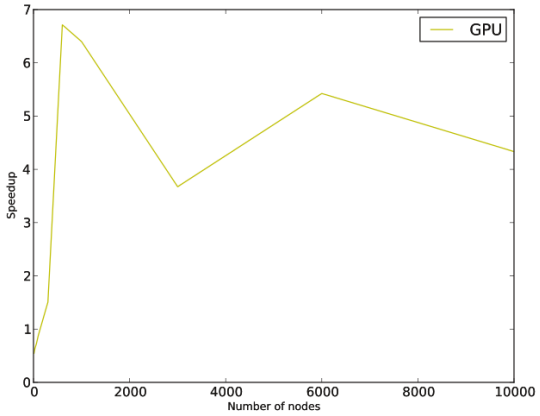


Figure 4. GPU speedup over CPU in the force-directed approach

GPU calculations curve is an upper bound for the GPU speedup at the current kernel implementation.

Next, the layouts computed were compared with the force-directed approach and the community multi-layer approach. For this purpose the so called 'uk' network as described in [27] is laid out with both methods and results after a different number of iterations are presented in Fig. 5. As it can be seen, this network has a strong disjoint community structure that is exposed by the presented multi-layer layout method and after 1000 iterations more the layout is near its final shape.

On the other hand in the pure force-directed approach from a random initial layout, the network's communities get slowly exposed but some of the network's 'tentacles' can remain untangled even after more than 4000 iterations.

Finally, the computational times of the two implementations of the presented multi-layered algorithm, one involving the GPU for computing the force-directed layouts for the different layers was compared. The same conditions as in the first experiment are applied. The results are plotted in Fig. 3 and Fig. 4. The speedup presented in the graphs of Fig. 3 and Fig. 4 is not as high as expected, since most of the computations were done on the CPU.

IV. Conclusion

Interactive visualization of large networks with a community revealing layout is a challenging task. We presented a GPU-based implementation of a multi-layer force-driven layout method that exploits the community structure of the network as revealed by the measure of modularity. This approach can produce a layout where communities are more distinguishable from one another in shorter time than traditional methods. Next, the massively parallel GPU architecture is enabling d to achieve speedups that range from 50 times to over a thousand times, compared to the CPU implementation of the force-directed layout method. The achieved speedups are 3 to 6 times higher when the multi-level approach is used for computation, in cases when the network size is larger than a couple of hundred of nodes. With this method larger networks can be visualized in more reasonable time and the community separating network layout enables the experts to easily inspect the community structure of a given network.

The limitations of our methods include the GPU memory size which is limited, compared to the expandable memory of computers. To overcome this, the algorithm needs to be converted to a scalable variant, where only part of the information is transferred to the GPU for processing. Next, the computation of the repellent forces in the force-directed layout method is the computationally most expensive part of our methods, and our naïve approach has a time complexity $O(n^2)$. This can be improved with the $O(n*log(n))$ approach as suggested [28]. Lastly, the community exposing multi-level approach does not necessarily produce a pleasing layout. Some networks do not exhibit strong community structure and in these cases the described multi-level approach fails to separate the particular communities in the resulting layout. In such cases, the achieved results are computationally and visually similar to standard force-directed layout methods.

Other future work in order to improve the speedup will involve porting as much method's computations as possible to the GPU. The mains parts that could gain significant speedup from GPU parallelization include the computation of the multi-level modularity heuristic and the multi-level expansion of the layout. Another approach that might produce a significant speedup would be to compute everything on the GPU in order to minimize the memory transactions from the main PC memory to the GPU memory.

To conclude, the GPUs are proven to be a powerful computational tool at a moderate price that speeds up significantly the computations required for network analysis and visualization of large networks. This fast growing area will play an important role in many areas, including the network analysis.

Reference

[1] T. M. J. Fruchterman, E. M. Reingold, "Graph drawing by force-directed placement," Software - practice and experience, 21(11), pp.1129-1164, 1991.

[2] M. E. J. Newman, "Modularity and community structure in networks," Proceedings of the National Academy of Sciences 103.23, pp.8577-8582, 2006.

[3] U. Brandes et al., "On Modularity Clustering," IEEE Transactions on Knowledge and Data Engineering, 20(2), pp.172-188, December 2008.

[4] V. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of community hierarchies in large networks," J. Stat. Mech., 10:10008, 2008.

[5] C. Walshaw, "A multilevel algorithm for force-directed graph drawing," Graph Drawing. Springer Berlin/Heidelberg, 2001.

[6] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proc. Natl. Acad. Sci. USA 99 (12): 7821–7826, 2002.

[7] O. Dürr, A. Brandenburg, "Using Community Structure for Complex Network Layout," arXiv preprint arXiv:1207.6282, 2012.

[8] F. B. Zhan, "Three fastest shortest path algorithms on real road networks: Data structures and procedures," Journal of geographic information and decision analysis, 1(1), pp.69-82, 1997.

[9] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, B. Bhattacharjee, "Measurement and analysis of online social

networks," In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, pp.29-42, October 2007

[10] G. D. Bader, C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," BMC bioinformatics, 4(1), 2, 2003.

[11] T. Von Landesberger et al., "Visual Analysis of Large Graphs: State ‐ of ‐ the ‐ Art and Future Research Challenges," In Computer graphics forum (Vol. 30, No. 6, pp. 1719-1749). Blackwell Publishing Ltd., April 2011.

[12] Y. Frishman, A. Tal, "Multi-level graph layout on the GPU," Visualization and Computer Graphics, IEEE Transactions on, 13(6), pp.1310-1319, 2007.

[13] A. Godiyal, J. Hoberock, M. Garland, J. Hart, "Rapid multipole graph drawing on the gpu," In Graph Drawing (pp. 90-101). Springer Berlin/Heidelberg, 2007.

[14] M. J. McGuffin, "Simple algorithms for network visualization: A tutorial," Tsinghua Science and Technology, 17(4), pp.383-398, 2012.

[15] J. Leskovec, K. J. Lang, A. Dasgupta, M. W. Mahoney, "Statistical properties of community structure in large social and information networks," In Proceeding of the 17th international conference on World Wide Web (pp. 695-704). ACM, April 2008

[16] J. Leskovec, K. J. Lang, M. Mahoney, "Empirical comparison of algorithms for network community detection," In Proceedings of the 19th international conference on World wide web (pp. 631-640). ACM, April 2010

[17] J. Parkhurst, J. Darringer, B. Grundmann, "From single core to multi-core: preparing for a new exponential, "In Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design (pp. 67-72). ACM, November 2006.

[18] "CUDA Community Showcase," http://www.nvidia.com/object/cuda_showcase_html.html

[19] P. Harish, P. Narayanan, "Accelerating large graph algorithms on the GPU using CUDA," High Performance Computing–HiPC 2007, pp.197-208, 2007.

[20] P. Harish, V. Vibhav, P. J. Narayanan, "Large graph algorithms for massively multithreaded architectures," International Institute of Information Technology, Tech. Rep. IIIT/TR/2009/74, 2009.

[21] D. Merrill, M. Garland, A. Grimshaw, "Scalable GPU graph traversal," In Proceedings of the 17th ACM SIGPLAN symposium on Principles and Practice of Parallel Programming (pp. 117-128). ACM, February 2012.

[22] A. S. K. Gautham, K. K. P. N. R. Govindarajulu, "Evaluating Centrality Metrics in Real-World Networks on GPU."

[23] "CUDA Programming Guide," http://docs.nvidia.com/cuda/index.html

[24] numpy, http://www.numpy.org/

[25] pyCuda, http://mathema.tician.de/software/pycuda

[26] P. Holme and B. J. Kim, "Growing scale-free networks with tunable clustering", Phys. Rev. E, 65, 026107, 2002.

[27] C. Walshaw, "The graph partitioning archive," http://staffweb.cms.gre.ac.uk/~c.walshaw/partition/, June 2011

[28] P. Sharma, U. Khurana, B. Shneiderman, M. Scharrenbroich, J. Locke, "Speeding up network layout and centrality measures for social computing goals," Social Computing, Behavioral-Cultural Modeling and Prediction, pp.244-251, 2011.

[29] J. Barnes, P. Hut, "A hierarchical 0 (N log iV) force-calculation algorithm," nature, 324, 4, 1986.

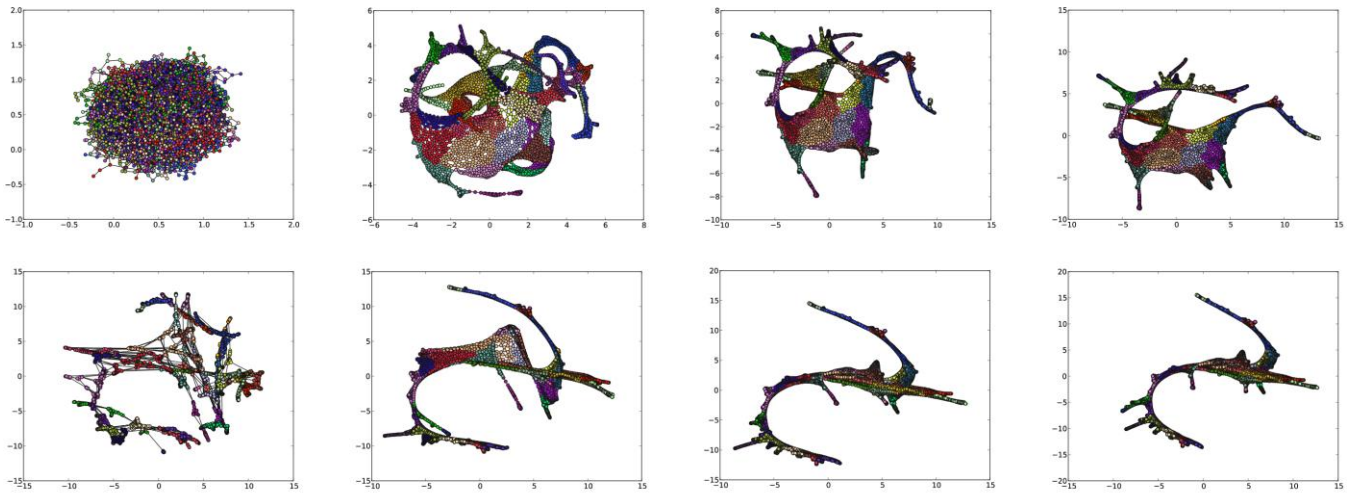


Figure 5. Comparison of the force-directed layout progression of the 'uk' network which has 4824 nodes, 6827 edges, and 38 communities. Nodes belonging to the same community are colored with the same color. In the top row, from left to right are the layouts produced from an initial random layout after 180, 1180, 2180 and 3180 iterations. In the bottom row, the first layout is the result of the described multilayer method as the initial layout, followed by 1000, 2000 and 3000 iterations of the force-directed method. The 180 initial iterations in the first layout took the same time as the initial multi-layer layout, thus the layouts in the same column are generated in equal time.

Realization of Natural Interaction over the Micro Soot Device Model

Petar Mrazović, Marko Pilipović, Mario Volarević, Željka Mihajlović
Faculty of Electrical Engineering and Computing, Zagreb, Croatia
petar.mrazovic@gmail.com, marko.pilipovich@gmail.com, mario.volarevic@gmail.com, zeljka.mihajlovic@fer.hr

Abstract - Paper explores methods of achieving natural interaction between the virtual and the real world, along with their application in industry. Significant emphasis is given to the Microsoft Kinect device, whose functionality was introduced in the field of powertrain systems industry. Collaboration with Austrian company AVL, which is engaged in development, testing and simulation of powertrain systems, has contributed with specific problem that can be solved in augmented reality domain. In order to facilitate maintenance of special device for continuous measurement of lowest soot concentration in the diluted exhaust from internal combustion engines, an interactive virtual service manual has been developed. The application uses interaction with the Microsoft Kinect device which enables users to control and interact with computer world through a natural user interface using gestures and spoken commands. The device finds greatest application in interactive entertainment, but this paper presents and exploits its great potential in industrial environments.

I. INTRODUCTION

With the rapid development of new technologies we are constantly searching for new methods which would affect our experience of interaction with computer world. *Augmented Reality Technology* (AR) superimposes a computer-generated image on a user's view of the real world, thus providing a composite experience of human-computer interaction. AR techniques are used widely in various spheres of human life, and are especially interesting and attractive to apply in the advertising and interactive entertainment. However, the great potential of AR technology has not yet been sufficiently exploited in modern production processes, which leads us to the question of how to use this technology in order to facilitate work in industrial plants.

AR technology is rather interesting and useful way of exploiting modern computer technology and high performance electronics to blur boundaries of real and virtual world. As explained in [1], AR is a variation of the more known concept of *Virtual Reality Technology* (VR) and it is important to notice the difference between the two. VR technology creates virtual environment in which user feels to be moving inside of a computer-generated world, but at the same time he cannot perceive the real world which still surrounds him. On the contrary, AR allows the user to feel the real world, augmenting it with superimposed virtual objects. Awareness of the real world allows user to act natural in complex environment, but also to feel safe at the same time. Precisely this feature is a great potential for AR application in complex or dangerous maintenance tasks in industrial environments.

As concluded in [2, 3], maintenance tasks in industrial environments are excellent domain for AR applications. However, the great potential has not yet been exploited as well as in the field of entertainment. The basic idea of AR is to bring additional information as seamlessly possible into the view of a user, thus providing new perspective and better understanding of real-world situation. Another advantage of AR systems in industrial environments is fault tolerance which can greatly reduce number of errors during maintenance tasks. Economical point of view is also discussed in [1] where authors agree that industries can use AR to lower processes' operational costs and thus sustain their growth and innovation.

There are many available solutions for AR which can be easily adapted for industrial application. Most AR systems make use of simple handheld devices such smartphones, which are usually equipped with compasses, global position system sensors, gyroscopes and cameras. These kinds of well equipped devices provide a large amount of information about real world which can be easily supplemented with virtual features, thus creating a simple AR system. This paper will address more complex AR system which enables user to interact through *natural user interface* (NUI) using gestures. NUI is alternative to a *command-line interface* (CLI) or *graphical user interface* (GUI), but also their possible successor. It allows user to act and feel more natural in interaction with computer world, and enhances their experience of using computer systems. Motion sensing input devices are usually needed to achieve proper natural interaction, and therefore Microsoft Kinect will also be introduced in the paper.

II. INTERACTIVE MANUALS

Collaboration with Austrian company AVL, which is engaged in development, testing and simulation of powertrain systems with internal combustion engines, has contributed with specific problem which is solved in AR domain. The paper follows the process of developing interactive virtual service manual which facilitates maintenance of *AVL Micro Soot Sensor* (Figure 1).

AVL Micro Soot Sensor is device for continuous measurement of lowest soot concentration in the diluted exhaust from internal combustion engines. Its maintenance presents an extremely demanding and responsible job which is carried out as a series of careful tasks. These tasks are animated using supplied 3D model of the Micro Soot device, and present the main feature of the developed interactive software. Software also enables user to manipulate with the model, e.g. rotate, zoom, disassemble device model and explore all of its parts.

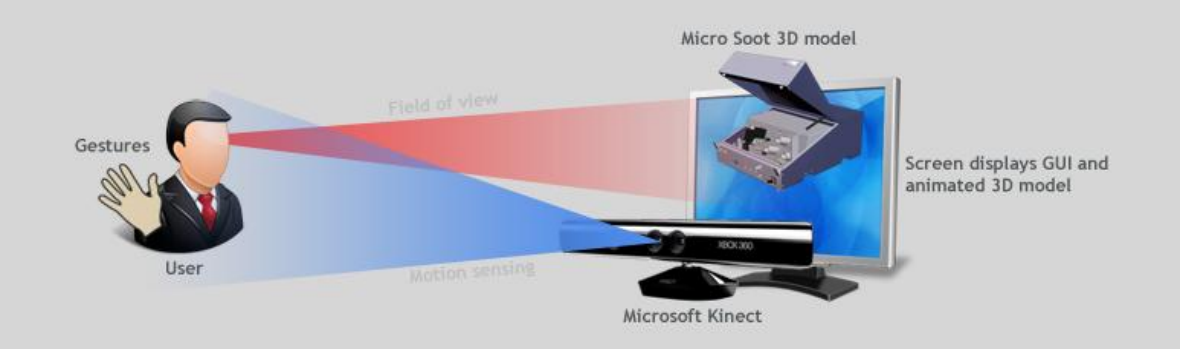


Figure 1 AR system as an interactive virtual service manual

Entire human-computer interaction is implemented using Microsoft Kinect as a motion sensing input device, i.e. user can interact with the device through NUI using gestures. The most of successful Kinect application concern the field of entertainment [1, 4], but this AR system introduces Microsoft Kinect to an entirely new field of application – maintenance in industrial environments.

III. TECHNICAL IMPLEMENTATION DETAILS

During planning and development of this project several problems and lack of features in the standard SDK were encountered, so for this program to function correctly they had to be implemented. Some of the notable examples of these problems and how they were solved will be presented in this segment. The biggest issues were creating an augmented reality environment that uses real 3D graphics engine as a backbone and creating a system that enables intuitive interaction and use of natural user interface.

A. Creating Augmented Reality

To create an augmented reality view where real and virtual world are interconnected, rendering engine and live video feed have to be somehow connected. Since various 3D models (Micro Soot sensor and its parts) and animations (service manual steps) had to be loaded, standard WPF interface couldn't be used because these features aren't supported there so 3D rendering engine has to be used. As it was mentioned before, technologies and SDK*s* that were used are official Microsoft technologies, and were chosen because of compatibility, portability and their ease of use, and for that reason *XNA Game Studio* was chosen as a 3D rendering engine. But creating AR view isn't straightforward. Although Kinect SDK enables the user to get frames from various video feeds (color, depth, infrared) that are retrieved from the Kinect sensor they come in the form of pixel arrays and first have to be converted to a compatible format. For example, bytes from color video frames come in the BGRA format (blue, green, red and alpha channel) and XNA uses RGBA format, so the order of red and blue bytes has to be swapped. When using standard WPF framework color format can be specified as a parameter and conversion takes place automatically but in XNA it has to be done manually and for that purpose small HLSL program was written (since it's a shader everything is done on the GPU so the performance hit is negligible). After the image is converted it has to be drawn to the screen, and to do that first the array of pixels has to be transferred to a 2D texture and then this texture is drawn as a sprite over the full screen. After that, models and GUI (graphical user interface) can also be rendered. Since XNA has an unusual way of deciding what is rendered in which order in the graphical pipeline, special attention has to be given to manual tweaking of rendering order so that the video image is always in the background, models in front of it and GUI on top or problems like drawing invisible or semi-transparent models and GUI may arise.

B. Natural user interface and *Graphical user interface*

Opposed to classical user interfaces where some kind of hardware has to be used (mouse, keyboard, joystick etc.) in combination with various buttons, menus and images to perform a certain action, natural interfaces allow use of voice or intuitive gestures like waving, reaching or grabbing to interact with program with the same purpose like they are used in real world. For example waving might be interpreted as "Hello" or "Goodbye" and could be used to start or stop a program.

In this case grabbing was used as a way to rotate, zoom or disassemble the model. For zooming both hands have to be used and it works in a similar way like a "pinch to zoom" gesture on modern multi-touch smartphones. Problem with using the grabbing gesture is that Kinect doesn't support most gestures "out of the box" so a method to detect whether the fist is open or closed had to be developed [5, 6, 7]. Data that Kinect does provide and which helps in accomplishing this task were skeleton info (contains positions of 20 major joints in the body) and depth image (for each camera pixel its distance from the Kinect is given). First the joint position of each hand is acquired and then using this position as center, rectangular region approximately the size of the open hand is extracted from the depth image. This was done to minimize the amount of scanning (lowers performance costs a lot, instead of checking 640x480 pixels only around 30x30 pixels have to be used). In the next step horizontal and vertical scanning of each line is performed where depth information is used to isolate hand from background objects (no complex computer vision algorithms are needed for edge detection like with conventional 2D cameras). Scanning is used to determine the amount of gaps between fingers (Figure 2) and this information is then used to conclude whether the fist is open or closed (when it is closed there should be no gaps detected). This primitive method works well for a

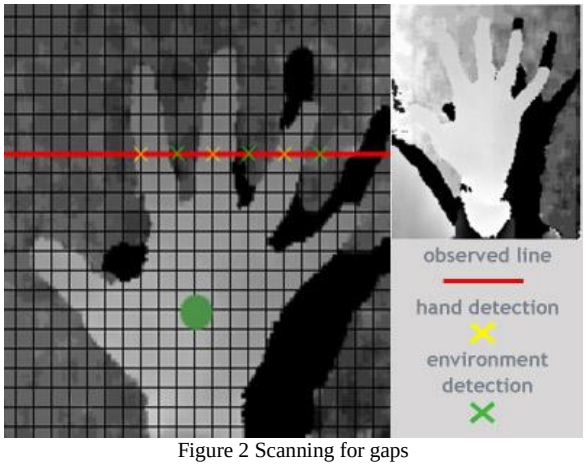


Figure 2 Scanning for gaps

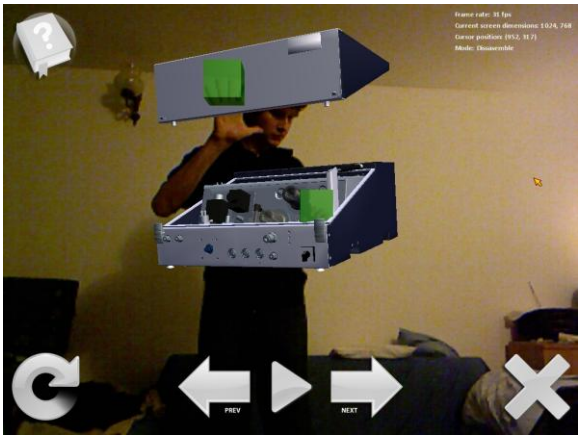


Figure 3 Part moved, hand released

prototype application where false positives can be tolerated but later versions will use some of the more robust algorithms (probably "blob detection" from *OpenCV* library).

After the status of the hand has been detected it can be used for grabbing virtual objects. In the case of model disassembly, grabbing is combined with mesh collision features of the XNA Framework. Usually invisible spheres or boxes are used to approximate the model because of performance reasons. Regarding the box based methods XNA only supports axis aligned bounding boxes and using them would give very bad results when the parts are rotated. Other box based collision method is called object oriented bounding boxes and it is much more precise because it does rotate itself with the model but it is also more computationally expensive and difficult to implement and for our purposes it wouldn't be cost effective to implement it by ourselves. Since the model and parts are rotated a lot, spheres were chosen because they are always the same, regardless of rotation. Minor annoyance is that XNA doesn't automatically calculate the dimensions and positions of the collision spheres depending on the model part but they have to be created manually. Collision detection is activated only if the user has closed his hand and then one sphere is assigned to the position of the hand. The sphere is then checked in a loop with all generated spheres of all parts in the scene. If they intersect then while moving the closed hand grabbed part also moves in the same direction and by the same amount (just like grabbing and moving an object would happen in real world). (Figure 3)

The other two gestures, zoom and rotate, were used in a more 2D way: we take note of the hands vertical and horizontal movement, and ignore the forward/backward movement. The reasoning behind this reduction in dimensions is user oriented – the user is viewing his actions on a 2D surface, it's not intuitive and it's hard to see your hands position in 3D relation to the virtual model. Because of tracking errors we had to include a threshold to our movement detection, and we also had to use fixed values for frame-to-frame scale and rotation updates as to avoid jagged movements of the model. One more point that distinguishes tracking hands with Kinect from tracking with a webcam is that Kinect preserves spatial distances. For example, Kinect will detect that you moved your hands 10cm apart if you're standing 2, 3 or 5m away from the device, while a webcam would detect different distances depending how far away you are form the camera.

Since the program has rich functionality not every action can be performed by gestures. Some of the actions wouldn't be intuitive and would be easily forgettable. this way the meaning of using NUI for its simplicity would be lost, so a classic GUI also had to be created, but it can be used with Kinect and with a mouse. Selection of buttons is performed by placing the hand or mouse on top of it and actions are confirmed, with the mouse by clicking, or hovering for a second over the button with a hand (Figure 4). Hovering was chosen because it is one of the methods described in [8] that can be used with classic GUI. It is also stated that a clear visual feedback of what is happening has to be provided. Therefore the buttons get highlighted when they are selected and while activating the button, the status indicator for the hand is getting more transparent while time passes, which indicates that something is happening.

C. Animations

In order to provide illustrative maintenance instructions, each task was animated in Autodesk 3D Studio Max, and then loaded into XNA program. This way the user is enabled to walk through animated service steps using intuitive navigation buttons.

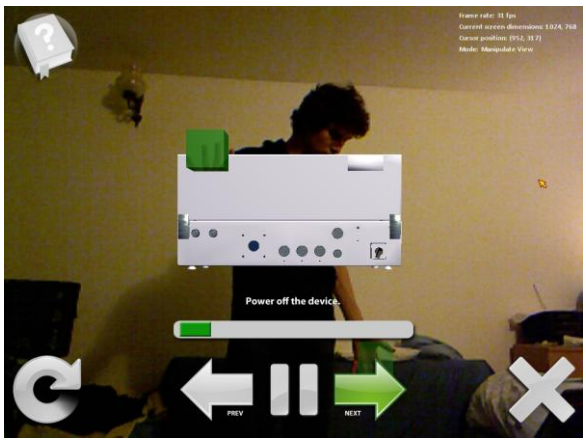


Figure 4 Button selection and activation

Seven animations were created and exported into separate FBX files. These files contain only animations and no geometry. Model meshes were culled down to one triangle reducing the size of the files. These "naked" files are loaded into program where transform information is extracted from them and applied to a preloaded unanimated model with complete meshes, i.e. reference model. The principle is described by Dr. Charles Owen in [9]. Owen developed this solution for loading pre-animated models into XNA programs. The solution extends *Content Pipeline*, which is responsible for loading assets into an object that can be used in XNA code. A custom made content processor class processes imported models and extracts animation data. Special animation data classes were used by the animated model processor to store the animation data and by the software application to load this data at runtime [10].

IV. CONLUSION

The most successful AR applications concern the field of entertainment. However, its ability to provide new perspective and better understanding of real-world situations guarantees wider future application in any kind of industry. Particularly promising areas of application are maintenance tasks in industrial environments. AR technology has proved to be highly suitable for development of interactive service manuals which can greatly facilitate complex maintenance tasks.

Introducing Microsoft Kinect to a new field of application has provided new exciting features in interactive manuals such as NUI. NUI implementation has proved to be very attractive, highly useful and profitable. It allows the user to act and feel more natural in interacting with computer world, and enhances his experience of using interactive manuals. The user can explore task problems more deeply, while feeling extra safe in a fault-tolerant environment. Both users' hands are free while interacting with manual which is sometimes of utmost importance in harsh environments of industrial plants. In the near future NUI will most certainly find greater use in industrial environments, while today its potential is still limited to the field of entertainment.

ACKNOWLEDGMENT

Special thanks are due to all partners from AVL Company, especially to Croatian representatives in Zagreb.

REFERENCES

[1] M. Hincapie, A. Caponio, H. Rios and E.G. Mendivil, "An introduction to Augmented Reality with applications in aeronautical maintenance," *Transparent Optical Networks (ICTON), 2011 13th International Conference on* , vol., no., pp.1-4, 26-30 June 2011.

[2] S. Bernd and L. De Blandine, "An augmented reality system for training and assistance to maintenance in the industrial context," *Journal of Winter School of Computer Graphics*, vol. 11, pp. 101-110, 2003.

[3] N. Zenati, N. Zerhouni and K. Achour, "Assistance to maintenance in industrial process using an augmented reality system," *Industrial Technology, 2004. IEEE ICIT '04. 2004 IEEE International Conference on* , vol.2, no., pp. 848- 852 Vol. 2, 8-10 Dec. 2004.

[4] T. Leyvand, C. Meekhof, Yi-Chen Wei, Jian Sun and Baining Guo, "Kinect Identity: Technology and Experience," *Computer* , vol.44, no.4, pp.94-96, April 2011.

[5] M. Tang, "Recognizing Hand Gestures with Microsoft's Kinect," Stanford University, Department of Electrical Engineering, Technical Report, March 16, 2011., unpublished

[6] A. Drake, "Kinect Hand Recognition and Tracking," Washington University in St. Louis, Kinect Hand Recognition and Tracking, Project Report, May 1, 2012., unpublished

[7] F.T. Cerezo, "3D Hand and Finger Recognition using Kinect," University of Granada, Project Report, 2011., unpublished

[8] *Human Interface Guidelines for Kinect for Windows*, Vers. 1.5.0., 2012.

[9] C. Owen, "A Better XNA Skinned Sample," *http://metlab.cse.msu.edu/betterskinned.html*, October 24, 2012.

[10] A. S. Lobao, B. Evangelista, J. A. Leal de Farias and R. Grootjans, *Beginning XNA 3.0 Game Programming: From Novice to Professional*, New York: Springer-Verlag, 2009

Accelerating Mean Shift Image Segmentation with IFGT on Massively Parallel GPU

J. Sirotković*, H.Dujmić ** and V. Papić**
* Siemens CVC d.o.o.
Address: Put Brodarice 6, Split, Croatia
E-Mail: jadran.sirotkovic@siemens.com
** Faculty of electrical engineering, mechanical engineering and naval architecture, University of Split, Croatia
Address: Ruđera Boškovića 32, Split, Croatia
E-Mail: hdujmic@fesb.hr, vladan.papic@fesb.hr

Abstract – Mean shift algorithm is a popular technique in many machine vision applications including image segmentation. Main drawback of the original algorithm is its quadratic computational complexity, the problem approached with many acceleration methods developed by researchers so far. One of the most effective is usage of the Improved Fast Gauss Transformation (IFGT) to accelerate Gaussian summations of the mean shift, resulting with linear computational complexity. Despite such advances, mean shift segmentation on larger images can still be too expensive for time critical applications. However, recent rapid increase in the performance of general purpose graphic processing unit (GPGPU) hardware has opened opportunity for significant acceleration of the algorithms by parallel execution. This paper introduces first parallel implementation of IFGT-MS segmentor based on many core GPGPU platform. The emphasis is placed on adaptation of the core algorithm to efficiently exploit benefits of underlying GPU hardware architecture. Numerical experiments have demonstrated considerably faster segmentation execution compared with alternative CPU and GPU based mean shift variants.

I. INTRODUCTION

Image segmentation is an important part of many image processing applications. Segmentation is a process of partitioning digital image into multiple segments in order to change the representation of the image into format that is simpler and easier to analyze. By their nature, the image segmentation operations are also computationally demanding and therefore require efficient algorithms running on powerful hardware in order to meet desired processing speed.

Mean shift [1] is a well established method for image segmentation, but is also widely used in other application areas including, but not limited to, video segmentation, object tracking, image filtering, 3D reconstruction and data analysis. It is an iterative non-parametric clustering algorithm which locates modes of attraction (local maximums in probability density) in multidimensional feature space considered as an empirical probability density function. In the case of image segmentation, each point in feature space represents spatial and color characteristics of the single pixel. Mean shift algorithm iteratively uses kernel density estimator (e.g. Gaussian) to estimate probability density at each point and moves it in direction of the local maxima until convergence. Points which have converged to the same mode are grouped together and pixels they represent are assigned to the common segment of the image. Described approach results in high data adaptivity and robustness compared to parametric clustering algorithms (e.g. k-means). Although, mean shift based systems are known to produce quality image partitions, high computational complexity of iterative kernel evaluation still make them unsuitable for time critical applications.

In the original mean shift algorithm [2], Gaussian kernel density estimator is evaluated by direct computation of multivariate Gaussian sums. Unfortunately, such summation has quadratic computational complexity and it is primary slowdown factor of the whole algorithm. As a faster alternative to direct calculation approximate summation based on Improved Fast Gauss Transformation (IFGT) was proposed in [3]. Such IFGT accelerated mean shift algorithm (IFGT-MS) achieved linear time computational complexity resulting with significant speed-up.

Graphics processing units (GPUs) are increasingly used as co-processing units for central processing units (CPUs) with ability to accelerate various numerical and signal processing applications [4]. GPUs may consist of hundreds of highly decoupled processing cores capable of achieving immense parallel computing performance. It must be emphasized that a critical implication of this architecture change to applications is that they must be parallelized to take advantage of multicore architectures. However, the fastest sequential algorithm may not remain the fastest algorithm after ported to many-core platform [5].

In our paper we present first parallelization strategy for GPGPU accelerated image segmentation using IFGT-MS algorithm. Experimental results show that our approach provides considerable performance improvement over several CPU and GPU based mean shift variants.

The rest of this paper is organized as follows. Section II briefly explains principles of GPGPU, while Section III introduces related materials and explains our contribution. We discuss mathematical background of mean shift and

IFGT algorithms in Section IV. Section V presents our IFGT-MS parallelization strategy. Experimental results are discussed in Section VI and finally Section VII concludes the paper and points out some possible future venues to explore.

II. GPGPU PROGRAMING FRAMEWORK

Our segmentation system is developed for CUDA platform, the latest software and hardware architecture for managing computations on GPU, abstracting it as parallel data processing device [6]. The example of overall system architecture is illustrated in Fig. 1. The GPU is viewed as a set of multiprocessors executing high number of threads in parallel. A typical CUDA application consists of two parts (phases). First part is executed on the host CPU responsible for transferring input data and results between GPU and CPU. Additionally, it is responsible for set-up of device code execution (kernel) by specifying quantity and logical organization of threads to be spawned on the GPU. Second part is completely executed on GPU and it is known as "kernel". Kernels are implemented in CUDA programming language, which is basically C programming language extended with few additional key words. Kernel code segment is executed by multiple threads that run on GPU hardware in parallel. CUDA threads are organized in groups of threads called blocks. Number of blocks and threads each block contain is specified in host code prior to kernel execution. Furthermore, the blocks are grouped into a grid. Execution order of blocks inside the grid is undefined. Mechanisms for communication between threads are shared memory and limited synchronization mechanisms. However, as extensive synchronization can cause performance problems, its usage should be as limited as possible.

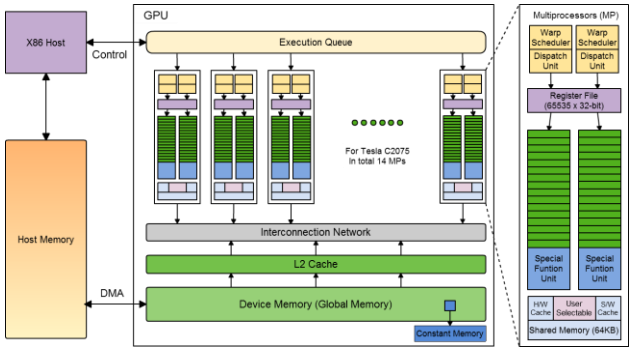


Figure 1. GPGPU system architecture sheme

Threads have access to different types of memory, as illustrated on Fig. 1. In addition, each thread has very fast registers assigned to it. On the block level all threads have access to shared memory assigned to a particular block. If memory access pattern can avoid cache conflicts, shared memory can be accessed with low latency, just as registers. Constant memory is very limited, read-only portion of device memory. Access to constant memory is cached and it can be accessed from any thread running on the device. All threads running on device also have access to cached global memory, which has highest capacity, but higher latency then other types.

Memory access patterns should be carefully designed to avoid bottlenecks, otherwise performance of the GPU can be degraded significantly. Shared memory access by multiple threads in parallel can lead to memory bank conflicts serializing execution of affected threads [7]. Also, simultaneous access from multiple threads to global memory addresses should be arranged so that multiple memory requests can be coalesced into single continuous memory access. The number of threads active on the device is limited by the capacity of shared memory and registers. Device code should be optimized to run as many threads in parallel as possible and to mask relatively high latency access to main memory by having enough threads to execute while others are waiting on memory operation.

III. RELATED WORK AND OUR CONTRIBUTION

Mean shift method was first introduced in [2], and further investigated in [1] and [8]. So far, many improvements in the performance have been proposed, but the majority is developed for Single Instruction Single Data (SISD) processors (CPUs) which partly ignore potentially parallel characteristics of the algorithm [9] [10][11].

However, some recent publications have introduced GPGPU based approaches for accelerating mean shift. Parallel mean shift algorithm accelerated by GPU has been presented in [12], which also includes feature space simplification by sub-sampling and some cluster merge heuristic. Mode seeking is performed entirely on GPU and results with significant speed-up compared to the CPU executed version. Another paper [13], simplifies feature space by partition with Gaussian kd-tree and using it as base for reduced space mean shift computation. Algorithm is also executed on the GPU and has reported significant performance gain.

Fast summation IFGT algorithm itself is developed and successfully applied to mean shift in [3]. IFGT-MS combination resulted with high speed-up, especially for medium and large kernel windows. Method is further explored in [14] resulting with definition of improved error criterion. Problem of automatic parameter selection is addressed in [15]. Mentioned publications have been used as starting point for our work.

IV. MEAN SHIFT CLUSTERING WITH IFGT

A. *Review of mean shift*

Mean shift uses Parzen window [16] kernel to evaluate probability density directly from data. Given n data points $x_1 \dots x_n$ in d-dimensional space, the kernel density estimator with kernel function $K(x)$ and window bandwidth h, is given by the expression:

$$\hat{f} = \frac{1}{nh^d}\sum_{i=1}^{n} K(\frac{x - x_i}{h}) \quad (1)$$

where the d-variate kernel $K(x)$ is nonnegative and integrates to one. It is possible to use various kernel functions, but Gaussian is often popular choice. In its essence, the mean shift algorithm is a steepest descent procedure which requires estimation of the density gradient. Mean shift expression (2) is proportional to normalized density gradient and always points toward the steepest accent direction of the function.

$$m(x) = \frac{\sum_{i=1}^{n} x_i\, g(\left\|\frac{x - x_i}{h}\right\|^2)}{\sum_{i=1}^{n} g(\left\|\frac{x - x_i}{h}\right\|^2)} - x \quad (2)$$

Algorithm iteratively repeats computation of new position (3) until convergence criterion (4) is met:

$$x_j^{k+1} = \frac{\sum_{i=1}^{n} x_i g\left(\left\|\frac{x_i - x_j^k}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{x_i - x_j^k}{h}\right\|^2\right)} \quad (3)$$

Convergence criterion can be defined with term:

$$\left\|x_i^{k+1} - x_i^k\right\| < threshold \quad (4)$$

The algorithm can handle different types of feature spaces such as edge maps or texture, but in most cases of image segmentation, a composite feature space consisting of spatial and color information is used. Each pixel position is represented in the feature space with 5D vector $x = (x_s, y_s, l, u, v)$ where (x_s, y_s) correspond to the spatial coordinates, while (l, u, v) correspond to the color components of LUV color system. In our case, all components are normalized to fit in range [0,1].

B. Fast summation with IFGT

Improved Fast Gauss Transformation (IFGT) aims to approximate Gaussian summations by utilizing space partitioning scheme and series expansion. Quadratic computational complexity is reduced to linear with investment in some additional overhead. The IFGT algorithm is executed in following steps [3]:

1) Assign N points into K clusters using subdivision based on k-centers farthest point algorithm [17]. The final radius of clusters should be less then precision criterion hp, defined by kernel bandwidth h and Taylor series truncation coefficient p. Example of spatial partitioning is visualized on Fig. 2.

2) Choose expansion truncation coefficient p sufficiently large so that that the error estimate is less then desired precision.

3) For each cluster S_k with center c_k and assigned points with weights q_i, compute the Taylor coefficients with expression

$$C_\alpha^k = \frac{2^{|\alpha|}}{\alpha!}\sum_{x_i \in S_k} q_i e^{-\frac{\|x_i - c_k\|^2}{h^2}}\left(\frac{x_i - c_k}{h}\right)^\alpha \quad (5)$$

4) For each target point y_i, find its neighbor clusters whose centers lie within radius hp. Then, sum of Gaussians can be evaluated by expression:

$$G(y_i) = \sum_{\|x_i - c_k\| \le hp_y}\ \sum_{|\alpha| < p} q_i e^{-\frac{\|x_i - c_k\|^2}{h^2}}\left(\frac{x_i - c_k}{h}\right)^\alpha \quad (6)$$

Now we can consider how IFGT based summation is applied to the mean shift. The numerator in the mean shift expression (3) is weighted sum of Gaussians expect that weights are vectors. Also, the denominator in same expression is uniform weighted sum of Gaussians. This means that both can be evaluated by IFGT as $d + 1$ independent sums of Gaussians [3].

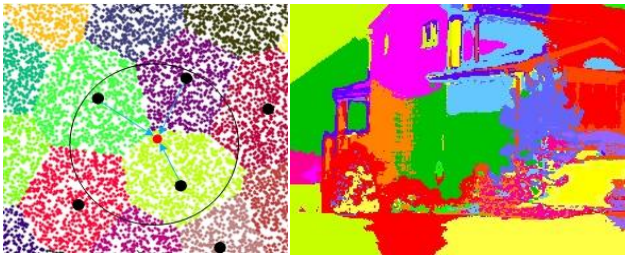


Figure 2. Spatial partitioning by k-centers algorithm. Left: influence of neighbor cluster centers toward evaluated point. Right: example of k-centers partitioning on the image

C. Computational complexity analysis

Computational analysis of IFGT-MS algorithm is presented for key phases of the algorithm:

1) K-centers spatial partitioning takes $O(NK)$ with Gonzalez [17] algorithm where N is number of points (pixels) and K is number of clusters. This can be further decreased to $O(NlogK)$ by using Feder-Green [18] algorithms but such approach will result with significant gains only if number of clusters K is relatively high.

2) Taylor's coefficients computation takes $O(Nr_{pd}(d + 1))$, where r_{pd} represents length of polynomial vector and $(d + 1)$ incremented number of dimensions.

3) Mean shift computation takes $O(Nnr_{pd}\tau(d + 1))$, where n represents maximum number of clusters near each evaluated point, τ number of iterations and $(d + 1)$ dimensions of mean shift sums.

Overall computational complexity is defined with expression:

$$O = NK + r_{pd}(d+1) + Nnr_{pd}\tau(d+1)$$

Number of points N is by far most significant factor, so contribution of other terms in applications such as image segmentation is expected to be insignificant compared to N, i.e. number of pixels. Therefore, computational complexity can be approximated as linear time [3]. In our experiments described in following sections, each phase of the algorithm has been executed on P independent processing elements resulting with significant reduction in computation time. Efficiency and scalability of our parallel implementation is discussed per phase in following chapters.

V. GPU IMPLEMENTATION ANALYSIS

Architectural style of GPU does impose certain limits and programming strategies for efficient exploitation of available computational capabilities. In order to produce efficient code, it is mandatory to have in mind recommended optimization practices and specifics of target hardware architecture (in our case NVIDIA Kepler G104). Original sequential algorithm was analyzed in details and sections with expected high gain from parallel execution have been identified and rewritten as GPU kernels. This resulted with implementation of 5 different CUDA kernels, each accelerating particular phase of the algorithm. Overall program flow is presented with pseudocode below.

Algorithm 1 GPU accelerated IFGT-MS

```
0. Read input data and transfer to GPU
1. K-centers spatial partition
     p = choose(error_limit)
     start with random point as center c_0=rand(p)
     for each c_0 ... c_K do
       dist = find_distance(c_k, p)<gpu_kernel>
       c_k = find_max(dist)<gpu_kernel>
     end
     kcenters = calc_cluster_centers<gpu_kernel>
2. Compute Taylor coefficients
     precompute_term_c()
     compute_taylor_coeffs(p, kcenters,c)<gpu_kernel>
3. Mean shift
     for defined number of iterations
       mean_shift(p,coeffs,kcenters)<gpu_kernel>
4. Transfer results from GPU
5. Merge converged points into image segments
```

A. CUDA memory management

Very important aspect of GPU programming is efficient exploitation of available memory bandwidth. GPUs are characterized by enormous throughput of memory transactions, but fully utilization can be achieved only if algorithm is respecting design constrains imposed by underlying hardware. Therefore, our implementation attempts to follow recommended memory management principles and best practices [19] [7].

Memory transfer between GPU and CPU, as limited with throughput of PCI-E bus, is by far slowest memory transaction. Our implementation minimizes such transfers by doing all operations on the GPU after initial data transfer in early phase of the program. Additionally, memory used for transfers is excluded from operation system paging in order to avoid performance overheads caused by indirect addressing or dislocation of the segment (e.g. on storage device).

GPU memory controllers are designed to coalesce sequential global memory accesses of the thread warp into a single memory access. This means that if each thread of the warp is accessing 4 bytes from main memory, all accesses can be grouped into single memory transaction. Having in mind that single transaction can take up to 400-800 clock cycles, it is obvious that frequent uncoalesced memory transactions can have very degrading impact on the performance. Therefore, it is important to organize data into memory structures which enable efficient coalescing, for this reason our implementation use SoA (Structure of Array) for storage of feature vectors on the device. Each dimension of the vector is stored in separate array, enabling warp threads to coalesce memory access during execution of memory instructions involving feature vectors. Such data organization is visualized on Fig. 3.

Additionally to global memory, our NVIDIA Kepler generation GPU has configurable 48/64 kB of very fast shared memory physically organized as part of SM chips. This memory is accessible only from thread blocks assigned to the SM for execution. One of our design decisions is usage of shared memory for caching Taylor coefficients and cluster centers. As those arrays are intensively accessed by the kernels in second and third phase, we expected performance gain from reducing high latency communication toward global memory. This proved valid decision in case that available memory capacity is adequate to accommodate all coefficients without reduction of occupancy. Unfortunately, this is not possible for high resolution images, so it is preferred to offload some coefficients in the global memory in order to preserve maximum occupancy. Lower occupancy due limited shared memory or register resources means that number of blocks assigned for parallel execution on SM will be reduced, causing lack of stand-by threads valuable for masking arithmetic and memory latency.

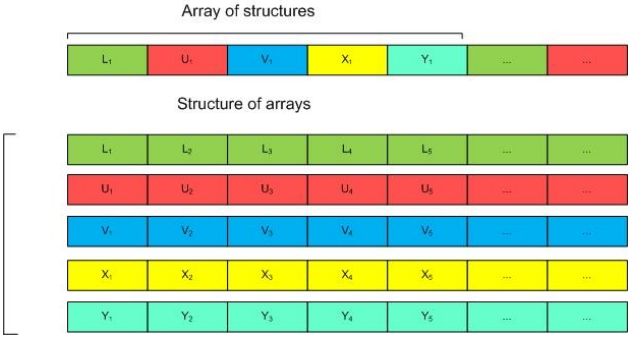


Figure 3. Array of structures vs. Structure of arrays format: examples of in-memory organization of vector components

B. K-centers spatial partition

Gonzales k-centers algorithm [17] is implemented as combination of two kernels which are iteratively invoked for K times and third kernel responsible for calculation of cluster center positions. Expected outcome is uniform partitioning of feature space in K hypercubes, as visualized on Fig. 2. Phase starts by randomly choosing initial cluster center and kernel iteration process begins. First kernel calculates distance between evaluated cluster center and all points in the feature space. Second kernel is in essence parallel reduction of the maximum which finds point with highest distance from evaluated center. That point becomes new center and algorithm is iteratively repeated till all K centers are formed. This effectively divides feature space vectors among K hypercubes. In last phase, cluster centers C_K are calculated by sum reduction kernel. Our algorithm avoids performance inefficient transfer of data back to CPU, instead we are using synchronization pattern based on thread-fence reduction [20][7] to finalize center calculation on the device itself. Meaning, intermediate block specific results are aggregated to final result by the last block of the grid without help of the CPU. Parallel execution of k-centers algorithm achieved significant gains in performance compared with sequential implementation. However, profiler analysis shown that this phase takes disproportional high part of overall execution compared to its computational cost, we find that reason for this lies in high launch overhead of the kernels. Seems that in order to fully exploit GPU acceleration, it would be necessary to find alternative method for spatial partitioning which could allow us to avoid this drawback.

C. Taylor coefficients computation

This phase is fully performed on the GPU by single kernel. Attraction of each cluster center is approximated by Taylor expansion based on contribution of all points assigned to the cluster. Each CUDA block is responsible for calculating transformation coefficients for assigned group of localized pixels based on its distance from cluster centers. Each thread block independently calculates coefficients specific for assigned pixels, in the last step of the procedure, block results are aggregated into set of coefficients valid over complete feature space. Initially, we attempted to perform aggregation with CPU, but this proved not to be very performance effective. Such pattern may be more efficient for older GPU models, but instruction set of newer generation (e.g. NVIDIA Fermi, Kepler) has opened possibility for implementation of more effective paradigms. Aggregation based on thread-fence reduction [20][7] proved to be more efficient solution, even as such reduction has to be performed by only one thread block, it still results with high gain compared to CPU based aggregation. Reason for this lies in fact that we managed to avoid slow transfer of large coefficient arrays between CPU and GPU. Taylor coefficients remained in the GPU memory to be used in subsequent mean shift calculation phase. For smaller images this phase can be effectively accelerated by keeping coefficients in shared memory, but for larger images this is not possible as occupancy issues would occur. We could overcome this problem, but also increase computational efficiency with introduction of spatial indexing. This would give algorithm awareness about position of data it is processing, which would enable thread blocks to consider and reserve storage space only for cluster centers in vicinity of pixel group assigned for processing.

D. Mean shift

Mean shift analysis is executed by dedicated kernel taking results of preceding phases as arguments for calculation. Each CUDA block is responsible for processing group of neighbor pixels, each thread calculates mean shift for one pixel based on density approximated by Taylor expansion around cluster centers. Algorithm is stopped after fixed number of iterations. This approach results with high benefit from parallel execution, but we have also noticed few drawbacks and we can suggest few improvement directions. A lot of overhead is introduced by consideration of all clusters in feature space, which is not efficient as most of influence is limited to clusters in vicinity of evaluated pixel. For this reason, we also believe that introduction of spatial indexing would be beneficial. Another aspect with potential for improvement is implementation of stopping criterion (4), which is causing divergence among threads of the warp (e.g. when few threads of the warp has meet stopping criterion while remaining continues to iterate). Assignment of neighboring pixels to the same thread block can mitigate this to some level, but divergence still remains as slow down factor.

VI. EXPERIMENTAL RESULTS

Our segmentation framework is tested and verified on color images of multiple resolutions, including standard Berkeley database [21] format. As platform for our experiments, we used newer generation NVIDIA GeForce 660 GPU with 960 CUDA cores (streaming processors) organized in 5 streaming multiprocessors and 2048 MB of RAM memory. Sequential algorithms are tested on Intel i3 3.06 GHz host CPU and used as the reference for comparison. Code is compiled on VS 2008 compiler with maximum optimization settings, enabled SIMD extensions and "speed over size" option.

In order to get insight in effectiveness of our solution, we compared it in matter of performance against multiple mean shift implementations based either on GPU or CPU platforms. In lack of publically available code for similar GPU algorithms, in order to increase relevancy of out comparison, we have implemented direct summation mean shift algorithm for GPU based on sequential algorithm [1]. This algorithm is marked with abbreviation DMS (GPU) on performance charts below. Additionally, we did comparison against highly optimized EDISON algorithm [22] and original implementation of the IFGT-MS [3]. Both algorithms are designed for the CPU and are publically available as open source.

First series of experiments is considering performance of image segmentation on different image resolutions. It is clear that both GPU implementations clearly outperform CPU counterparts for order of magnitude on all resolutions. Comparison between direct and IFGT based implementation is more interesting. Parallel IFGT-MS method is performing more efficiently on all resolutions and from Fig 4. it is clear that relative effectiveness of this method is proportional with image resolution.

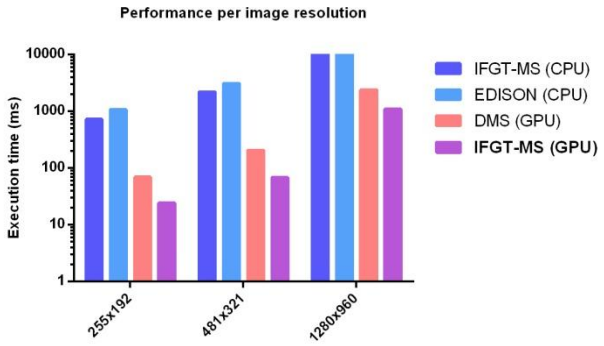


Figure 4. Performance scaling per image resolution for all considered algorithms. Kernel bandwidth in the experiment was set to h_r=0.7 and spatial to h_s=1.4.

Following experimental results reveal more insight in the benefit from GPU execution per algorithm's phases. Execution time of IFGT-MS algorithm is compared per phase between CPU and GPU based implementations. While both Taylor coefficient and mean shift calculation have high benefit from GPU execution, in case of k-center clustering phase performance gain is not scaling as expected. As already pointed out in our analysis, it can be explained with large overhead from high number of iterative kernel launches.

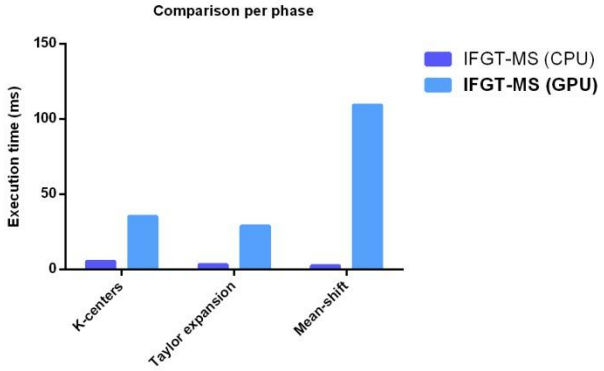


Figure 5. Comparison visualizes per phase acceleration gain between CPU and GPU executed IFGT-MS methods.

Detail evaluation of segmentation quality is out of scope of this paper, but we are planning to address it in subsequent work. Nevertheless, from resulting frames on Fig.7 it is noticeable that IFGT-MS method is producing segments of comparable quality with direct implementation of mean shift.

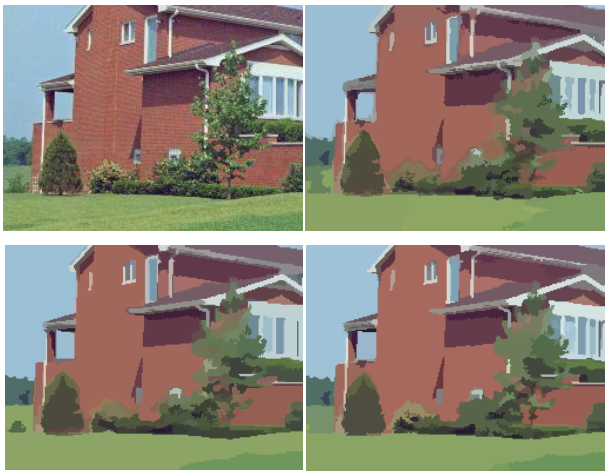


Figure 7. Original test image from Berkeley database (top left), segmentation results produced by IFGT-MS (GPU) (top right), DMS (GPU) (bottom left) and EDISON (CPU) (bottom right) methods

VII. CONCLUSION AND FUTURE WORK

We have presented parallelization strategy for accelerating IFGT-MS image segmentation by execution on massively parallel GPGPU platforms. Achieved results are for order of magnitude better compared to state of art mean shift based algorithms for CPU. Also, significant speed-up is achieved compared to direct summation mean shift executed on the GPU.

Even as considered acceleration strategy has resulted with significant speed-up, our analysis has also revealed few directions for further research. In order to fully exploit full processing power of the GPU, we are of opinion that some parts of the algorithm have to be reformulated. We are planning to address those open topics with our further work.

REFERENCES

[1] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach Toward Feature Space Analysis", in IEEE Transactions on pattern analysis and machine elarning, vol. 24, pp. 603-619, May 2002.

[2] K. Fukunaga and L. Hostetler, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition", in IEEE Transactions on Information Theory, vol. 21, pp. 32-40, Jan 1975.

[3] C. Yang and D. Duraiswani, "Improved fast gauss transformation and efficient kernel density estimation", in International computer vision conference, vol. 1, pp. 664-671, Oct 2002.

[4] D. Kim, V.W. Lee and Y.K. Chen, "Image Processing on Multicore x86 Architectures", in Signal Processing Magazine, IEEE, vol. 27, no. 2, pp. 97 - 107, Mar 2010.

[5] N. Cheung, X. Fan, O. C. Au, M. Kung, "Video Coding on Multicore Graphics Processors", in Signal Processing Magazine, IEEE, vol. 27, no. 2, pp. 79-89, Mar 2010.

[6] S. Che, "A Performance Study of General-Purpose Application on Graphics Processors Using CUDA", in Parallel-Distrib. Comput. Vol. 68, pp. 1370-1380, Oct 2008.

[7] NVIDIA CUDA programming guide, version 5.0, 2012.

[8] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering", in IEEE Transactions on pattern analysis and machine aintellignece, vol. 17, pp. 790-799, Aug 1995.

[9] M. Carreira-Perpinan, "Acceleration strategies for Gaussian mean-shift image segmentation", in IEEE Computer society Conference Computer Vision and Pattern Recognition, vol. 1, pp. 1160-1167, Jun 2006.

[10] D. Freedman and P. Kisilev, "Fast mean shift by compact density representation", in IEEE Conference Computer Vision and Pattern Recognition, pp. 1818-1825, Jun 2009.

[11] P. Wang, D. Lee, A. Gray and J. Regh, "Fast Mean Shift with Accurate and Stable Convergence", in Workshop on Artificial Intelligence and Statistics (AISTATS), pp. 604-611, 2007.

[12] B. Varga and K. Karacs, "High resolution image segmentation using fully parallel mean shift", in EURASIP Journal on Advances in Signal Processing, pp. 1-17, Nov 2011.

[13] C. Xiao and M. Liu, "Efficient mean-shift clustering using gaussian KD-tree", in Computer Graph Forum, vol. 29, pp. 2065-2073, Sep 2010.

[14] V. Raykar, C. Yang, R. Duraiswami and N. Gumerov, "Fast computation of sums of Gaussians in high dimensions", in UMIACS-TR, Nov 2005.

[15] V. Morariu, B. Srinivasan and V. Raykar, "Automatic online tuning for fast Gaussian summation", in Advances in Neural Information Processing Systems, 2008.

[16] R.Duda, P. Hart and D.G. Stork, "Pattern classification", Wiley-Interscience, Oct. 2000.

[17] T. Gonzalez, "Clustering to minimize the maximum intercluster distance", in Theoretical Computer Science, vol. 38, 1985

[18] T. Feder and D. Green, "Optimal algorithms for approximate clustering", in 20th ACM Symp. Theory of computing, Chicago, Illinois, 1988.

[19] NVIDIA Kepler Optimization Guide, version 5.0, 2012.

[20] W. Feng, "To GPU synchronize or not GPU synchronize", in Circuits and Systems (ISCAS), International Symposium IEEE, pp. 3801-3804, May 2010.

[21] "Berkeley Segmentation Dataset"
http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

[22] C. Christoudias, B. Georgescu and P. Meer, "Synergism in low-level vision", in 16th International Conference on Pattern Recognition, Quebec City, Canada, vol. 4, pp. 150-155, August 2002.

[23] H. Guo, P. Guo. and H. Lu, "A fast mean shift procedure with new iteration strategy and re-sampling", in IEEE International Conference Systems, Man and Cybernetics, vol. 3, pp. 2385-2389, Oct 2006.

[24] W. Fang et al. , "Parallel Data Mining on Graphics Processors", Technical Report, Oct 2008.

[25] D.A. Forsyth and J. Ponce, "Machine vision: Modern approach", Prentice Hall, 2007.

[26] S. Xiao-Gu and L. Man-Chun, "Accelerated segmentation approach with CUDA for high spatial resolution remotely sensed images based on improved Mean Shift", in Urban Remote Sensing Event, pp. 1-6, May 2009.

[27] P. Li and L. Xiao, "Mean Shift Parallel Tracking on GPU", in 4th Iberian Conference on Pattern Recognition and Image Analysis, pp. 120-127, Springer, 2009.

[28] B. Fulkerson and S. Soatto, "Really quick shift: Image segmentation on a GPU", in European Conference on Computer Vision, Sep 2010.

[29] M. Liang, M. Huang and J. Gauch, "Accelerating Mean Shift Segmentation Algorithm on Hybrid CPU/GPU Platforms", in International Workshop on Modern Accelerator Technologies for GIScience, Sep 2012.

ARIS MashZone KPI Visualisation of Simulated Routing Protocol Operations for Smart Dust Environments

Kemal Hajdarevic*, Almin Civa **
*,** Faculty of Electrical Engineering, University of Sarajevo,
Sarajevo, Bosnia and Herzegovina
*khajdarevic@etf.unsa.ba, **ac14830@etf.unsa.ba

Abstract – Recent history showed that world needs more mobility, smaller mobile devices, and if possible mobile devices which have multiple sensors. Wireless ad-hoc sensor network technology in 1997 by K. Pister called Smart Dust is on implementation of mobile ad hoc network (MANET) and wireless sensor network (WSN) which might be a solution for future applications. Smart Dust is a network which contains tiny sensor nodes called motes which are able to, use sensors to collects data, to communicate with each other and transfer collected data. While Smart Dust components have relatively high price (WSN classroom Kit starts at 6000 $) commercially available, they are still not in wide usage as education assets at universities because of their price, and due to technology changes. In this paper are presented simulated operation performances of three wireless ad-hoc routing protocols: AODV, DSDV, and OLSR (using NS-3 simulator) which can be used for Smart Dust applications. We used previous research results to recreate similar results of simulated AODV, DSDV, and OLSR and we their Key Performance Indicators (KPI) them with Aris MashZone. Aris MashZone allows easy KPI interpretation and represents them in Dash board as soon as data become available by quick filter adjustments.

I. INTRODUCTION

Ad hoc wireless networks are decentralized networks where nodes themselves serve as intermediaries in forwarding data between source and destination node. One of implementations of mobile ad hoc wireless networks (MANET) are multi-hop wireless sensor networks (WSN). Smart Dust [1] networks are one implementation of WSN which can be used for wide range of applications. Tiny sensor nodes called motes (communication particles called also I-motes) builds Smart Dust networks which operates in the same time as communication nodes and data collection points. Because of mobile and flexible nature of operation for these networks it is an important decision which routing algorithm to use for finding appropriate path between nodes which are implemented in the routing protocols at the network layer. Because of unavailability of WSN or Smart Dust components simulation is an option to realise modus operandi for different routing protocols. One important issue to realise how these networks work is to see appropriate simulation results in appropriate form. Main purpose of this paper is to compare different routing protocols for WSN applications by displaying their KPIs in for this occasion created Aris MashZone [2] appropriate dashboard. Aris MashZone allows visualising KPIs by combining and manipulating data from different sources and in that way to discover new knowledge from existing data. Statement that "MashZone breathes new life into data graveyards" [3] well describes Aris Mash zone capabilities.

A. *Smart Dust WSN*

Smart Dust networks applications are almost limitless comparing it with conventional stationary sensors which are mostly used in industrial and business environments. In the essence two categories for Smart Dust applications exist: environmental and habitat monitoring is first category and second one is structural. Typical Smart Dust applications are asset tracking, monitoring for medical diagnostics, manufacturing process flow, indoor climate control, surveillance, monitoring for agriculture purposes, condition-based equipment maintenance, urban terrain mapping, wildlife habitats, disaster management, emergency response, and many other.

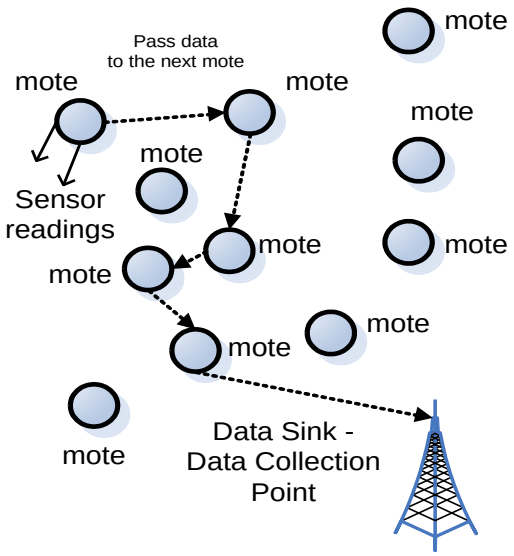


Figure 1.

In practice Smart Dust and I-motes use TinyOS [4] operating system to operate. I-motes can be in size of grain of sand as explained in Hoffman [1] and they can be equipped with different kinds of sensors as it is presented by Culler [5] and [6]. In Figure 1. is shown typical Smart Dust environment which is built of I-motes. Typical Smart Dust application is to collect data such as environmental data readings which are then communicated to the data sink point via neighbouring motes. It can be predicted that prices of I-motes and variants of this technology will be less than it is now because mobile devices such as smart phones or future variants can be used as host to motes or motes itself. Idea that different technologies will be emerged is supported by Edholm's [7] law. Smart phones have sensors, they running capable operating system, they have robust hardware, and in urban surroundings fidelity of mobile phone devices is high because of large number of people that carrying them. Another fact that support idea that Smart Dust will be more available is that we are all aware that every smart phone is equipped with sensors such accelerometers, some other smart phone devices with electromagnetic force meters, noise detector or other sensors such as bar code reader etc. [8]. Earth images from space are already taken by these devices and they are used to control specific satellites [9]. This is all possible be because smart phones are already powerful with their software and hardware equipped with sensors. Competition among manufacturers in smart phone arena is high to create cheap smart phones equipped with sensors resulting in prices reduction. Until we will be able to use smart phones as motes in Smart Dust environment we can simulate routing protocols and mote operations by using network simulators such as NS-3 [10] simulator.

II. SIMULATION OBJECTIVES, SCENARIOS, DESCRIPTIONS, METRICS, HARDWARE AND SOFTWARE COMPONENTS

Through the appropriate choice of simulators and simulation aspects it is possible to display some of the characteristics of wireless networks. Here were analyzed routing protocols generally in ad hoc wireless networks and their performances in the tested conditions. Because of the tendency to have the results as relevant as possible and close to real, NS-3 [10] simulator was selected for simulation purposes to recreate results already presented by [11], [14], [15], [16], [17], [18] to simulate Route Discovery proactive protocols Destination-Sequenced Distance Vector (DSDV), Optimized Link State Routing (OLSR) and Ad Hoc On-Demand Distance Vector (AODV) and present their KIPs in Aris MashZone dashboard.

The aim was to measure the performance of the mentioned protocols in different conditions of network mobility (static nodes, scheduled breaks, faster or slower movement of nodes - described in detail in the next section). Also to test the adaptation ability to changes of network topology, network load, with the correct continuation of forwarding packets between source and destination nodes, and to minimize the time delay of the delivered packets and present results in Aris MashZone presenting wanted scenario by using appropriate filters. Because Aris MashZone contains part for data import and manipulation called data feeds we developed data feed to import data from xls simulation file with appropriate arithmetic operation and data format change.

A. Simulation test bed environment description and simulation scenarios

In order to compare the performance of different protocols, we used consistent simulations. In this simulation model we used a virtual space size 1500x300 [10] meters and the choice of the simulation duration was 500 seconds, for all tested protocols. Such a rectangular area (different side sizes) in the simulations is determined to provide more test paths between nodes than would possibly allow rectangular area with same side sizes what is proposed by [11]. This is important in terms of the increased mobility of network nodes. The total number of nodes in this space in the simulation is fixed at 30 and 60, and the number of nodes that make connections through other nodes (i.e., use them as relay-e) is 10. The mobility of nodes is frequently defined *as random way point* model in which the nodes are moving up to a random point and at random speed for which are defined limits by setting appropriate parameters [11], [12], [13]. In this case, it is the maximum speed of the nodes that have a certain value in simulations 2 m/s 20 m/s (meters that nodes travelled per second) in order to provide a clearer differentiation of the results of testing protocols. Pauses are defined, and the time that nodes spend in a static condition before they start to move in a simulated environment to a randomly selected point. This is important in the ranking and noticing the differences between proactive and reactive protocols and test step value in this case is 100 seconds in total duration of the simulation (0s, 100s, 200s, 300s, 400s, 500s). Pause with 500s duration resulting in no movements of nodes so the results on this points may be the same in different simulations.

Having configuration settings it can be concluded that there are in general four different simulation scenarios:1) 2 m/s node change speed 60 nodes, 2) 20 m/s node change speed 60 nodes, 3) 2 m/s node change speed 30 nodes, 4) 20 m/s node change speed 30 nodes.

It was necessary to define the other parameters of the network such as network interface modules or transmit power in order to ensure a realistic simulation of the NS-3 simulator. The mobility of the network itself, as stated above, is possible to change by changing parameters of velocity and displacement of nodes, but the parameters of the physical layer of nodes as for example in the case of increasing the power of the transmitter or the transmission range reduction also have impact on mobility. In this simulation was used NS-3 RangePropagationLossModel parameter that defines the scope of the transmission of transmitters, with a default value of 250 meters [11], [13]. UDP packet size was used 64 bytes [10], in order to avoid potential congestion occurred due to the size of the packets which are sent at a rate of 4 packets per second [10]. These parameters may be but not necessarily crucial in the process of comparing the performance of the

protocols – they are important by watching simulation alone and certainly contribute to realism, but having that all protocols were configured with the same parameters in all simulations, recorded results show each protocol behaviour under specific circumstances. In the preparation process of simulation scenarios, test parameters were changed repeatedly, such as the duration of the simulation, to make it apparent difference in the results - although, in general, the simulation results in line with expectations of the framework and to make shorter duration of simulations.

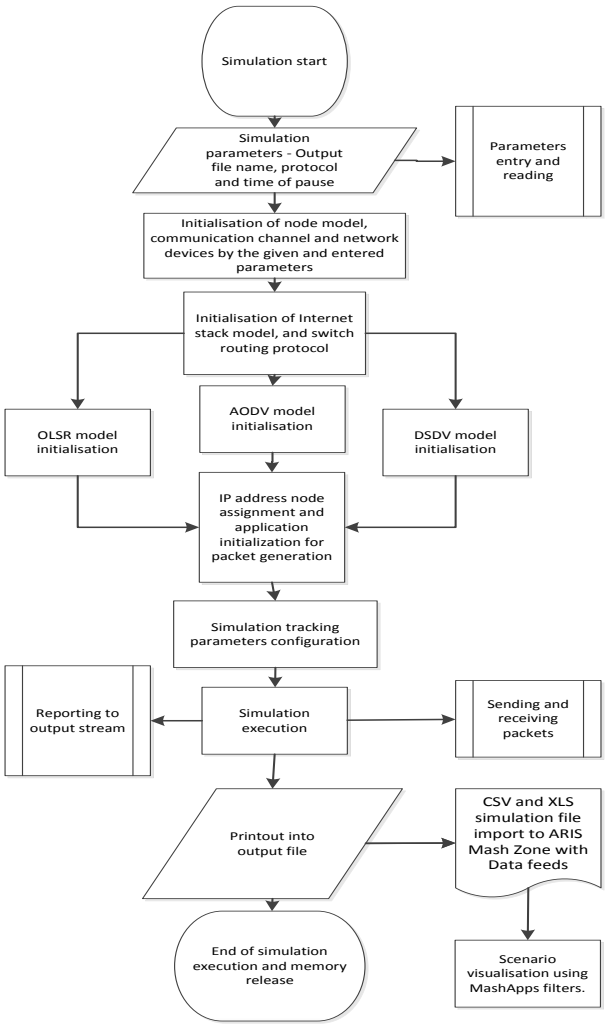


Figure 2. Simulation execution description

By running *waf* compiler, it runs the network and hence simulation results exports to file type of user choice. In this case - CSV and XML files were used, so it is still possible to graphically present results in different tools and we decided to use Aris MashZone.

B. Hardware used for simulation

Simulation itself is done on VM VirtualBox virtual machine with Ubuntu 11.04 installed version of Debian Linux to Windows 7 host and used NS-3:15 version of NS-3 simulator with preset Eclipse development environment. The basic specifications of the computer used for simulation purposes was 10GB HDD, 1.5GB allocated RAM and a processor Intel Core i5 mobile processor generation.

C. Simulation operation issues and problems

The problem that is commonly observed during simulation is high occupancy of virtual hard disk for trace files as the result of simulation. Because of that there was need to enlarge disk size, what is known to be a long process. However, given that demanding simulations occupied over 2GB, it was necessary to access analysis and validation of results before continuing with the second iteration and new simulation. Only selected test samples for animation simulations are used, and the process is largely automated with batch script. Totally 72 simulations were made, and the total size of the resulting simulation file was 64.5GB. For one iteration of the simulation in real time in the test conditions lasted 500 seconds took about half an hour in average - proportional to the number of nodes, wherein specific simulations which included 60 nodes, lasted for about 2 hours.

D. KPI Simulation metrics

KPI metrics used for presented protocols are already used by [11], [14], [15], [16], [17], [18] and we added two more metrics (total discarded and lost packets):

A) Packet Delivery Ratio (PDR) - ratio of the total number of transmitted and received packets.

B) Routing Overhead (RO) - total protocol's control packets transmitted during one simulation iteration.

C) Average Packet Delay (APD) - sending time from the moment of leaving the source to the moment of arrival to the destination.

D) Total discarded packets

E) Total lost packets

Packet Delivery Ratio is calculated using the formula:

Packet Delivery Ratio (PDR) = (packets received * 100) / (transmitted packets)

These packets are related to the packets sent by the NS-3 applications to generate traffic. PDR metric is particularly important because it describes the efficiency of the received or indirectly, lost packets and protocols are characterized by the completeness and correctness of the protocol.

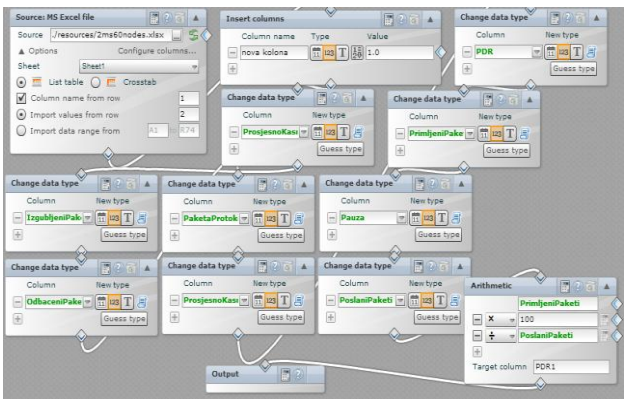


Figure 3. Aris MashZone Data feed

Total transmitted packets include all protocol packets in process simulation of ad-hoc networks. At different jumps, each packet has been transmitted via an intermediary node to its final destination is counted as a separate. The overall package of protocols is scalability because it describes the characteristics of the potential level of network congestion. Especially important for sensor networks where limited bandwidth or battery nodes.

A large number of control packet protocols can also cause collisions and delays in the flow of data in the network [11], [14], [15], [16], [17], [18]. The reliability and efficiency of the protocol with the temporal aspect just shows metric of average packet delay or APD from the sender to the recipient at the network layer. Total delay is the sum of the delays in processing, transmission (and retransmission due to loss of signal between nodes), propagation packet, and the time spent in the queue. Average packet delay is calculated based on the number of received packets, and the time delay is expressed in milliseconds.

Average packet delay = (total delay) / (received packets)

III. MEASUREMENT RESULTS AND PERFORMANCE COMPARISON

As previously described simulation scenarios were performed with maximum node travel speed set parameters to: 2m/s and 20m/s and with number of nodes 30 and 60. Results are shown in separate pie chart diagrams of graphics (from Figure 4. to Figure 8.).

A. Correctly received packets ratio - Packet Delivery Ratio (PDR)

It can be very clearly observed simulation results for metrics Packet Delivery Ratio (PDR) in Figure 4. **OLSR** by this KPI is the most efficient of the three protocols which where compared. At iteration with less mobility, whether it is slower position change or with longer pause break, OLSR achieves over 95% efficiency, because it discards packets received in non-sequential order.

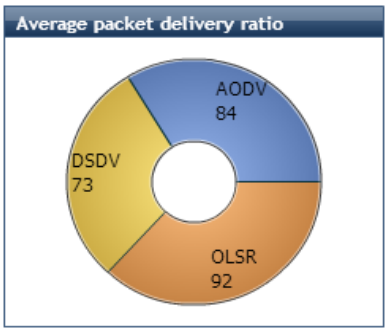


Figure 4. Average Packet Delivery Ratio

AODV has lower performance than OLSR and better than DSDV protocol. **DSDV** has the worst performance in all conditions. As a proactive protocol, resulting line in graph is similar to OLSR but with a 20-30% difference in efficiency due to a lot of packet loss and congestion caused by amount of lost control packets.

B. Total protocol's control packets - Routing Overhead

It can be seen from Figure 5. to Figure 8. that **DSDV** has more control packets than other two simulated protocols for each simulated scenario.

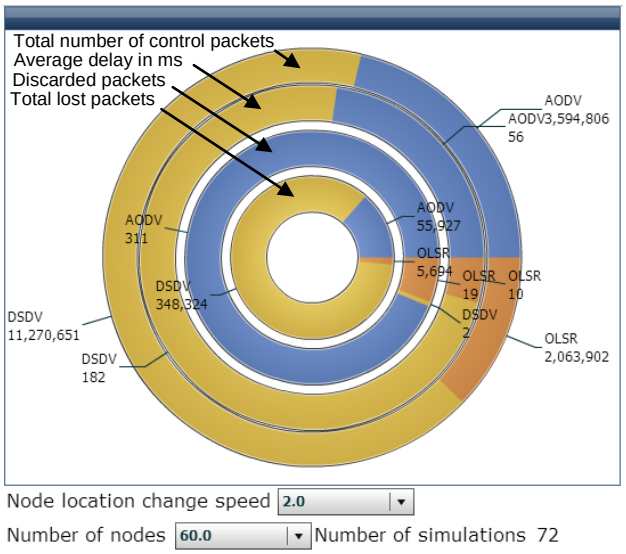


Figure 5. Scenario 1) 2 m/s node change speed, 60 nodes

It is shown large difference in the DSDV participation of control packets compared to AODV and especially to OLSR, and in simulations with 60 nodes, the number of DSDV protocol control packets amount is almost 2.5 million - more than 2 times higher than of AODV protocol, and an order of magnitude higher than in simulations with 30 nodes.

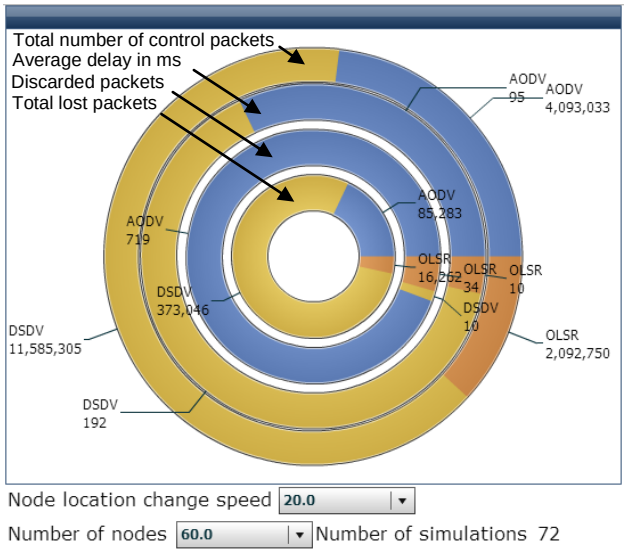


Figure 6. Scenario 2) 20 m/s node change speed, 60 nodes

This shows that DSDV is very bad for networks with a large number of nodes and high congestion. In such circumstances it is difficult to maintain paths for bulky tables for all nodes. **AODV** as a reactive protocol behaves as expected, which is clearly seen in simulations with higher number of nodes. With frequent changes of location nodes, AODV is constantly forced to find a new route to the destination node. The percentage relative to

the total number of packets, DSDV even has a better performance when the network is not too congested and was already familiar with routes (for the simulation parameters of the initial 10-second breaks to generate tables with paths). Regardless of the load, as a proactive protocol **OLSR** periodically broadcast control packets to refresh the routes. Using multipoint relay (MPR) optimized flooding mechanism of adjacent nodes OLSR creates greatest difference compared to DSDV protocol.

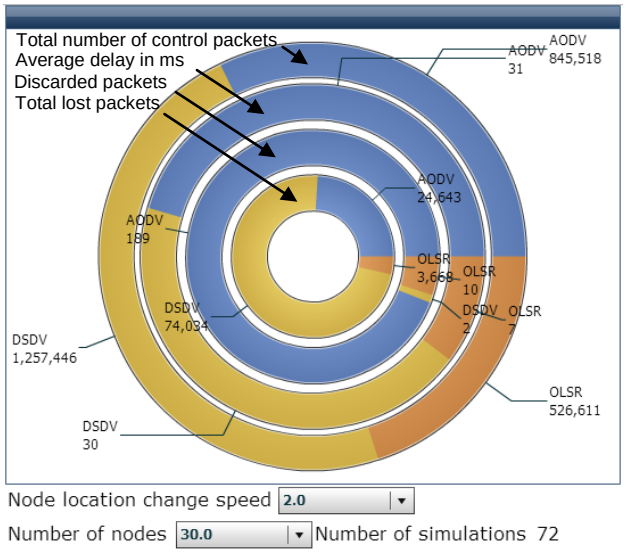


Figure 7.Scenario 3) 2 m/s node change speed, 30 nodes

C. *Average latency / delay*

For Average Packet Delay metric, there is an important difference between the total number of nodes in the network, and in situations with heavy network traffic.

As can be seen from the graphic (Figure 5. to Figure 8), **DSDV** behaves as a slow routing protocol in the conditions in simulations with more than 50 nodes. For scenario with network which has 30 nodes delay is smaller. The number of attempts updating large tables in terms of congestion is evident, while the simulations with 30 nodes showed better results than AODV because the delay of finding a path is lower.

AODV protocol obviously need a larger period of time to search for routes in terms of greater mobility, as demonstrated in simulations with 30 and 60 nodes. However, when the nodes are in a pause state or relative rest, the advantage of the mechanism of establishing the connection comes back into play, and the final results are similar to OLSR protocol.

OLSR here also has the best performance. Behaves very stable in all conditions, with below 2ms delay. As an optimized table-driven protocol, predictably faster in conditions with less mobility but shows a remarkable superiority in conditions of high mobility and network congestion as compared to the other two protocols.

D. *Total discarded packets*

In all simulated scenarios AODV has most discarded packets in all scenarios compared to OLSR and DSDV has lowest number of discarded packet.

E. *Total lost packets*

DSDV protocol has most lost packets compared to AODV and OLSR which has lowest number of lost packets

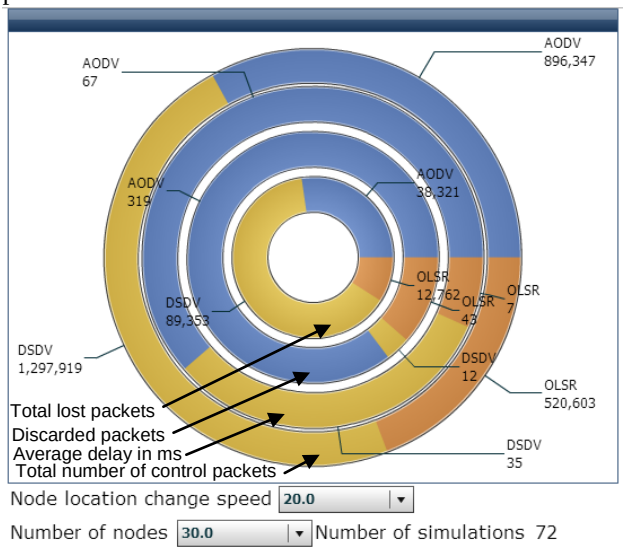


Figure 8. Scenario 4) 20 m/s node change speed, 30 nodes

IV. CONCLUSIONS

Progress in the development of wireless technology and the great advantages of their use, certainly contributed to the diversity of applications of wireless networks in many areas of work and living. Maximum utilization of resources is extremely important in wireless sensor networks where important parameters are throughput, power consumption, reliability, etc.

Regardless of the type of application an ad hoc networks, the ad hoc routing protocols according to the method of determining the path between source and destination can be divided into proactive (table-driven), which contain internal data at each node of the possible pathways and reactive (on-demand) protocols that define the path only when necessary.

Simulations of mobile ad-hoc networks operations provide comparison of effectiveness of selected protocols in different conditions. In addition, the main advantage of simulation compared to the physical implementation of wireless networks is to test the cost and if possible to minimize it. Through simulations and KPI visualisation we have presented some of the key characteristics of wireless networks, such as mobility, flexibility and scalability. We used an open-source simulator NS-3, which supports setting different parameters detailed simulation of networks and network protocols, as well as scalability using available modules. In the presented simulations were compared the performance of proactive protocols DSDV, OLSR, and reactive AODV protocol. The performance of the protocol depends on the simulation scenario, but it is possible to draw general conclusions from the scenarios used.

The simulation of the used parameters, showed that the worst simulated protocol was **DSDV** protocol. The main drawback is the large number of control packets,

which primarily affect the performance. Mechanisms which include routing table repopulation in networks with larger number of nodes, results that DSDV behaves very chaotic and causes a long delay due to packet loss. So it is not recommended for dynamic networks with higher network loads. Because of the need for continuous refreshing of large tables, it is not recommended for sensor networks with limited battery life, and network with bandwidth constraints.

AODV uses Advanced Distance Vector routing algorithm, and connections establishes on-demand instead of maintaining a table with all paths. That's why it needs more time to find a route to a destination at a higher mobility of nodes, but the way how connections are established there are no major delays when the small number of changes in the network topology occurs. Otherwise, it could occur redundancies by sending RREP packets. It is equally adaptable to the conditions of smaller and larger network load, because of reactive approach to establishing connections between nodes. This protocol has most discareded packets compared to other two.

OLSR showed the best performance among the compared protocols in almost all iterations and scenarios of the simulations, in almost all conditions. OSLR behaves very well in terms of larger and more congested networks. Relay multipoint approach obviously reduces the redundancy of control packet flooding in adjacent nodes and optimized update topological information through HELLO messages to quickly find a path to the destination, which of course affects the performance. As it is a proactive protocol, it represents a problem to network bandwidth limitations and rapid changes in the geographical nodes constellation. However, in networks with average mobility, achieved the lowest losses and efficiency of 95-100% correct forwarding packets.

Aris MashZone provides mechanisms to easily present KPIs dynamic changes as soon as they are updated in source files or data bases in web based dashboard.

In the future we hope to see WSN and particularly Smart Dust technology which will be implemented in nano technology small and controllable enough so that they can be placed in live organisms such as human bodies for proactive detection and prevention of specific illnesses in their early development stages.

V. REFERENCES

[1] Smart Dust, Thomas Hoffman, Computer World, 2003, Available from: http://www.computerworld.com/s/article/79572/Smart_Dust [Accessed on 09.01.2013]

[2] Aris MashZone MashUp Avaialble at: http://www.mashzone.com [Accessed on 16.01.2013]

[3] Computerworld, ARIS MashZone: Create the perfect Business Mashup in Just Ten Minutes, Available from: http://www.computerworld.com.au/mediareleases/10123/aris-mashzone-create-the-perfect-business-mashup/ [Accessed on 16.01.2013]

[4] TinyOS, Available from: http://www.tinyos.net/, [Accessed on 09.01.2013]

[5] Culler David, Estrin Deborah, Srivastava Mani, Overview of Sensor Networks, Comoputer Society 2004, Available from: http://www.archrock.com/downloads/resources/IEEE-overview-2004.pdf, [Accessed on 09.01.2013]

[6] Cross Bow, Available from: http://www.xbow.com/, [Accessed on 09.01.2013]

[7] STEVEN CHERRY, Edholm's Law of Bandwidth Telecommunications data rates are as predictable as Moore's Law JULY 2004, Available from: http://spectrum.ieee.org/telecom/wireless/edholms-law-of-bandwidth [Accessed on 09.01.2013]

[8] Richard MacManus, 2010 Trend: Sensors & Mobile Phones, January 17, 2010, Available from: http://www.readwriteweb.com/archives/2010_trend_sensors_mobile_phones.php, [Accessed on 09.01.2013]

[9] Phil Smith, 'Smartphone satellite' developed by Surrey space researchers, Mon, 24 Jan 2011, Available from: http://www.sstl.co.uk/news-and-events?story=1705, [Accessed on 09.01.2013]

[10] NS-3 Simulator, Available at: http://www.nsnam.org [Accessed on 09.01.2013]

[11] Broch,J., A.M. David and B. David, 1998. A Performance Comparison of Multi-hop Wireless Ad Hoc Network Routing Protocols, Proc. IEEE/ACM MOBICOM., Avaialble at: http://www.ages.cs.wisc.edu/~akella/CS740/S07/740-Papers/BMJ+98.pdf [Accessed on 09.01.2013]

[12] Christian Bettstetter, Giovanni Resta, and Paolo Santi, The Node Distribution of the Random Waypoint Mobility Model for Wireless Ad Hoc Networks, 2003.

[13] NS-3 Simulator Documentation, Available at: http://www.nsnam.org/doxygen-release/index.html [Accessed on 09.01.2013]

[14] Samba Sesay, Zongkai Yang, Biao Qi i Jianhua He, Simulation Comparison of Four Wireless Ad hoc Routing Protocols, 2004. Avaialble at: http://www.ic.uff.br/~celio/classes/cmovel/slides/Compara_ad_hoc.pdf [Accessed on 09.01.2013]

[15] S.A.Ade and P.A. Tijare, Perforamnce Comparison of AODV, DSDV, OSLR and DSR Routing Protocols in Mobile Ad Hoc Networks, International Journal of Information Technology and Knowladge Management, July-December 2010, Volume 2, No. 2 pp. 545-548

[16] Ikeda, M.; Hiyama, M.; Kulla, E.; Barolli, L. Mobile Ad-hoc Network Routing Protocols Performance Evaluation Using NS-3 Simulator, Third International Conference on Intelligent Networking and Collaborative Systems (INCoS), 2011, Publication Year: 2012, Page(s): 57 - 60

[17] Ikeda, M.; Hiyama, M.; Kulla, E.; Barolli, L.; Takizawa, M. Multi-hop Wireless Networks Performance Evaluation via NS-3 Simulator, International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA), 2011, Publication Year: 2011 , Page(s): 243 - 249

[18] Ikeda, M.; Kulla, E.; Barolli, L.; Takizawa, M. Wireless Ad-hoc Networks Performance Evaluation Using NS-2 and NS-3 Network Simulators, International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), 2011, Publication Year: 2011 , Page(s): 40 - 45

[19] Santoso, G.Z.; Kang, M.Performance analysis of AODV, DSDV and OLSR in a VANETs safety application scenari, 14th International Conference on Advanced Communication Technology (ICACT), 2012 , Publication Year: 2011 , Page(s): 14 – 20

Physical Simulation of 3D Dynamical Surfaces on Graphics Clusters

Adrian Sabou, Dorian Gorgan
Computer Science Department
Technical University of Cluj-Napoca
Str. G. Baritiu 28, 400027, Cluj-Napoca, Romania
Email: {adrian.sabou, dorian.gorgan}@cs.utcluj.ro

***Abstract*—Physically-based simulation represents a research field of great interest in computer graphics. Mass-spring systems provide a simple yet powerful solution to simulating soft, deformable bodies such as cloth, but have a very high computational cost, especially with high resolution models. Even with highly parallel architectures such as graphics clusters providing the computational power required for these simulations, efficient parallel techniques are needed in order to obtain the best performance in real-time. In this paper we experiment and evaluate a graphics cluster based solution for physical cloth modeling, highlighting issues that arise with model distribution, communication among nodes, distributed rendering and parallel techniques that can be integrated in order to obtain an efficient simulation method, capable of handling complex physical models.**

I. Introduction

Modeling and simulating three-dimensional dynamic surfaces represents one of the main research areas of Computer Graphics. Physically-based modeling techniques aim to offer improved realism and efficiency for real-time simulations, at the same time ensuring unequaled control through numerous simulation parameters and are the natural choice when simulating fluids, textile materials or biological tissue. However, applications employing such techniques utilize computational expensive methods, run-times for realistic scenarios often being too large and, obviously, enhancing the realism of the simulation by increasing the resolution of the surface further aggravates the problem. Taking into account the real-time execution necessity for these scenarios, large, complex models can no longer be simulated using sequential methods and hardware, so parallelization is required in order to continue using physically-based modeling to great success.

Advances in High-Performance Computing (HPC) in the last decade offer a very powerful alternative to traditional CPU-based processing in the form of Graphics Clusters, a parallel computing architecture that is gaining popularity in the scientific computing community due to the massively parallel hardware architecture and high performance of floating point arithmetic and memory operations on Graphics Processing Units (GPUs). The raw processing power and the fast network connections of GPU equipped nodes could make them suitable for the computation required by real-time physically-based simulations, providing that the challenges introduced by such distributed memory architectures can be overcome. Furthermore, modern remote visualization and interaction techniques allow using GPU clusters in order to decouple the rendering resources required for such simulations, thus allowing users using devices with limited graphics power such as notebooks and hand-held devices to simulate and view increasingly complex scenes. Among the challenges introduced by distributed and parallel computation we can mention remote visualization and interaction with the scene, efficient parallelization of physical models, collision detection, model distribution among individual GPUs in the graphics cluster and ensuring communication and synchronization between processes.

In our previous work towards physical simulation of three-dimensional dynamical surfaces we experimented and evaluated basic techniques for simulating cloth on graphics clusters [1]. Experiments carried out proved that our previous solution was highly scalable and efficient, taking into account only the distributed computation process. This paper expands on our previous work by experimenting and evaluating a hybrid GPU-CPU parallelization approach, as well as providing centralized remote visualization of the entire surface. The rest of the paper is organized as follows: In Section II we report on related work. Section III gives a brief account of physically-based modeling for cloth. Section IV discusses our approach to Graphics Cluster based cloth simulation. In Section V we present the strategies for distributed rendering employed. We report on performance measurements in Section VI.

II. Related Works

Using physically-based simulation for reproducing the behavior of dynamic deformable three-dimensional surfaces such as cloth is widely adopted. The research literature on this domain is abundant and we refer the interested reader to the textbook [2].

Paper [1] describes a method for simulating very large resolution textile surfaces (over one million particles) on a graphics cluster using OpenCL for GPU implementation and OpenMPI for communication between cluster nodes. It evaluates the performance of the simulation taking into account only the computational part, without centralized visualization.

Important work in the field of GPU accelerated mass-spring systems is presented in [3] and [4]. These papers explore parallel GPU computation frameworks such as the Compute Unified Device Architecture (CUDA) from NVIDIA, but the techniques used are limited to single GPU systems.

Regarding parallel cloth simulation on multi-core processors, [5] explores possibilities for accelerating physically-based

simulation algorithms on multi-core architectures. Two components of physically-based simulation represent a great potential for bottlenecks in parallelization: implicit time integration and collision handling and this paper addresses both issues. The same issues are addressed in [6] for parallel cloth simulation on distributed memory architectures.

Concerning parallel rendering applications for 3D visualization we can say that while the majority of rendering applications are serial, some applications have the ability to distribute the rendering load onto multiple GPUs. Such popular applications are ParaView [7] and CEIs Ensight [8], or parallel rendering frameworks such as Equalizer [9].

While most work on parallel physical simulation addresses critical performance issues such as numerical integration and complex collision detection, parallelization is mainly attempted at either single GPU or CPU cluster level, hence disregarding the possibility of parallelizing at both levels simultaneously. Furthermore, the models used are in the range of 20k particles and the problem of distributed visualization is therefore neglected. With a considerably larger number of particles (one million and above), due to the need for centralized visualization of the entire surface, distributed rendering becomes an issue that can no longer be ignored, being a potential performance bottleneck.

III. Physically Based Simulation for Cloth

As far as simulating dynamic deformable three-dimensional surfaces such as cloth, physically-based modeling is universally accepted as the best choice. Approaches in this area vary from simplified models for real-time applications such as video games to techniques designed to reproduce accurately certain measured material parameters. However, as HPC architectures such as GPU clusters built from commodity components become cheaper and more performant, those simplified models can considerably increase in size and complexity without affecting the real-time computing requirement. In the following paragraphs we will briefly outline the method which forms the basis for our parallel implementation, the model used being similar to previous work carried out in [1].

For the physical modeling, we rely on a simple mass-spring approach [10]. Our elastic model of cloth is a rectangular mesh of mass points, each of them being linked to its neighbors by massless springs of natural length greater than zero. Three different types of springs are used: springs linking neighbor vertices horizontally and vertically, called "stretch" or "structural" springs, springs linking neighbor vertices obliquely called "shear" springs and springs linking second degree neighbor vertices horizontally and vertically, called "bend" springs. As their name indicates, each type of spring resists to stretching, shearing and bending respectively. The system is governed by Newton's second law of dynamics:

$$\vec{F} = m\vec{a} \tag{1}$$

where $\vec{F}$ is the net force applied, m is the mass of the body, and $\vec{a}$ is the body's acceleration. Thus, the net force applied to a body produces a proportional acceleration.

In a typical force-based mass-spring system there are forces that arise from tension in the elastic springs, called internal and forces that can differ in nature, depending on the simulation, most frequently gravity, wind and collision response, called external. As a spring is extended or compressed, it generates a force of opposite direction as the force applied. This represents Hooke's law of elasticity and states that the force that arises in the spring is proportional to a stiffness coefficient and the rate of deformation.

To obtain the dynamic evolution of the system, Newton's equations of motion have to be stepped forward in time, a process called time integration. There are several types of integration schemes. For our implementation, we chose the velocity Verlet technique [11]. The Verlet integrator offers greater stability at no significant additional cost over simpler methods like the Euler integrator. By using this method, at each step of the simulation new particle positions and velocities are computed using the formulas:

$$\vec{x}(t+\Delta t) = x(t) + v(t)\Delta t + \frac{1}{2}\cdot a(t)\Delta t^2 \tag{2}$$

$$\vec{v}(t+\Delta t) = \vec{v}(t) + \frac{\vec{a}(t)+\vec{a}(t+\Delta t)}{2}\Delta t \tag{3}$$

where $\vec{x}$ is the position, $\vec{v}$ is the velocity, $\vec{a}$ is the acceleration, t is the current time and Δt is the time interval over which we integrate.

The standard steps for implementing this integration methods are as follows:

1) Compute:
$$\vec{v}(t+\frac{1}{2}\Delta t) = \vec{v}(t) + \frac{1}{2}\vec{a}(t)\Delta t$$
2) Compute:
$$\vec{x}(t+\Delta t) = \vec{x}(t) + \vec{v}(t+\frac{1}{2}\Delta t)\Delta t$$
3) Derive $\vec{a}(t+\Delta t)$ using the position of the particles computed at the previous step, $\vec{x}(t+\Delta t)$
4) Compute:
$$\vec{v}(t+\Delta t) = \vec{v}(t) + \frac{1}{2}(\vec{a}(t)+\vec{a}(t+\Delta t))\Delta t$$

By incorporating accelerations, velocity Verlet offers superior stability to the Euler integrator, however, as most explicit integration techniques, is susceptible to instabilities for large values of Δt. Since the computational procedure is identical for all particles and contains no variable loops, this technique is considerably well suited for implementation using highly data parallel architectures such as the GPU. The only observation is that due to the dependency between steps 2 and 3, more specifically the fact that $\vec{a}(t+\Delta t)$ (computed at step 3) is derived using $\vec{x}(t+\Delta t)$ (computed at step 2), positions must be generated for all particles in the model before continuing, thus the numerical integration for a single simulation step must be broken down into two distinct phases, phase one including steps 1 and 2, and phase 2 including steps 3 and 4.

IV. Parallel Physical Simulation

In this section we will discuss in detail our approach for parallel physical simulation on a GPU cluster, applied for simulating cloth. In order to provide context, the first part will be a short overview of work carried out previously and presented in [1] while the second part will describe parallel techniques used to improve on that work, more specifically we will discuss a hybrid GPU-CPU parallelization approach.

A. Cloth Modelling On a GPU Cluster

As presented in section III, using a simple force-based mass-spring system for cloth modeling, coupled with a velocity Verlet integration technique represents a problem displaying a high level of parallelism, since at each of the two integration passes, positions and velocities respectively can be computed independently for each particle in the model. This makes it possible to have the computational process for each particle handled by a different thread, while having virtually no communication required between individual threads. The only data required by a thread beside the position and velocity of its own particle are the positions of neighboring particles (i.e. particles directly connected through elastic springs to the current particle). Obviously, a high number of threads that can be executed in parallel leads to more data that can be processed at the same time and implicitly to better performances, hence such a simulation is extremely well suited to be implemented using the GPU. The versatility, high performance of floating point arithmetic and memory operations of the GPU and the fact that it became fully programmable through development kits such as Open Computing Language (OpenCL) [12] or NVIDIA's Compute Unified Device Architecture (CUDA) [13], provides us with cheap and powerful general purpose accelerators that can be used to perform the expensive computation required for physically-based simulations.

For our implementation we chose to use OpenCL due to its open-source nature and the fact that it is implemented solely as a library, which presents the advantage of easier integration with parallel development tools such as the Message Passing Interface (MPI), since GPU kernels are compiled by the OpenCL library itself at runtime instead of being compiled in batch along with the rest of the application. The idea was to use OpenCL for the numerical time integration (i.e. generate new positions and velocities for the particles at each simulation step) while displaying the cloth using OpenGL, thus using the GPU for both general purpose computing and rendering in the same application. This type of usage introduces several challenges mostly due to the transfer time and memory overhead existent if maintaining particle data both in OpenCL and OpenGL buffers. The solution is to take advantage of an OpenCL/OpenGL interoperability mode, through Vertex Buffer Objects (VBO). The VBO is an extension for OpenGL intended to improve performance by providing benefits of vertex arrays and display lists while avoiding downsides of their implementation. VBO allows vertex array data storage in high-performance graphics memory on the server side and efficient data transfer. Using VBO, the number of function calls and redundant usage of shared vertices can be reduced. Moreover, the OpenCL/OpenGL interoperability mode allows for OpenGL data manipulation by OpenCL without the need to transfer data back and forth, through switching ownership of a shared VBO.

This parallel implementation offers greatly improved performances over sequential ones, but is however limited in the number of particles it can effectively process at the same time by the number of threads that can be spawned and implicitly by the capabilities of the OpenCL device used. Thus, in order to obtain a highly scalable simulation, computation must be broken down into smaller equally weighted computation processes, with good data locality, between the GPU equipped nodes of a graphics cluster. This means partitioning the model of the cloth into groups of the same size and statically assign each group to a GPU. The positions of neighboring particles that belong to different GPUs has to be communicated, therefore, in order to minimize the communication overhead, it is essential to split up the model in a way that ensures minimal length borders. In our case, since the model implemented was a rectangular piece of cloth, this is easily achieved by dividing it into horizontal patches of equal size. In the case of irregular cloth surfaces efficient model distribution can be obtained by applying a multilevel k-way graph partitioning algorithm [14]. To ensure minimal modifications to the kernels used for the single node version of the application, each node keeps an extended model, consisting of its own patch and neighboring particles that belong to other nodes, so called ghost-points, whose positions have to be synchronized at each step of the simulation. Communication, synchronization and the transfer of data between the clusters nodes is ensured through the Message Passing Interface (MPI), a language independent protocol used in parallel computing that has become the *de facto* standard for communication among processes that run on a distributed memory system.

B. A Hybrid GPU-CPU Parallelization Approach

We have so far discussed about parallelization techniques that allow for efficient simulation of cloth on a graphics cluster with regard to the massively data parallel computation capabilities of the GPU. However, analyzing the steps required for generating each frame, as presented in algorithm IV-B.1, we discovered that the parallelization process can be taken even further by employing a hybrid CPU-GPU approach. The total processing time for simulation step *i* can be computed as:

$$\begin{aligned}\tau_{sim_step}(i) = \tau_{kernel}(i) + \tau_{sync_common}(i) \\ +\tau_{global_sync}(i) + \tau_{render}(i)\end{aligned} \tag{4}$$

Without any alteration of the simulation sequence, the kernel computation and the synchronization must be executed sequentially. However, waiting for the GPU to finalize kernel execution before continuing leaves the CPU idle for the entire duration of all GPU operations, thus the idea is to try to decouple these two phases in order to be able to execute them in parallel. This is possible since each one requires a different hardware resource (GPU and CPU), but cannot be achieved if all steps described in algorithm IV-B.1 are performed for the current simulation step, due to the fact that synchronization can only be performed after new particle positions have been computed. This can be avoided by modifying the sequence of operations in such a way as to allow kernel computation

Algorithm IV-B.1 parallel cloth simulation

partition mesh {distribute the initial model of the cloth between processing nodes}
loop
  pre-frame synchronization {synchronize viewing parameters}
  run kernel {run the OpenCL kernel in order to compute forces and generate new particle positions and velocities through velocity Verlet integration}
  sync common regions {communicate the positions of ghost points between processing nodes}
  global sync {communicate geometry data or raster data in order to render the entire model of the cloth on a single rendering node}
  compose the frame {render the cloth model and the other objects in the scene}
end loop

and nodes synchronization to be out of phase, i.e. compute positions for simulation step i, while performing global synchronization for step i-1. Obviously the synchronization of ghost points is tightly coupled with the kernel run and therefore cannot be executed out of phase and rendering can only be performed for the step which has been globally synchronized. This leaves us with the following execution time:

$$\tau_{sim_step}(i) = MAX(\tau_{kernel}(i), \tau_{global_sync}(i-1)) + \tau_{sync_common}(i) + \tau_{render}(i-1) \quad (5)$$

with a small added overhead for creating and joining threads.

As can be seen by observing equations 4 and 5, this hybrid parallel approach yields the greatest increase in performance when τ_{kernel} and τ_{global_sync} are of similar magnitude. Experimental performance measurements for this technique will be presented in Section VI.

V. Distributed Rendering

Distributed visualization is one key aspect of designing powerful graphics cluster based applications. Efficient transportation of render data or of visual images is needed in order to allow users to visualize and interact with a scene distributed among more processing nodes. The basic idea behind distributed visualization is the same one regardless of its implementation: distributing a single render job within a single frame across many computers in a network. Data must be sent over the network from the GPU equipped processing nodes to the visualization server to obtain the image on a single display device.

There are two main approaches to distributed rendering, as identified in [15], based on where the sort from object coordinates to screen coordinates occurs, which is fundamental whenever both geometry processing and rasterization are performed in parallel. When distributing sets of pixels in screen space among processing nodes, we have sort-first rendering and when distributing objects, we have sort-last.

Even though parallel physical simulation of cloth on graphics clusters is not entirely a distributed rendering problem, it greatly resembles one if we consider the analogy between the great volumes of data involved in typical parallel rendering and the high demanding real-time computation involved in simulations. Moreover, the rendering process strictly falls into the category of problems before mentioned, thus the concepts of sort-first and sort-last also apply in our case. The challenge is to apply these distributed rendering concepts to our problem, which combines traditional rendering with General Purpose GPU computation (GPGPU). Obviously a sort-first approach would suffer greatly in terms of load balancing, since we decided to split the computational process into equally weighted parts among nodes, thus such an approach would greatly unbalance the GPGPU process. A hybrid approach that would consider splitting primitives (particles in our particular case) but using primitive data for centralized visualization (i.e. the position of particles) would perform better in terms of load balancing, but would suffer from very poor scalability since the amount of data that has to be transmitted to the visualization node depends directly on the size of the model. Moreover, this approach would ignore the rendering capabilities of individual processing nodes, putting all the stress involved in this process on the single rendering server. This leaves us with a pure sort-last process, splitting primitives among nodes and using each node for both the GPGPU computation and rendering.

The network traffic generated by this pure sort-last approach does not directly depend on the resolution used for the cloth model, but it does depend on the resolution of the final raster image since it implies rendering locally on each node and centralizing and depth sorting the raster data on the visualization server. However, we must take into account that each node processes and renders a small portion of the model, thus relevant pixel data will only be generated for the small area of the scene containing the cloth partition assigned to each node, making it therefore highly inefficient to merge full frames. Instead, we experimented with merging only select rectangular regions from each node that fully enclose each cloth partition. This comes with the extra overhead introduced by determining the screen coordinates of each such rectangular area, starting from the scene coordinates of all particles, but fully compensates through decreasing network traffic and, implicitly, processing time for each frame. Moreover, since physical simulations need to be performed in real-time, this implies a frame rate greater than 30 frames per second (FPS) as universally accepted in the video game industry, and since the explicit numerical integration chosen for our application allows for relatively small time steps, it can be assumed that the variation of the coordinates of the areas enclosing the individual cloth patches are extremely small from one frame to another, thus making it unnecessary to compute them for each step, allowing a reduction of the associated overhead. Experimental performance measurements for this approach will be presented in Section VI.

VI. Experiments

The experiments were carried out on a graphics cluster consisting of four nodes with NVIDIA GeForce GTS 250 and NVIDIA GeForce 8800 GTS GPUs and Intel Core 2 Duo E8400 CPUs running on Debian 6.0.5 Squeeze, connected to a Gigabit network. The scenarios devised were aimed at evaluating the performance of the proposed solution taking into account factors such as the resolution of the model (i.e. number

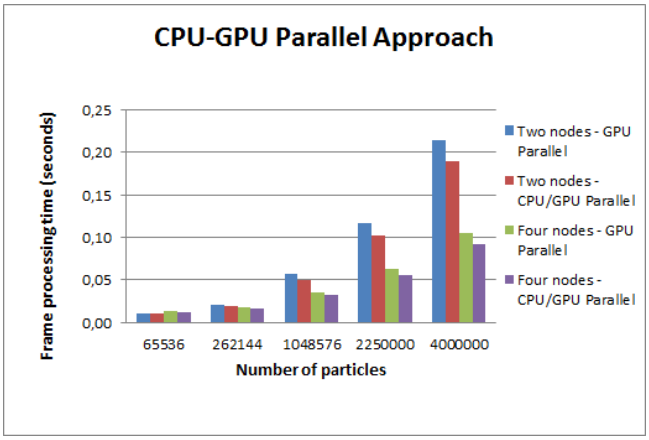


Fig. 1. Frame processing time for the hybrid CPU-GPU parallel approach compared to the simple GPU parallel approach

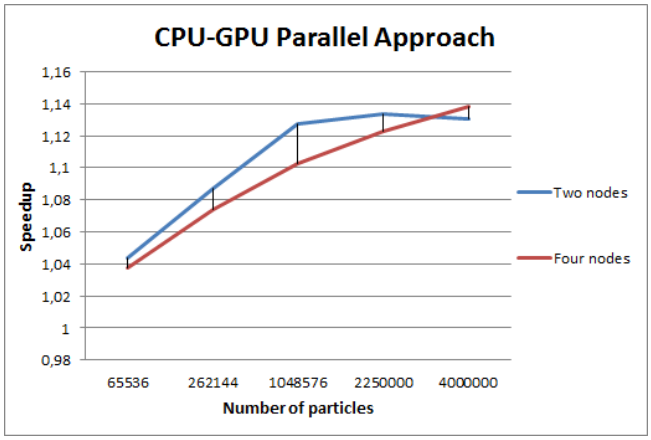


Fig. 3. Speedup in terms of model size for the hybrid CPU-GPU parallel approach

of particles) and the number of nodes used for the simulation. As criteria for measuring overall performance we chose the number of Frames per Second (FPS) which was considered relevant for a graphical application and a measure commonly used for evaluating parallel computation, the speedup. Each FPS sampling was performed after double the number of frames of the previously computed frame rate, with an initial threshold of 90 frames. To account for performance variations due to the load on the cluster, the final computed frame rate was the average of 10 successive samples over five consecutive runs. For evaluating the hybrid CPU-GPU parallel approach, we computed the average computing time for one frame over a total of 1000 frames. The test scenario considered for all experiments was the simulation of a piece of cloth colliding with other objects in the scene, in our case four spheres as shown in figure 2.

A. Hybrid CPU-GPU Parallel Approach

The aim of this experiment was to evaluate the increase in performance obtained by using the hybrid CPU-GPU parallel approach, compared to the simple GPU parallel one. Figure 1 shows the average measured processing time for generating a frame using both methods, when running the simulation on two and four nodes with regard to the resolution of the model (i.e. number of particles). As can be observed, the hybrid parallel method yields better results for all test cases. Moreover, the difference in performance becomes more significant as the model resolution increases due to the higher processing load on the GPU. Figure 3 shows the corresponding speedup for the different resolutions of the simulated model. The late occurrence of the switching point between the speedup on two and on four nodes is due to the difference in magnitude between the kernel processing time and the network transmission and frame merging time, as mentioned previously in the article. However, as the computation for the model becomes more demanding (ex. the model resolution increases, the simple numerical integration method is replaced with a more complex one, more collision detection, etc.), the difference between the two values will become less significant, thus better speedup on more nodes will be obtained earlier with regard to increasing model resolution.

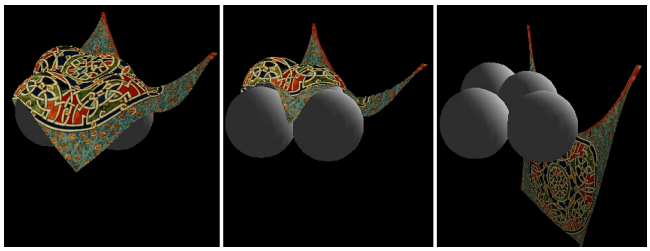


Fig. 2. Experimental scenario

B. Sort-last With Partial Frame Merge

The aim of this experiment was to evaluate the increase in performance obtained by using the sort-last rendering approach, with partial instead of full frame merging. Figure 5 shows the measured frame rates for both the original and the optimized method, with regard to model resolution (i.e. number of particles). The results prove that it is more efficient to obtain centralized visualization by using only those areas of each processing node's frame buffer which hold pixel data belonging to its own cloth patch than to have full frame buffer merging, despite the considerable computational overhead introduced by the need to continually recompute the coordinates of all those rectangular regions. Experiments have also shown that, for explicit numeric integration, which allows for small simulation steps, it is possible to skip this recomputing process for up to four frames without noticeable visual consequences.

C. Overall Performance

Figure 4 shows the overall performance improvement when using the hybrid CPU-GPU parallel approach in conjunction with the sort-last with partial frame rendering technique on a graphics cluster compared to running the application on a single node or compared to using the hybrid-sort alternative described in section V. As can be observed, the single node implementation offers better performance for relatively small model resolutions (65536 - 262144 particles) due to the fact that there is no need for synchronizing ghost points or transmitting data for obtaining centralized visualization. However, as the resolution for the cloth model increases

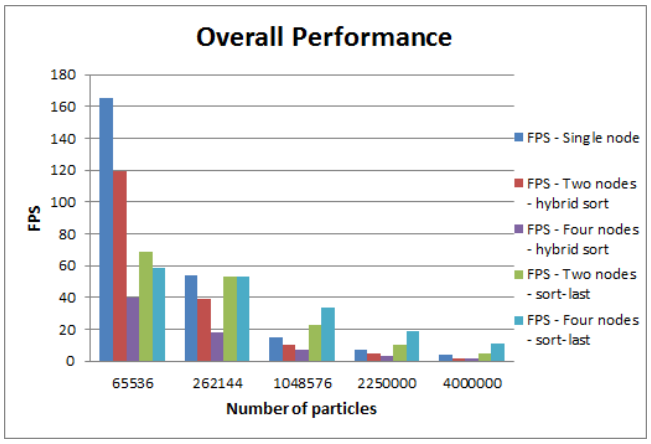


Fig. 4. Overall performance measurements

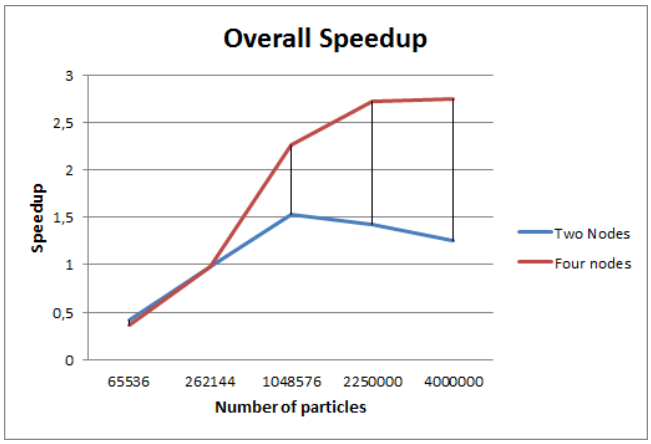


Fig. 6. Overall scalability

past 1m particles, the computation becomes overwhelming for a single node and the advantage of distributed physical simulation becomes obvious. The hybrid-sort alternative yields poorer results than even the single node implementation mainly due to the large amount of data regarding the position of particles that must be transmitted through the network and thus greatly underperforms the better, model size independent, sort-last approach. As shown in figure 6, the solution offers good scalability in terms of speedup obtained when running the simulation of four cluster nodes. As more nodes are used, maximum speedup is obtained for a larger number of particles, and a method for determining the optimal number of nodes for a certain model size may also be derived from empirical data, providing that there are a sufficient number of nodes available in the graphics cluster used for testing.

VII. Conclusion

In this work we have presented key techniques for simulating cloth on a graphics cluster. We focused on a hybrid CPU-GPU parallel approach and on a sort-last with partial frame merging technique for distributed rendering. Our performance measurements confirm the parallel efficiency of these methods and indicate that graphics clusters may be well suited for carrying out complex tasks such as real-time physical simulations.

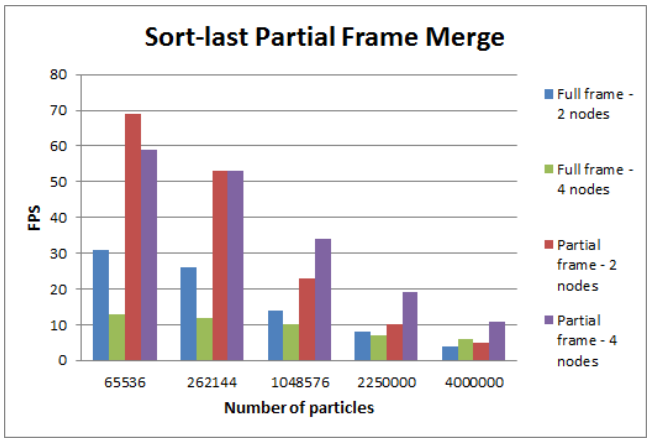


Fig. 5. Computed frame rates for sort-last with partial frame merge compared to full frame merging

Because scalability is encouraging, we would like to further explore these methods using more processing nodes, switch to implicit numeric integration and detect self collision.

References

[1] A. Sabou, C. Mocan, and D. Gorgan, "Particle based modelling and processing of high resolution and large textile surfaces," in *Intelligent Computer Communication and Processing (ICCP), 2012 IEEE International Conference on*, 30 2012-sept. 1 2012, pp. 355 –360.

[2] D. H. House and D. E. Breen, Eds., *Cloth modeling and animation*. Natick, MA, USA: A. K. Peters, Ltd., 2000.

[3] Z.-Y. Yuan, Y. Zhang, J. Zhao, Y. Ding, C. Long, L. Xiong, D. Zhang, and G. Liang, "Real-time simulation for 3d tissue deformation with cuda based gpu computing," *JCIT*, vol. 5, no. 4, pp. 109–119, 2010.

[4] A. Rasmusson, J. Mosegaard, and T. S. Sørensen, "Exploring parallel algorithms for volumetric mass-spring-damper models in cuda," in *Proceedings of the 4th international symposium on Biomedical Simulation*, ser. ISBMS '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 49–58.

[5] B. Thomaszewski, S. Pabst, and W. Blochinger, "Parallel techniques for physically based simulation on multi-core processor architectures," *Computers & Graphics*, vol. 32, no. 1, pp. 25 – 40, 2008.

[6] B. Thomaszewski and W. Blochinger, "Physically based simulation of cloth on distributed memory architectures," *Parallel Comput.*, vol. 33, no. 6, pp. 377–390, Jun. 2007.

[7] (2012) ParaView. [Online]. Available: http://www.paraview.org

[8] (2012) CEIs Ensight, engineering visualization software. [Online]. Available: http://www.ensight.com

[9] (2013) Equalizer: Parallel rendering. [Online]. Available: http://www.equalizergraphics.com/

[10] X. Provot, "Deformation constraints in a mass-spring model to describe rigid cloth behavior," in *IN GRAPHICS INTERFACE*, 1995, pp. 147–154.

[11] L. Verlet, "Computer "experiments" on classical fluids. I. Thermodynamical properties of lennard-jones molecules," *Phys. Rev.*, vol. 159, pp. 98–103, Jul 1967.

[12] (2013) OpenCL - The open standard for parallel programming of heterogeneous systems. [Online]. Available: http://www.khronos.org/opencl/

[13] (2013) CUDA Parallel Computing Platform. [Online]. Available: http://www.nvidia.com/object/cuda_home_new.html

[14] G. Karypis and V. Kumar, "Parallel multilevel k-way partitioning scheme for irregular graphs," in *Proceedings of the 1996 ACM/IEEE conference on Supercomputing (CDROM)*, ser. Supercomputing '96, 1996.

[15] S. Molnar, M. Cox, D. Ellsworth, and H. Fuchs, "A sorting classification of parallel rendering," *IEEE Comput. Graph. Appl.*, vol. 14, no. 4, pp. 23–32, Jul. 1994.

Road Line Detection based on Optical Correlator

T. Harasthy[1], J. Turán[2] and Ľ. Ovseník[3]
[1,2,3] Technical University of Košice, Department of Electronics and Multimedia Communications, Košice, Slovakia
tomas.harasthy@tuke.sk[1], jan.turan@tuke.sk[2], lubos.ovsenik@tuke.sk[3]

Abstract –Tracking of objects in video sequences, respectively in video streaming is necessary part of video surveillance systems. Tracking and detection of objects in video sequences is very important in automobile industry as well, concretely in video based Driver Assistance System. Processing of video in real time puts high demands to methods of processing of video stream. In this paper is presented detection of road line using optical processor. Video is captured with digital camera and preprocessed video sequence is analyzed by Optical Correlator. System presented here helps keep the car between the road lines and to prevent accidents caused by driver inattention.

I. INTRODUCTION

Video surveillance systems are nowadays very interested and requested topic [3,5,6]. Object recognition (faces, traffic signs etc.) is very helpful in many sectors of industry [3,5,6,7]. Video surveillance system is effective helper in security as well [4,5,8,9,10].

Usage of optical processors can improve Video surveillance system to objects recognition and make them real time system. Optical correlator is optical processor which can compare two images in speed of light. Match of two captured images is interpreted by correlation peak in correlation plane [1,2,7,11]. There are two types of optical correlators [1,2,7,8,9,10]:

1. Mach Filtered Correlator, use Fourier transformed reference image as a match filter. Match between reference and input images is interpreted by one correlation peak in place where match occur.
2. Joint Transform Correlator (JTC) compares reference and input images in same time. These two images are aligned each other on Spatial Light Modulator (SLM) and enter to correlation process together as one image. Match between these two input images is represented by pair of correlation peaks. Position of correlation peaks in correlation plane is relative represented by position of input and reference image in input plane of correlation process [1,2,7,9].

In this paper is described system to detection road lines with using optical correlator. Road is monitored with digital camera and captured road is compared with reference road line [1,2,7,10].

In Chapter 2 is described JTC which is used in Optical Correlator, (Optical Processor used in proposed system). Chapter 3 contains design of proposed system. Systems scheme is showed and their main parts are described. In this capture is described preprocessing of captured video by threshold as well. Experiments and results are presented in Chapter 4. Conclusions are covered in Chapter 5.

II. JOINT TRANSFORM CORRELATION

In JTC are input (captured video) – *s(x,y)* and reference image – *r(x,y)* aligned together in input plane of correlation process. Input plane is now Fourier Transformed using lens system and produce Joint Power Spectrum (JPS). JPS is displayed in (SLM) and preprocessed, (mostly thresholded).

Preprocessed JPS is captured by CMOS or CCD image sensor and then is Fourier Transformed again. Output of second Fourier Transformation contains autocorrelation of input plane, (input and reference image). This plane contains pair of correlation peaks per match [1,2,7,8,9,10].

Principle of Joint Transform Correlator is showed in Figure 1.

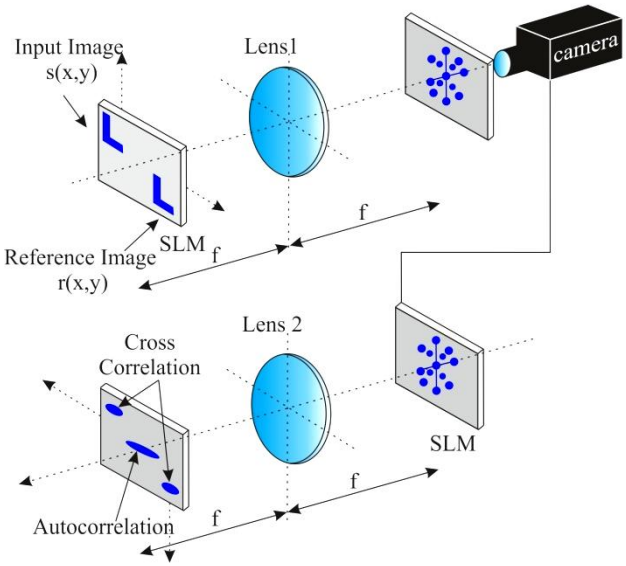


Figure 1. Captured and preprocessed scene of road line video

JTC compare input data in two criteria. Intensity of correlation peaks give information about measure of similarity between captured Road Line by camera and reference road line. Position of correlation peaks give information about relative position Input and reference data in input plane of correlation process. Advantages of this method are the simplified optical train (and therefore much cheaper to produce) and strict alignment criteria [1,2,7,9].

The Cambridge Correlator, used in our experiments is type of JTC, known as 1/f Phase Only Joint Transform Correlator. The 1/f term applies to the fact that a single

Fourier Transform stage is used for both stages of JTC process – i.e. the same optics are used twice to produce the correlation. The Phase-Only term relates to the images being displayed on the SLM in phase (i.e. +1/-1 rather than 1/0 in the case of a binary device) which significantly improves system performance [9].

III. Design of System

Road Line detection System was designed and main parts of system are showed in Figure 2.

a) The road is monitored by digital camera from car. Captured video sequence is preprocessed in Computer. In our system (and experiments) the CCD camera was used. Camera was situated on left side of car and was focused to middle road line.

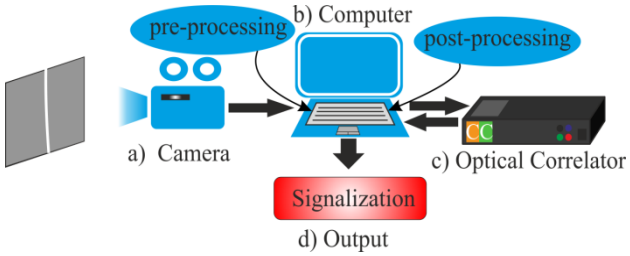


Figure 2. Proposed system to Road Line Detection

b) Captured road line is showed in Figure 3a and preprocessed video is in Figure 3b. To preprocessing was threshold technique used. This technique is a simplest way to minimize unwanted noise of background [4].

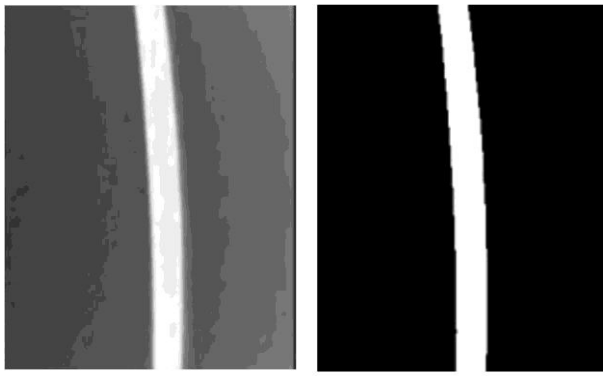


Figure 3. Captured and preprocessed scene of road line video

Another ways, how the road line could be filtered are Edge or Line detection [4].

c) Captured and preprocessed scenes from camera are then compared using optical correlator with database of road lines. Output of Optical Correlator is then post-processed in Computer to detect correlation peaks in correlation plane. In this case is most important describe how are correlation peaks moving in correlation plane in depend to moving car on the road.

d) If car is out of lines and any dangerous can occur, driver will be warn, or car can correct position by information from system. In this case, System just warns driver if any dangerous can occur (car is out of lane).

Whole proposed system use these four sub-blocks to complete detection Road Line and this system can very helpful in drive process. Experiments with proposed system are demonstrated in next capture.

IV. Experiments and Results

The experiment demonstrates system to keep car between road lines. In Figure 4 is showed input plane of correlation process. The output (correlation) plane is showed in Figure 4b. Correlation plane consist of three peaks. The biggest peak is situated in the middle of plane and it represents DC noise or autocorrelation of input scene. Two correlation peaks represent detected match or correlation between input images (captured and reference road line).

In this case the more relevant knowledge is that correlation peaks change their position in relation to moving captured line. Camera is focus on middle road line. Moving captured line in input plane (Figure 4.a) is represented by moving correlation peaks in correlation plane (Figure 4b).

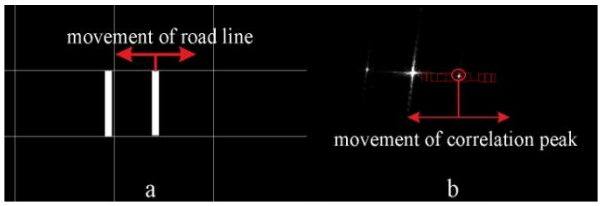


Figure 4. Demonstration of moving car (road line)

Demonstration influence of moving car on road and also moving road line in input scene is showed in Figure 5.

Position of correlation peaks in different Region of Interest (ROIs) gives information about moving car on road. In Figure 5a is a car situated nearest middle road line and correlation peaks are situated nearest the DC correlation peak. In Figure 5b and 5c is demonstrated move off the car from middle road line.

Correlation peaks in these cases are situated further away from middle peak. Moving of correlation peaks represent moving car on the road. As is showed in Figure 5, in correlation plane were chosen 10 Regions of Interest (ROIs) where would correlation occur. Number of ROIs can be different, number ten was chosen by experiments. More ROIs can help in better localization of car on road, e.g.: if correlation peaks is in 3rd ROI and next correlation peaks occur in 4th ROI, we know, that the car move to right (if we follow correlation peak on right side of correlation plane).

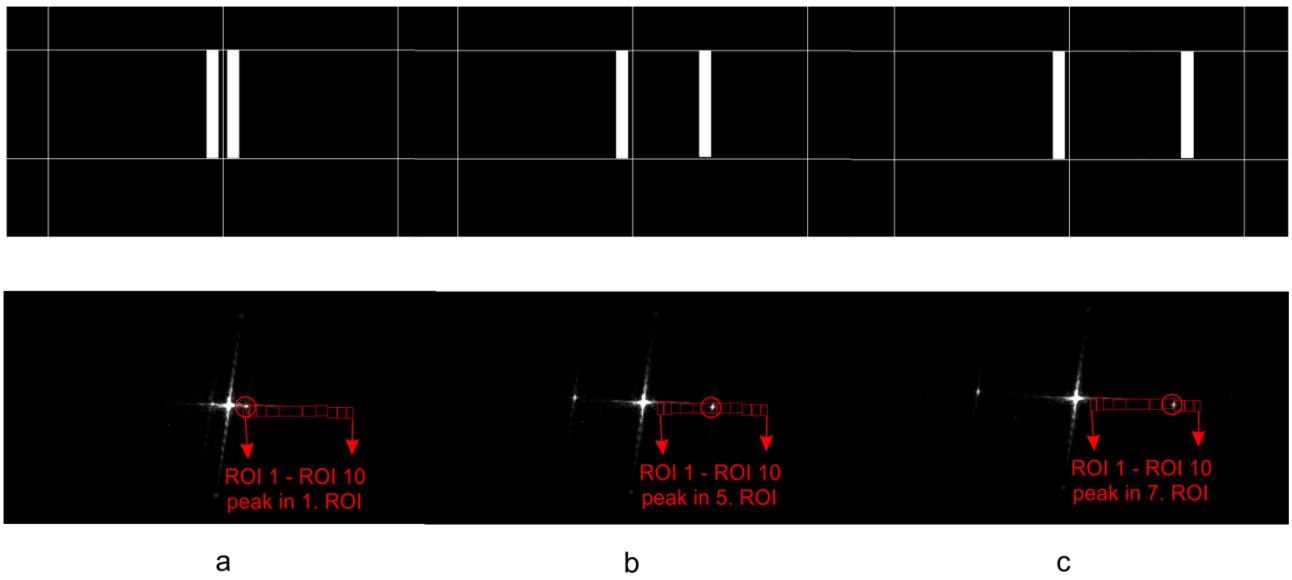


Figure 5. Principle of Joint Transform Correlator

For most simple situation, we should determine just one larger ROI. If correlation peak occur in this ROI, car is between road lines and it is safe, and if in ROI no correlation peak occur, car is out of line, and dangerous can occur. Using more than one ROI we can say exactly how is car moving in the road and then warn driver which way should turn the car.

V. CONLUSION

In this paper The Road Line Detection System based on optical Correlator is presented. System was designed with Cambridge Correlator, which is based Joint Transform. Traffic scene was captured by digital CCD camera and captured video was preprocessed. The threshold technique was used. Preprocessed video goes to optical correlator to compare with reference road line. Match of these inputs was represented by correlation peaks. Moving of these correlations peaks represent moving of captured Road Line in input scene depend to moving car on the road. Experiments were done. Proposed system track car on the road and can help driver in driving process. This System can warn to driver if car leave lane. To improve this system, we will track correlation peaks in correlation plane with more ROIs and then we can tell to driver where the car is located and turn the car back to correct lane. Proposed system is designed only to solid line, but can be modified to detecting dotted lines or both (dotted with solid lines). This improvement can be done using more ROIs in correlation plane and correct detection of correlation peaks.

ACKNOWLEDGMENT

We support research activities in Slovakia / Project is co-financed from EU funds. This paper was developed within the Project "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055.

REFERENCES

[1] J. W. Goodman, “Introduction to Fourier Optics,” Colorado: Roberts & Company, Publishers, 2005.

[2] Yu, F. T. S. and Yang, X.: “Introduction to Optical Engineering,” Cambridge University Press, 1997.

[3] F.P. Paulo and L.P. Correia, “Automatic detection and classification of traffic signs,” Image Analysis for Multimedia Interactive Services, WIAMIS, 6-8 June, 2007.

[4] S. E. Umbauch, “Computer Vision and Image Processing and A Practical Approach Using CVIPtools”, Prentice Hall PTR, Upper Saddle River, NJ 07458, June 1999.

[5] T. Harasthy, J. Turan, L. Ovsenik and K. Fazekas, “Traffic Signs Recognition with Using Optical Correlator”, Proc. IWSSIP 2011, Sarajevo, Bosnia and Herzegovina,16-18 June, 2011, 239-242, 2011.

[6] D. S. Kang, N. C. Griswold and N. Kehtarnavaz, “An Invariant Sign Recognition System Based on Sequential Color Processing and Geometrical Transformation,” Department of Electrical Engineering, Texas A&M University, College Station, Texas, 1994.

[7] P. Ambs, “Optical Computing: A 60-Year Adventure, ” Hindawi Publishing Corporation Advances in Optical Technologies, Volume 2010.

[8] L. Guibert, Y. Petillot and J.-L. de Bougrenet, “Real-time demonstration of an on-board nonlinear joint transform correlator system”, Society of Photo-Optical Instrumentations Engineers, Optical Engineering, Vol. 36 No. 3, Brest Cedex, France, March 1997.

[9] Cambridge Correlators, http://www.cambridgecorrelators.com, December 2012

[10] L. Guibert, G.Keryer, A. Servel, M. Attia, H. S. MacKenzie, P. Pellat-Finet, J.-L. de Bougrenet de la Tocnaye, “On-board optical joint transform correlator for real-time road sign recognition”, Optical Engineering, Vol. 34 No 1, France, January 1995.

[11] W. Steingartner, “The role of categorical structures in integral calculus”, International Scientific Conference on Computer Science and Engineering CSE 2012, Stará Lesná, High Tatras, Slovakia, October 3-5, 2012,. 24-30.

BIOMEDICAL ENGINEERING

The specific approach for establishment of Innovation Hospitals in Gorenjska region

Stanič Uroš *, Dolinšek Slavko **, Andreja Cerkvenik Škafar***, Irena Grmek-Košnik****, Jana Stanič*****, Gregor Cerinšek******

*Biomedical Research and Innovative Society, Ljubljana, Slovenia
**Innovation Development Institute University of Ljubljana, Ljubljana, Slovenia
***Gynaecology and Obstetrics Hospital Kranj, Kidričeva cesta 38a, 4000 Kranj, SI
****Institute of Public Health Kranj, Gosposvetska 12, 4000 Kranj, SI
*****Kosezi d.o.o., Iskra Techno, Ljubljana, Slovenia
*kosezi@siol.net

Abstract - The objective to convert hospitals in EU regions to innovative type is of high priority for their economic growth, employment and wellbeing of EU citizens. The results presented in the paper were obtained in Gorenjska region through the implementation of the InTraMed C2C Central Europe project. The system approach consisted of identification of the state of the art in innovation at each of the involved hospitals, SWOT analysis, seminars and workshops related to innovation process and culture. Among large spectrum of ideas gathered at these events the managers of hospitals agreed that first priority should be given to most needed biomedical engineering (BME) prototypes, which were supported also by small and medium enterprises (SMEs) business interest. Finally, through open innovation approach five ideas were selected for development of prototypes and later testing. Some of these selected BME ideas will be presented in the paper.

I. INTRODUCTION

The literature on innovation in hospitals has recently started to observe hospitals as providers of complex services and healthcare system hubs. In this approach, hospitals are regarded as combinative providers of diverse and dynamic services, able to go beyond their own institutional boundaries by becoming part of larger networks of healthcare provision, which are themselves diverse and dynamic (inter-hospital agreements on the shared use of equipment, community health networks set up to tackle a range of pathologies or even social problems). This approach makes it possible to extend the model of hospital innovation to incorporate new forms of innovation and new actors in the innovation process.

In this paper the results will be presented that were obtained through the implementation of the InTraMed C2C Central Europe project [1] during second and third (last) year. According to the project Action Plan the Innovation Seminars and Workshops were done, the final selection of the innovation ideas were selected and the introductory workshops with SMEs and research institutes for five ideas of the three participating hospitals took place.

II. MAIN CONCLUSIONS FROM THE INNOVATION SEMINARS AND WORKSHOPS

In the context of the InTraMed project, Institute for Innovation and Development (IRI UL) has organized and delivered 4 innovation seminars and 3 innovation workshops within 4 different health care institutions in Gorenjska, Slovenia, i.e. University Clinic of Respiratory and Allergic Diseases Golnik, General Hospital Jesenice, Institute of Public Health Kranj and Hospital for Gynecology and Obstetrics Kranj. The participants were the decision makers within each health care institution (i.e. top management and deputes, specialists, financial management, human resource management).

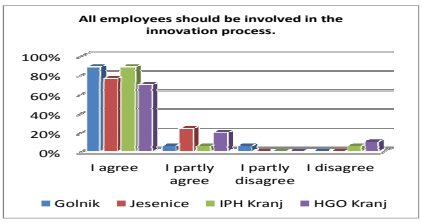
The main objectives of **the innovation seminar** were: 1) introduce the key characteristics of the innovation and innovation processes; 2) introduce the innovation model focusing on idea generation and idea filtering; 3) introduce the main mechanisms, tools and best practices to develop and maintain the innovation system within the health care institutions.

All participants of the innovation seminar received the "idea generation form" in order to describe several different ideas related to the identified problems that are emerging in the context of their daily work. All the forms were gathered and analyzed by the general director of the respective health care institution and the expert from the IRI UL which acted as the process facilitator. After this preliminary filtration of ideas, the authors of ideas were guided step by step through the innovation process (from idea to realization) in the context of the innovation workshop. Therefore the main objectives of **the innovation workshop were:** 1) general overview of the gathered ideas; 2) using the innovation model and idea management in practice, focusing on the gathered ideas;

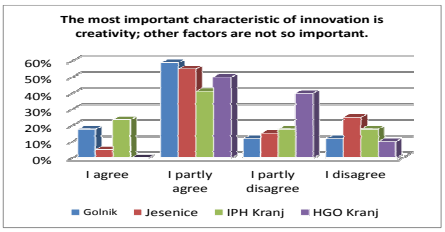
3) introducing the supportive mechanisms for idea development and commercialization.

At the very beginning of the innovation seminar, the participants received a questionnaire in order to identify their current understanding of innovation. Some main results are as follows:

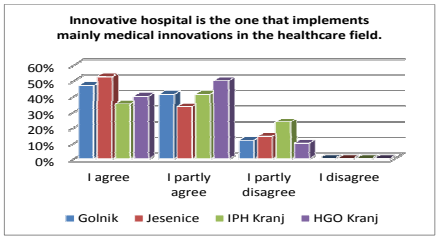
1) The majority of participants agreed that innovation is of crucial importance to enhance the quality of services and achieve the competitive adventage of the healt care instituion. The concept of innovation has a strong value for them and they furthermore agree that all employees should be involved in the innovation process.



2) Innovation as a concept was mainly understood as "somethin new, not including the added value yet". Furthermore, the participnats considered creativity as a key and main characteristic of the innovation process; other factors were not considered so much important when compared to creativity. It was explained to the participants that innovation is understood as a useful novelty, which brings new solution(s) to the identified problem(s) and therefore always has the added value. Innovation always involves creativity; however, it also involves successful ralization, implementation and commercialization of new ideas.

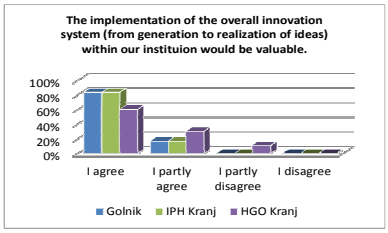


3) The majority of participants also considered innovative hospitals as those focusing mainly on medical innovation in the health care field. These innovations are crucial in order to enhance the quality of primary services provided. However, other (supportive) functions can also greatly contribute to better innovation performance of the helath care instituion.

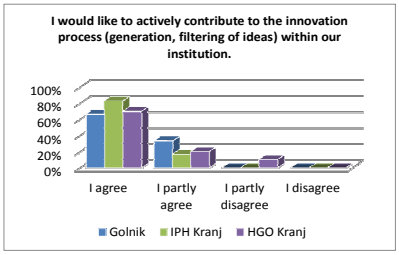


The innovation seminar was followed by the innovation workshop. At the end of the workshop participants received a new questionnaire in order to receive the feedback in terms of reaching the objectives set by the innovation seminar and innovation workshop. Some main results are as follows:

1) The majority of participants agreed that their institution would benefit from implementing the innovation system. They could benefit using the tools for generating, filtering, sharing and storing the ideas that could furthermore be ICT-enhanced.



2) Despite the high burdens and responsibilities of their day-to-day job, the participants would like to actively participate in the innovation process. This also gives a clear signal to the top management in terms of the necessary implementation of the innovation system and active encouragement of all employees to contribute to its realization.



III. METHODOLOGY

The methodology consists of analyses of the innovation state of the art in hospitals in Gorenjska, Slovenia and definition of innovative hospitals through 10 innovation enablers. The synthesis consists of Action plan for evolution from Non-innovative to Innovative hospitals and its implementation, what is presented in the paper [2].

A. IMPACT of innovation seminars and workshops on SWOT analysis

The objective of InTraMed project is to activate the innovation potentials of hospitals to become major actors in regional development, employment and wellbeing through transfer of knowledge to SMEs and Industry. In Gorenjska region 5 key health care actors were actively involved in the project: 4 hospitals (among them one university clinic) and National Institute of Public Health.

SWOT analysis considerably changed in favour of innovation:

Strengths:

- Good quality of services at primary, secondary and tertiary level
- Good coverage of services adapted also to specific needs of life and sports activities in Alps
- Awareness of innovation potentials
- Enhancement of competitiveness due to introduction of innovation culture
- Establishment of long-term cooperation with economy(SMEs and industry) and research

Weaknesses:

- Legal status of clinics and hospitals not adapted to the market economy,
- Unexploited private-public partnership
- Lack of the managers of clinics with innovation skills

Opportunities:

- Introduction of innovation culture in the management of clinics
- Transparent motivation scheme for innovative achievements
- Strengthening of long–term business cooperation with industry/ SMEs and academic sphere/ research
- Establishment of local Medical and Technological Technology Parks on the grounds of clinics
- Clinics could become the key actors for economic growth, employment and wellness of the region
- Establishment of BME Alliance in Gorenjska/ Slovenia

Threats:

- Absence of willingness of political actors to change the legal status of clinics adapted to the market economy and the need of innovation culture
- Resistance of staff to the introduction of innovation culture to clinics
- Faster growth of BME innovation in other EU regions

B. Impact of seminars and workshops on 10 innovation enablers

The positive effects of the implementation of the InTraMed C2C project might be assessed also through methodology of 10 innovation enablers originally developed for innovation level of SMEs:

Table 1: 10 Innovation enablers Changes due to the InTraMed C2C project

Innovation Enablers definitions	**Hospitals starting point**	**Hospitals today**	**Innovative hospitals (ideally)**
Strategy	xx	xxx	xxxxx
Structure & Organisation	xx	xx	xxxxx
Innovation Culture	x	xxx	xxxxx
Financial Resources	x	xx	xxxxx
Human Resources	xxx	xxxx	xxxxx
Information & Technology	xxxx	xxxx	xxxxx
Idea Generation & Creativity Process	xx	xxx	xxxxx
Implementation of Innovation	x	xxx	xxxxx
Market Orientation & Operation	x	xx	xxxxx
Exploitation of the innovation	x	x	xxxxx

C. Realisation of Action plan for evolution from Non-innovative to Innovative hospitals

TO DO list of actions to be performed in each of involved hospitals to build innovation organization:

Actions to be accomplished later:

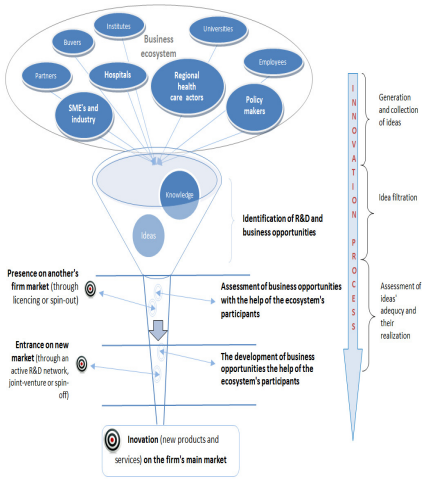
1. Amend the Hospital Long -term Strategy document with the Innovation articles
2. Amend the Hospital Statute with innovation elements
3. Establish The Innovation & Transfer office or at least nominate the responsible person for the management of innovation
4. Introduce Transparent Motivation scheme document
5. Test the present legislation environment for innovation friendliness

Actions accomplished during implementation of the InTraMed C2C project:

6. Prepare the templates for the submission of innovation ideas
7. Categorise innovation ideas
8. Identify main circles in categories of experts/employees
9. Prepare the list of SMEs and/or industrial partners
10. Prepare the list of academic (research, education) and business partners
11. Gather the most innovative and urgent ideas for further processing to innovation projects

It is evident that in order to raise the interest for innovation among the managers and doctors we first did the actions for which they were personally interested and involved. Realisation of the cited actions the Open Innovation approach was extensively used.

Figure 1: Gathering and selection of innovative ideas –



IV. RESULTS

The practical "learning by doing" open innovation approach [22, 23, 24] was the shortcut to introduction of innovative culture to hospitals in Gorenjska region. We focused on medical problems identified by doctors and selected by leading hospital staff. In total 5 ideas were selected for development through involvement of SMEs/industry and RTDI organisations:

IDEA NO 1: **The granny birthing rocking chair**

Humankind has existed for millennia and women have been giving birth for millennia. The positions in which they gave birth have changed over history, with developments in society and medicine. For decades, opinions have clashed regarding the advantages of the upright position (sitting up, using a birthing stool, being on all fours) versus lying down during labor [3].

In her book *Ina May's Guide to Childbirth*, Ina May Gaskin reports that in traditional societies all over the world women have given birth in an upright position. Seeking greater control over labor, greater comfort of the staff that supervised the labor, and the use of anesthetics during labor, women began to give birth lying in bed, on their backs. Gaskin points out that as early as 1882, in his book *Labour among Primitive People*, George J. Engelmann emphasized the importance of women's instincts and spontaneity in selecting the best position when giving birth [4].

Our maternity hospital, Hospital for Gynecology and Obstetrics Kranj, already has some experience with women giving birth sitting on a stool. This method is used in the second stage of labor, during the expulsion of the baby, but women are advised not to sit on the stool for too long. We identified the need to design an advanced birthing stool in order to further develop our doctrine, named the Granny Rocking Chair as a birthing chair. This would enable women to give birth in a position that is much more comfortable and natural than lying on their backs [5]; this is also what modern women increasingly demand. There is increasing evidence that, if women move around [5,6,7] and use a comfortable position in a homey environment supported by people they know well and trust [8], the first stage of labor (dilation of the cervix) is shorter and less painful due to more intense contractions. Having a bath or shower during labor [9], using steady movements, flexing muscles, dancing, using birthing balls, stools, squatting, and so on [10] are all methods that relax the woman, reduce her labor pain, and enable faster and easier labor with fewer injuries to the birth canal [4,11]. However, other studies indicate no statistically significant reduction of instrumental deliveries when giving birth in the upright position [12].
The second stage of labor (full dilation of the cervix to the birth of baby) on the Granny Rocking Chair takes place with the woman kneeling. This makes good use of gravity because it shortens labor compared to lying down—which is, however, more convenient for the medical staff helping with the delivery. Giving birth in an upright position (sitting, squatting, or standing) is more natural and physiological, and thus also easier and less dangerous for both the baby and the mother [5]; in addition, it is also an important factor that reduces birth canal injuries and the need for episiotomy [11]. The disadvantages reported include a greater loss of blood during labor—that is, more than 500 ml, but less than 1,000 ml [3, 12].

Experience shows that touching the mother's skin and nipples, gently massaging her, and her touching and looking at the baby's head also stimulate the mother to push her baby out more effectively [4]. A camera would be attached to the Granny Rocking Chair, so that the mother and her partner could watch the gradual crowning of the baby's head (and see their first glimpse of the baby's hair).

One of currently available aids enabling delivery in the upright position include the birthing stool is shown in Figure1.

Fig.1 Birthing stool

IDEA NO 2: **Functional electrical stimulation (FES) supported labor**

Clinical practice shows that many women in labor cannot or do not have enough strength to participate in activating the muscles that help expel the baby, and so midwives and obstetricians must help them by pressing against their abdominal muscles. The concept is based on the FES of the abdominal muscles, the pelvic floor muscles, or all the muscles that help expel the baby and enable easier and faster expulsion.

Description of the problem:

Towards the end of the second stage, when the head has already reached the pelvic floor and started to press against the rectum, during contractions the woman gets a "pressing" sensation down below and starts to push voluntarily. The uterine contractions are thus accompanied by contractions of the mother's abdominal muscles. Nature uses these two forces to complete the delivery as soon as possible. The innovation idea is to support the cited natural forces by FES of abdominal muscles FESAM. The FESAM methodology was first developed and applied to support ventilation in SCI and tetraplegic patients [13, 14].

IDEA NO 3: **Machine for specimen management in clinical microbiology**

Nothing is more important to the effectiveness of the laboratory than a specimen hat has been appropriately selected, collected, and transported. If specimen collection and management are not priorities, the laboratory can contribute little or nothing to patient care or related investigations. It stand to reason, then, that if laboratory data are used to supply critical information that either confirms or leads to a diagnosis and successful treatment, those involved in selecting, collecting, and transporting these microbiology specimens must understand the needs of the laboratory regarding the specimens. In addition, laboratories must know the needs of the physician in order to direct their technical efforts toward providing results that are accurate, significant, and clinically relevant.

Specimens for microbiology analysis are likely to contain living organisms whose recognition depends on rapid, professional specimen management. Understanding this simple concept should motivate specimen collectors to select the correct anatomic site from which to obtain the specimens, collect the specimen using the proper technique and supplies, package the specimens in a container designed to promote survival of the causative organism and to eliminate leakage, which might pose a safety hazard and transport the specimen to the laboratory expeditiously or make sure that if it is stored, storage is brief, properly done, and at a temperature that will not damage the suspected organism [15].

Specimens for microbiology analysis should be promptly delivered to the laboratory for accessioning and plating. Everyone involved should remember that microorganisms are living things. They multiply and die very rapidly. The microbiology division is different from other sections of the laboratory. Maintenance and manipulation of living organisms require processes very different from those used in chemistry and hematology [16]. Our innovation is a machine for providing the proper supplies and collection of diagnostics specimens. It allows 24-hour service, guaranteed data protection, discretion of personal data. Due to the dispersion of health care institutions full coverage of all the provinces, accessibility to the Public health institutions, nursing homes, health centers and hospitals is required. There is great opportunity to export Bioavtomat especially in the EU region, high added value, new jobs.

IDEA NO 4: **Humanoid robot for vaccination**

In the event of a pandemic there is a major problem on how to vaccinate big capacity of people in a very limited time, so a continuous operation is required to guaranteed the same high quality of vaccination. It is impossible for vaccination team to fulfill that requirement [17,18]. Humanoid robot would allow more of a rational packaging and the using of the vaccines. Vaccination is the most efficent achievement of modern medicine. Vaccination is compulsory in Slovenia, starting at the age of three months old. Over the course of life people receive many doses of vaccine. In the event of outbreaks of communicable diseases, immunization is often a measure to prevent the spread of diseases such as individual outbreaks of measles, diphtheria, as well as in cases of terrorism. Vaccination is also important in travel clinics and deprived areas such as Gorenjska and Carinthia for tick borne encephalitis [19]. The advantages of our inovation are: High capacity vaccination coverage, repeatable high quality and quick response time in the event of a pandemic or other disaster. Humanoid robot will have a sensor manipulation and positioning of the extremities-hands, cooling and disinfecting the injection site, injection of the vaccine dose and installing the patch. Robot should allow traceability services to the telecommunications link with patient records and future national vaccination program and E-Health. Humanoid robot for vaccination, according to the application and

testing in Slovenia represent a great potential for global marketing especially in the developing countries of Africa and others developing countries. High added value, high income and new employment.

IDEA NO 5: **EKG based system for 24 hours monitoring of life function**

University Clinic of Respiratory and Allergic Diseases Golnik, University Clinical Centre Ljubljana, General Hospital Celje and many other hospitals in Slovenia identified the need for continuous monitoring of vital functions: electrocardiogram (ECG), respiratory rate (RR), transcutaneous oxygen content in the peripheral blood (satO2), heart rate (HR), blood pressure (BP) and body temperature (TT). With close monitoring of these parameters, patients can be closely monitored and according to dynamic changes necessary medical assistance can be provided.
The objective of the innovation idea is development of pilot EKG based monitoring system that could be implemented in the health centers all over Slovenia [20, 21].

Short summary:

Penetration of innovation culture through the implementation of InTraMed C2C project at chosen hospitals in Gorenjska, Slovenia was fast and caused strong interest among the leading doctors, directors and nurses. To achieve invaluable good results, the experienced team of BME external consultants was engaged as moderators. Next step will be dedicated to the preparation of research and development project proposals as the engaged SMEs are not fully able to financially cover the development costs. In hospitals the innovation organization should be established that will be capable to manage any innovation ideas proposed by their staff. The good practice of "Innovation hospitals" is planned to be disseminated to other 29 hospitals in Slovenia and also to hospitals in the neighboring countries and other EU regions.

ACKNOWLEDGMENT

The research presented in the paper was co-financed by the InTraMed – C2C project which is implemented through the CENTRAL EUROPE Programme co-financed by ERDF. The authors express their appreciation to Franc Preložnik, Jože Jelenc and Franc Setnikar for their valuable engineering, design and marketing contributions as well as Helena Cvenkel, Nives Justin and Bogo Filipič for their technical support.

REFERENCES

[1] InTraMed – C2C project, CENTRAL EUROPE Programme co-financed by ERDF

[2] Activation of Innovation Potentials of Medical Staff in Design of New Telemedical Applications in Slovene Hospitals U. Stanič, F. Strle, J. Stanič, S. Dolinšek, MIPRO2012 Invited paper

[3] Gupta LK, Hofmeyr GJ, Shehmar M. Position in the second stage of labour for women without epidural anaesthesia. Cochrane Database Syst Rev. 2012 May 16;5.

[4] Gaskin IM. Modrost rojevanja. Sanje, Ljubljana, 2007.

[5] Pajntar M, Novak Antolič Ž. Nosečnost in vodenje poroda. Cankarjeva založba, Ljubljana, 1994.

[6] McCarthy A, Hunter B. Obstetrics and Gynaecology. Churchill Livingstone, London, 1998.

[7] Lawrence A, et al. Maternal positions and mobility during first stage labour. Cochrane Database Syst Rev, 2009 Apr 15;(2).

[8] Schuiling KD, Sampselle CM. Comfort in labor and midwifery art. Imag J Nurs Sch. 1999;31(1):77–81.

[9] Stark MA. Therapeutic showering in labor. Clin Nurs Res. 2013 Jan 15.

[10] Sutton J, Scott P. Die Optimierung der Kindslage. Hippokrates, Stuttgart, 2010.

[11] Da Silva FM, et al. Risk factors for birth-related perineal trauma: A cross-sectional study in a birth centre. J Clin Nurs. 2012 Aug;21(15–16):2209–18.

[12] Thies-Lagergren L, et al. No reduction in instrumental vaginal births and no increased risk for adverse perineal outcome in nulliparous women giving birth on a birth seat: Results of a Swedish randomized controlled trial. BMC Pregnancy Childbirth. 2011 Mar 24;11:22.

[13] STANIČ U, Lenart L, Šorli J, Kandare F, Jaeger R. Augmentation of pulmonary ventilation by electrical stimulation of a abdominal muscles. V: Neuroprosthetics from basic research to clinical application / A. Pedotti et al. (eds.). Berlin [etc.]: Springer, cop. 1996. Str. 635-638.

[14] Kandare F, Exner G, STANIČ U, Jaeger RC. Functional electrical stimulation of respiratory muscles in spinal cord injury. V: Mechanics of breathing / A. Aliverti et al. Milano: Springer, cop. 2002. Str. 291-313.

[15] Miller JM. A guide to specimen management in Clinical microbiology. ASM 1999. Washington 2 and ed.

[16] Murray PR. Baron EJ, Pfaller MA, Tenover FC, Yolken RH. Manual of clinical microbiology. 7ed. Washington: ASM; 1999.

[17] Rizvi SZ, Yousuf S, Kanwal N, Akbar S. Automation of Vaccination and Blood Sampling Procedures. In : Engineering Sciences and Technology, 2005. SCONES Conference publications.: Available from http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4382894&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4382894

[18] Yousuf S, Kamran R. Emerging Technologies, 2007. ICET 2007. International Conference. Conference publications.: Available from http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4516338&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4516338

[19] Grmek Kosnik I, Krt Lah A. A campaign to increase the vaccination rate in highly endemic tick-borne encephalitis region of Slovenia. Vaccine 2013;31:732-4.

[20] Telehealth :A Myth or Reality?, R.Trobec and U.Stanič,Proceedings 34.Conf. Mipro 2011/GVS pp.326-331, May 23-27.2011, Opatija, Croatia, ISBN 978-953-233-060-1

[21] R. Trobec and I. Tomašić, "Synthesis of the 12-lead electrocardiogram from differential leads", IEEE Trans. Inf. Technol. Biomed. 15, 615-621, 2011.

[22] Bigliardi, B., Dormio, A. I., & Galati , F. (2012). The adoption of open innovation within the telecommunication industry. European Journal of Innovation Management, 27-54.

[23] Chesbrough, H. W. (2003). Open innovation. The imperative for creating and profiting from technology. Boston: Harvard Business Press.

Multifunctional protective clothing system: Development vs. Functionality

D. Z. Pavlinic*, A. Oder**, V. Grm***, U. Stanič*
* BRIS – Biomedical Research Innovative Society, Ljubljana, Slovenia, dzpavlinic@gmail.com
** Prevent Deloza d.o.o., Celje, Slovenia
*** Etra d.o.o., Celje, Slovenia

Abstract - Development of personal protective equipment (PPE) requires the knowledge of fire fighter's working environment and working activities. The PPEs function is the protection of the user from danger in the environment. Additional protection, monitoring of vital life functions and of surrounding environment could be added afterwards. Besides above mentioned functions the PPE clothing system must be functional, e.g. easy to use and care for, and above all it should not limit the range of movement of the fire fighter. The primary function is fulfilled by the choosing of optimal protection materials, because non compatible materials with non-optimal thermal balance can impact the fire fighter's working performance. Secondary functions are presented in the paper in the form of guidelines for development of fire fighter's PPE clothing system. The issues of PPE development problems in connection with the functions of cooling and multi-layering are presented in the paper. The presented engineering approach takes into account the requirements of fire fighters, who are exposed to the extreme environmental working conditions and who risk their lives daily.

I. INTRODUCTION

During the summer time people like to spend their free time in the outdoors –- camping, hiking, having –picnics and parties. Because of this the natural environment is exposed to increased fire danger than usual. Even small fires or flame particles that can spread into the natural surroundings can cause a huge fire. The vegetation, speed and wind direction, terrain, as well the air temperature and relative humidity impact the fire ignition. Additionally the fire spread can -be faster more than 100 km/h in very short time. While burning of organic substances different noxious gasses get released into the air and they add to the air pollution. Because of the destruction of the first earth layer the vegetation could be changed, the erosion could be increased. A lot of animal species could be endangered too. The most frequently causes for fire ignition in the outdoor environment are: the use of open fire (barbecue, camp fire, to kindle a candle flame and flambeau, rocket, petard,…), coal firing and burning, to throwing away cigarette ends, depositing of embers and burning substances. More than 95% of all wild land fires are caused by humans due to their carelessness and/or ignorance. In 2011 there were 2,990 wild land fires in Slovenia and 19,245 professional and volunteer firemen participated at rescue activities [1].

Developed personal protective equipment (PPE) available on the market is manufactured from different materials. PPE is used to protect people against hazards in the environment and usually has a limited life time. An example of PPE is protective clothing, which can be designed to protect against rain, temperature, explosions, chemical or biological agents, etc. Its protective performance will decrease due to environmental exposure. Therefore it would be very useful to incorporate sensors inside the clothing, to monitor whether the protective performance is still adequate. Current PPE is designed without any additional monitoring and controlling functions, or without any cooling ability. Furthermore the sustainability of protective materials while exposed to heat and flame several times is not known. Consecutively the resistance to mechanical loads of these flame damaged textile materials has to be investigated as well. Since high protection is a key factor in the quality of firefighter lifes, optimal maintenance of PPE is essential. Therefore it is obvious that the protection and safety of firemen at work are key determinants of the European directive 89/686/CEE [2].

Combining the different protective textile materials with sensor matrices into PPE can cause a problem in view of their behavior during handling and care. Integration of sensor matrices into textile layers can omit the firemen movements if the positioning of the matrices inside the garment would not be optimal. Thus, knowledge concerning the coupling of sensors into textiles is envisaged to have important implications for the firemen society and for manufactures of PPE as well. A growing number of firemen suffer from heart attack (44% in USA) due to firemen interventions. The findings of these studies (U.S.Fire Administration, 2007) show the need to improve the safety of firefighters and reduce the risks they take to save lives [3]. Understanding what causes the heart attacks will help to understand, predict and prevent them from happening as often.

II. DEVELOPMENT OF THE PERSONAL PROTECTIVE EQUIPMENT

Development of PPE is vital for human safety at work. PPE protect humans from environmental impacts, however its costs incurred by the manufacturer of protective materials are often a significant element of their final product cost. As PPE prices are rising across the world, the drive to reduce the costs and remain

competitive is stronger than ever. The cost of the PPE is passed to the Fire Association, who usually buys the PPE according their functional abilities. Protection problems are experienced by the firemen, and for some protective materials, especially when combined into sandwiches, this is problematic, both in terms of the product costs and in terms of the quality of protection.

Over the past ten years the need to adequately protect the humans in various dangerous situations has forced the development of clothing industry capable of providing highly innovative products, in some cases equipped with sensors or so called smart technologies. Beside basic functions of protection (including the thermal and ergonomic comfort), this new generation of protective garments have served the functions like monitoring and/or protection, both only partially accepted by the end users (firemen).

From the firefighting activities point of view there are different protective materials on the market. For wild land firefighting one layer garments are suitable made of PBO (Poly (p-phenylene-2,6-benzobisoxazole) is a rigid-rod isotropic crystal polymer) and of aramide materials that differ regarding to the protection quality (high or standard). PBO materials have been developed by Japanese company Toyobo [4]. It has superior tensile strength and modulus to Aramid fibers, such as Kevlar, Technora and Twaron. It also has outstanding high flame resistance and thermal stability among organic fibers. PBO fiber shows excellent performance, in such properties as creep, chemical resistance, cut/abrasion resistance, and high temperature abrasion resistance - far exceeding other Aramid fibers. PBO fiber's moisture regain is low (0.6%) and it is dimensionally stable against humidity. PBO fiber is quite flexible and has very soft touch, in spite of its extremely high mechanical properties. For multilayer garments different protective materials are in use under different commercial names. For example the material named XENIOS, developed and produced by partner Europrotect, stands for high quality protection PBO in extreme environment. Competitive materials come from anEuropean company TenCate, USA and Japanese (Toyobo) companies. From the above statement it is obvious that only 2 manufacturers in Europe produce protective materials.

The PBO material developed by Europrotect has been developed to improve weak points of the blend PBI/Para-aramide (Polybenzimidazole=PBI fiber is a synthetic fiber with a very high melting point that also does not readily ignite, because of its exceptional thermal and chemical stability) and therefore offer best solution underlining the advantages of PBO fiber. Twin system's main advantage, as patented fabric of Europrotect, is that the para-aramid yarn had been weaved on the back side allowing to reach better light fastness according to the ISO 105-B02 as well as improve tensile and tear strength.

A. Multilayer protective garment

The middle layer of multilayer protective garment is composed of membranes of different types. The role of the membranes in multilayered garment is to regulate the body temperature and enable the evaporation of excessive perspiration caused by heat. Such membranes provide the most effective and efficient protection from fire and heat. They are also waterproof and watertight. Like the outer shell, they must fully comply with the requirements of the EN 469 standard and must be certified.

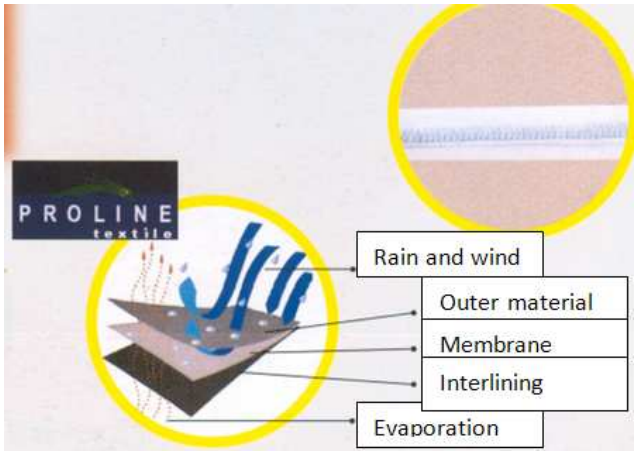


Figure 1 Multi-Layers of protective garment

The outer material is made of 100% NOMEX with Kevlar and anti-static fibers. This material is made according to the following standards: EN469 (firefighter clothing), EN 531 (flame retardant clothing) and EN 1149-1 (antistatic clothing). The surface of the outer layer is finished by water repellent aperture.

The second middle layer is made of PROLINE membrane; this is waterproof membrane made of Nomex and of polyurethane porous film. The membrane is made in accordance with the standard EN 343.

The third inner layer has an important role. It is made of blend of Nomex and Viscose FR fabric, with Kevlar threads in the warp direction. Using the Kevlar threads the fabric become microenvironments for air movement. This is important in order to avoid the accumulation of heated air and humidity in the garment microenvironment.

Protective garments are produced from the above described outer shell materials combined with a middle layer of membranes and with a lining, which –should be light and makes clothing extremely comfortable to wear. Several studies were performed regarding the investigation of particular protective materials and their compatibility in finished garments. Important guidelines are combined in the publications of Horvat & Regent [5] and in the handbook of Goldman [6]. Different directions on the field of the PPE development for firefighters were defined by Trovi and Abbott [7, 8]. Very important role in the development of optimal clothing ensembles has been given to Umbach with his 5-stage model [9].

Compatibility of textile materials from clothing extensibility point of view is important for all textile layers – garments that are combined into one clothing ensemble. The adding of clothing layers one on top of each other is a problem in the development process. A single layer cannot be developed on its own because of the size and also because of rapid changes in thermal

insulation over the entire body. The upper layer should allow enough space for the internal layers of garment, but it may not be too large when worn separately.

Outside layers demand larger extensibility in extreme positions, but this problem was considered from construction point of view during the development process. However the extensibility as the basic mechanical property of fabrics could be measured using expensive measuring systems or predicted using numerical models which describes the textiles by nonlinear behavior when exposed to mechanical loading. The relationship between the geometric structure and mechanical properties of fabrics has always received much attention since the early days of textile engineering science.

In order to obtain the optimal combinations (sandwiches) of protective materials to ensure the thermal comfort of the protective garment scientists have developed human thermal models to provide a theoretical framework for understanding human responses to work under various environmental conditions [10, 11]. The human thermal model provides powerful information for establishment of thermal and clothing comfort. PPE should be carefully designed to allow heat transfer from a wearer to the environment at an appropriate rate during various work intensities over a range of environmental conditions [12, 13, 14, 15]. Properties of a garment can be measured in the laboratory using heated manikins, but those values are usually not sufficient to define how well the garment will perform under field conditions. Mathematical human thermal models help to bridge the gap between laboratory studies and field performance, although their use only reduces the requirement for field trials; it does not completely eliminate the requirement.

B. Cooling layer in the protective garment

Climatologists worldwide are predicting a possible 1.8 to 2.3 °C rise in temperatures by the end of this century. The health consequences will be dramatic. According to a study of the Kiel Institute for World Economics (IfW) and the World Wide Fund for Nature (WWF), the number of heat deaths would increase by 5,000 to 15,000 per year if adaptation measures are not taken. Especially firemen and security people suffer from heat in their working environment.

The reason: heat causes enormous stress on the human body and can lead to death. The cardiovascular system is primarily affected; even worse, when wearing protective garments such as flame retardant clothing. In studies it is found that firemen lose up to 1 liter of water in 25 min, they often have a body temperature of 38-39°C and the maximum heart rate is over the recommended level for almost the whole duration of the firefighting operation. The immediate provision of direct, suitable for daily use and quickly effective protective measures is necessary.

E.Cooline products (men vest) are made from functional textiles that help regulate and cool the body at high temperatures, Figure 2. The cooling effect is a result of the evaporative cooling of water. To accomplish this, the product does not need an external source of energy. Thus E.Cooline cools the body in hot environments and contributes to the protection of the body.

The secret behind E.Cooline is a functional piece of clothing that is ready to use within seconds through the use of water and afterwards, delivers an immediate and long-lasting cooling effect. The body temperature is optimally regulated through evaporative cooling that develops at high temperatures, Figure 3. The cooling textiles can significantly mitigate decrease in concentration and performance and the associated endangerment to health and safety due to heat stress. The exterior of the product is dry and hence, can be simply worn under or over the usual clothing (under shirt or under functional T shirt) [12].

Figure 2 E.Cooline vest and T-shirt [12]

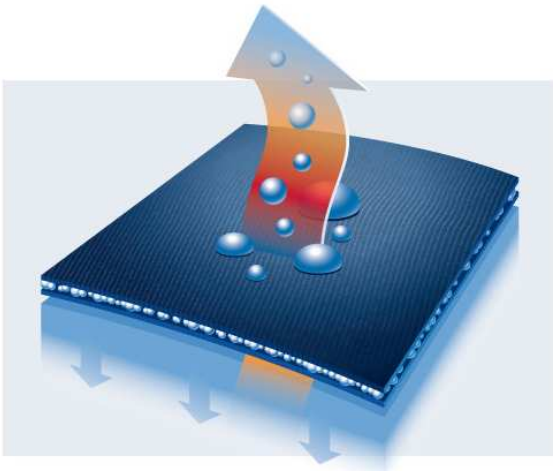


Figure 3 E.Cooline sweat management system [12]

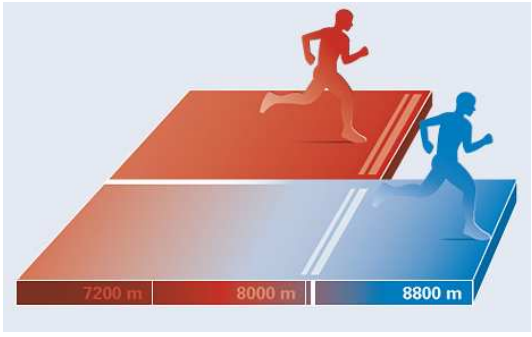


Figure 4 Human performances while wearing E.Cooline vest [12]

E.Cooline is a system based on innovative functional textiles, which stabilizes and checks the body temperature to save body energy and health. The system works in an active way even at high temperatures for up to 20 hours. E.Cooline system works always to prevent health and

performance as shown in Figure 4. The cooling using E.Cooline vest increase human performance for more than 10% [16]. These results were obtained over a wide study of E.Cooline products using human testing [16].

Cooling products have important role in PPE for enabling the optimal thermal balance. The workers who perform high intensive work in the extreme hot environment should use the optimal PPE from thermal balance point of view to eliminate heat stress.

C. Sensory layer in the protective garment

One of the research focuses in the development of innovative PPE is to embed them with the materials allowing them to cool and simultaneously to protect the humans in dangerous situations. Advanced PPE are embedded with sensors allowing them to monitor the vital life functions [17 – 22], but mostly such products have no written care procedure and even they are still not found on the market.

In the frame of research, that was performed by the authors some preliminary results were obtained. The aim was to combine the temperature sensors with flame retardant textiles and to take into account the care protocols. Preliminary results of the integrated sensors into textile layer are presented on the Figure 5.

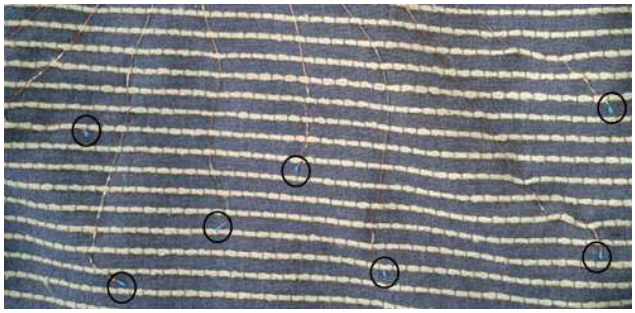


Figure 5 Sensor net on the flame retardant (Twin Spacer from Europrotect France) fabric layer

This has led to the development of smart textiles and textile materials with different functions: heating, cooling, antibacterial, with sweat management ability, etc. as well as some “sensorial garments” for different use. The ease of handling, robustness and comfort during wearing as well as the care of such garments has been largely ignored, although this was one of the most important requirements provided by firemen (end users). This statement was defined after several interviews made with Slovenian Firemen, both professionals and volunteers, as well by the Slovenian Firefighter Association.

III. DISCUSSION

Development of PPE has for the most part been dismissed as a sensory modality, that would not enhance monitor performance, neither would it add any significant operational utility to the protection systems. Since overheating and monitoring of vital life functions are the prevailing problem in PPE, the regulation of the temperature within the microenvironment of the PPE is achieved by different textile protection isolative materials with integrated breathable membranes. Some of the membranes on the market have limited breathability when high levels of sweat are occurred. However the monitoring of vital life functions is still not recognized and presented to the market as robust and flexible enough while exposed to the extreme working environments. The resistance of protective materials over real time situations was not measured yet, neither was the care protocol established in the way the firemen want it to be.

The impact from the environment (heat, mechanical loads, water, dust, oil, etc…) on the outer layer of protective garments must be investigated with the aim to develop a new care protocol of PPE that would be accepted by firemen. Only the development of such a complex system for monitoring and protection within PPE based on the sensor matrices and on the powerful software tools would serve as an appropriate approach. However, equipping PPE with the ability to monitor vital life functions and protect individual fireman, would be a significant milestone in efforts to develop PPE that would find application in many areas of human endeavor, especially on the field of firefighting and rescuing. Further areas of applications are in medicine, rehabilitation and care of elderly people.

IV. CONCLUSION

The impetus for the development of innovative PPE may be attributed to the recent trends in textile and clothing industry, personal protection, safety regulation, sensor techniques, communications and biomedical engineering. Which require a PPE incorporating the sensors for several practical applications. A common denominator in all the areas listed above is ensuring clothing comfort. Comfort is a complex factor, and comprises thermal comfort, ergonomic comfort, and mechanical comfort. The PPE developed in the future should focus on the problem of all three comfort factors

ACKNOWLEDGEMENT

The research presented in this paper is a part of the research project "Innovative modular system for monitoring and protection of rescue personnel in extreme environments" which is financially supported by the companies: Prevent Deloza d.o.o. and Etra d.o.o., both from Slovenia; Pervormance International GmbH Germany and Europrotect France SA.

REFERENCES

[1] http://www.stat.si/vodic_oglej.asp?ID=511&PodrocjeID=27. In Slovene.

[2] European directive on Personal Protective Equipment (PPE) 89/686/CEE, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1989L0686:20031120:EN:PDF, accessed Nov. 9. 2012.

[3] U.S. Fire Administration, Fire fighter Fatalities in the United States in 2007, June 2008.

[4] http://www.toyobo-global.com.

[5] Horvat J., Regent A.: Osobna zaštitna oprema, Veleučilište u Rijeci, 2009, ISBN 978-953-6911-43-1.
[6] Goldman, R.: Hanbook on clothing: Biomedical effects on military clothing and equipment system, 2007; Research Study Group 7 on Bio-Medical Research Aspects of Military Protective Clothing.
[7] Trovi, D.A. & Hadjisophocleous G.V.: Research in Protective Clothing for Firefighters: State of the Art and Future Directions, Fire Technology, 35 (1999) 2, 111-130, ISSN 0015-2684
[8] Abbott, N. J. & Schulman, S.: Protection from Fire: Nonflammable Clothing – A Review, Fire Technology, 35 (1999) 2, 205-218, ISSN 0015- 2684.
[9] Umbach, K.H.: "Physiological tests and evaluation models for the optimisation of the performance of protective clothing", In: Environmental Ergonomics. Sustaining Human Performance in Harsh Env. Philadelphia, 1987, pp. 139-161.
[10] Wissler E.H. (1985), Mathematical simulation of human thermal behaviour using whole body models. In: Shitzer A and Eberhart RC (ed) Heat transfer in medicine and biology - analysis and applications, Plenum Press, New York and London, pp. 325-373.
[11] Wissler, E.H. (2003). Probability of survival during accidental immersion in cold water. Aviation Space Environmental Medicine Vol. 74: pp. 47–55.
[12] C. Hertleer et al.: A textile antenna for fire fighter garments, AUTEX Conference, Portorož, Slovenia 2007.
[13] Mendelson Y.: Wearable Wireless Pulse Oximetry for Physiological Monitoring, Department of Biomedical Engineering Worcester Polytechnic Institute, Workshop 2008,USA.
[14] P. Brasser, M.P. Sobera: "Optimizing Balancing protection and physiological burden of CBRN-protective clothing", acc. for publ. Ch 11 in "Advances in military textiles and personal equipment" Editor: E. Sparks, ISBN 1 84569 699 9, Woodhead Publishing 238-259 (2012).
[15] P. Brasser: "Optimizing the protection against the physiological burden of CBRN-clothing", International Journal of Occupational Safety and Ergonomics (JOSE) 16(2), 153-168 (2010).
[16] http://www.e-cooline.de/
[17] *SINETRA* with leader partner Interactive Wear: Smart Integration for rescue Teams with end of the project in 2011. (http://www.mstonline.de/mikrosystemtechnik/smart-textiles/medien/pdf/sinetra)
[18] *IMST project* with Mehler Vario, Lodenfrey and Beurer, (KLIMAJACK) more information on http://www.mstonline.de/foerderung/projektliste/detail_html?vb_nr=V3TEX044;
[19] *ITA/Profitex* is EU-Project with partners a RWTH Aachen and Fraunhofer FIT (S. 10 der Präsentation), more information on (http://www.ima-zlw-ifu.rwth-aachen.de/forschung/aktuelle_projekte/projektansicht/projekt/profitex.html); from 2009 to 2012.
[20] *USA Project*, for Army Tools, developed by University of California San Diego. (http://www.rapydconnex.com/solutions/disaster-communications); 2011.
[21] VIKING fire fighter clothing http://www.viking-life.com/viking.nsf/public/products-firefighting.html
[22] "System for following of the persons and of vehicles in extreme environments« that was developed by the Slovene company Harpha Sea and supported by Slovene Ministry of Defense and was finished in 2010.

Functional textiles have important role on the field of human life

D. Z. Pavlinic
* BRIS – Biomedical Research Innovative Society, Ljubljana, Slovenia
dzpavlinic@gmail.com

Abstract - Textiles are elements which coexist with humans. Some are called functional; because they assist humans in every way, ensure their comfort and protection. They are divided regarding the environment in which humans live, move and/or perform their working activities. These textiles touch human skin irrespective of environment and work type; therefore it is logical that some functions are fulfilled by the textiles placed next to the skin surface. Some of these functions are sweat management system, while other ensures therapeutic, healing, antibacterial and antimicrobial effects and prevent the infections. The latter is important in medicine where patient's bodies are enveloped by different textiles with the aim to cool, heat and/or maintain body temperature as well as other specific care effects, for example the prevention of infections that are spread by textiles used in the hospitals. The inner textile layer with sweat management ability is the most important for establishing the thermal balance in the multilayer system "human-clothing-environment", but it is usually ignored. However, it is obvious that functional textiles that touch our skin impact human life, their comfort, well-being and performance at working and daily activities. In this paper we will see how human life could depend on the different textiles (positive or negative way).

I. INTRODUCTION

Humans first started to use textiles for the protection of their bodies from outside influences, in the winter for protection from the cold and wind and in the summer for protection from the heat and sun. The materials were harvested from their natural environment from tree bark, pelts and skin of animals and other plant material. They knew how to process these materials and by wrapping those around their bodies made use of them for at least one season. They searched for their textiles in the vicinity of their living environments. They chose materials which were most useful by sight and touch. There were no modern production technologies and long transport routes from order to the finished product yet they still managed to make "comfortable" protection layers around their bodies.

Through the timeline of development of the textiles they began to know an increasing number of textile fibers, which were grown and harvested in different ways. They discovered manual manufacture technology, made products according to their wishes and sizes and provided for themselves and their families. The changes brought about in the global markets with the start of intensive industrialization which manifested in textile industry as well ruined the precious hand craft and by doing so significantly weakened the quality of textiles at the time. The fast development of technology procedures for large scale industrial production of textiles after the year 1900 has impoverished some regions and given them the stigma of under development even then. In Slovenia some regions still bear the stigma particularly in Prekmurje and Bela Krajina where they were once in the foreground of woven linen fabrics, while now they are some remnants kept for the preservation of cultural heritage and tourist purposes [1].

From the above mentioned we can gather that humans possessed much knowledge about the products for personal use – clothes and textiles for home, probably more than they know now. The market was being flooded by cheap mass produced linens, while hand made products from linen and hemp were disappearing despite having known medicinal effects. Fast changes in the development, technology and multitudes of chemically produced fibers are the reason for devaluation of genuine hand made products. With the mass development of different textiles the quality control over products we wear in direct contact with our skin was being lost.

Numerous basic and applied researches in the field of textile have today brought us numerous different textiles. These are worldwide grouped according to their functionality i.e. increased ability to dissipate moisture from the skin surface, vapor permeability or "breathing", antibacterial performance, antistatic, protection from UV radiation, wind, chemical influences, etc.

Usually the development of new materials is focused on one kind of material which is then developed to superior quality and finished with a specific added value. Actually this kind of material should gain added value when we find out its compatibility with other similar and also different materials. In this area a void in the knowledge in the textile and clothing industry is felt in basic research as well as in the development of multi-purposeful functional products.

The paper considers different textiles with different functions. The aspects of end users target groups (medicine, work in extreme working environments, every day use for different professions), their needs and criteria for the understanding of functionalities of textile materials are considered. The focus is on the fact that the knowledge of textile materials and their influence on human skin is diminishing. The selection of textiles is based on the

powerful market society where adverse effects on health caused through textiles are generally not regarded.

II. FUNCTIONAL TEXTILES ARE BESIDE US

A. Functional textiles in medicine

The basis for the use of textiles in medicine is in care and maintenance. Clothes as well as all other textiles must adhere to the rules for maintenance and sterilization procedures. The maintenance is performed with industrial wash procedures, which are prescribed from the manufacturers of washing machines and washing agents. These textiles are bought primarily in accord with above mentioned proscribed procedures, while the comfort of the end user is seldom met. The comfort mentioned above is the comfort of wear of the medical staff as well as the patients.

Today the maintenance of textiles in health care institutions is implemented in line with established protocols. Despite the standardized procedures of textile maintenance (bed linens, towels, clothes, etc.) in use in health care institutions the fact that in Europe 37.000 people per year die of Hospital infection - MRSA needs to be pointed out. According to ECDC (European Centre for Prevention and Control) the patients most frequently suffer from urinary tract infections, infections of lower pulmonary tract, infections from caused by surgeries or infections from blood transfusions. Other infections are infections of digestive tract, skin and soft tissues infections. The microorganisms which most commonly cause the infections are Escherichia coli and Staphylococcus aureus, followed by the bacteria Pseudomonas aeruginosa. From the above mentioned microorganisms MRSA (Staphylococcus aureus) causes the most problems because it became resistant to last generation antibiotics. In the last three decades MRSA has caused huge problems in hospitals worldwide. MRSA spreads primarily with direct contact through hands of colonized or infected patients and staff. The prevalence of MRSA in hospital is therefore the best sign for the suitability of hygiene and the success rate of measures for prevention of transfer and spread of epidemiologically important microorganisms. Hospital infections have and will apparently keep shoving up (from 5%-15% and in intensive units even up to 87%). However they can be significantly lowered also by the use of textiles with reduced stain ability (bacteria and viruses). These kinds of infections present one of the most common causes for high mortality rate in intensive care units [2].

This kind of information is worthy of research in the field of textile maintenance system as well as on the textile materials which could act as carriers of the infection. Monitoring such infections is difficult from the medical as well as textile point of view. The field is quite un-researched because modern materials cannot withstand the standard cleaning protocols.

Clothes for patients are also part of the medical textile group. The patients are hospitalized for various infections (bacteria and viruses). They are treated in intensive care units and are experiencing fever conditions. In these cases one would expect that the textile provide additional care, Figure 1. Possibly also the ability to cool or heat the patients with the purpose of maintaining optimal body temperature. Within individual systems “clothing-human-bed linens” materials which are used have the function of regulation of body temperature within the (37±1°C) temperature zone – the temperature neutral zone. At the same time it would be beneficial if these textile materials would also be antimicrobial and antistatic and have a calming and therapeutic effect on the patient.

There are some cooling systems on the market but they are expansive and mostly single use. Because of the cost the intensive care units are often limited on a smaller number of such products. The development of optimal textiles with the above mentioned functions is lagging behind mostly because of care procedures as well as compatibility with other textiles, Figure 2.

Figure 1 Bed clothes in hospitals [3]

Figure 2 Medicine clothing and equipment [3]

B. Functional textiles in extreme working condition

The selection of clothing for work in extremely hot or cold environment is proscribed by standards depending on the nature of the work environment in which the person is

performing its working duties, Figure 3 [4]. Numerous rules in the scope of ISO, EN, SIST standards proscribe testing methods and measuring procedures which the textile should comply with. Usually one layer textile materials from which one piece of clothing or one component of personal protective equipment is tested. The measuring equipment is also included in the appropriate standards. A certain clothing part or part of personal protective equipment complies with standard however there is no written rule of compatibility of the clothing as a whole to the standards or other clothing or other pieces of personal protective equipment.

Figure 3 Functional clothing for extreme working environment

The lack of compatibility of textile materials and furthermore clothing cuts is increasing. A superb increased vapor permeability function of a textile material which will form a middle layer of the clothing could be completely neutralized by an unsuitable top layer of textile material which could be completely no permeable for vapor. The size of clothing within a clothing system is of vital importance in systems that utilize the onion system of layering (multiple layers - each top of the other). A lot of protective systems demand additional personal protective equipment for which compatibility with the clothing system is crucial for example a clothing system worn bellow a fire extinguisher.

C. Functional textiles for daily use at work

Most individuals have free rein when it comes to choosing clothing that they wear while working (office personnel, civil servants, etc.). For this target group comfortable clothing is of primary importance. Of course the comfort zone is defined differently in regards to the individuals own perception of temperature and moisture. Every individual perceive these two factors differently which means that this perception has a different influence on behavior and well being of one person than another. Long term exposure of the workers to different temperatures and humidity can have a detrimental effect to their health although at the time of the exposure they might not be aware of this fact. Many times the detrimental health effect can be seen even years after the exposure.

The guidelines for healthy work environments and wellbeing at the workplace should be seen through clothing as well. However employers are primarily focused on the monetary value when purchasing protective equipment and the health of the employees is often second place. Only in the cases where the rules, regulations and standards are stringently enforced there is compliance to the strictest criteria. The case of protective clothing in construction (outside work) can be seen as a bad practice example. The complexity and work demands (environment and work type) are extremely strenuous, the protection from outside influences in protection of health are not consistent in most cases.

III. CRITERIA FOR OPTIMAL CLOTHING

Every individual piece of clothing still has an effect on a person despite modern clothing manufacture technology, high quality fashion clothing as well as clothing systems for the protection from fire and flames and also less demanding protective clothing. The clothing influence individuals comfort, the level of their work efficiency, protection etc. because they are the element in direct contact with the persons largest organ – the skin. All three factors (efficiency, protection, comfort) are closely interlinked, their negative extremes are known only in given real environment for example in war or noncombat operation in crisis areas where their deficiencies are shown very quickly. The aforementioned three factors influence the wellbeing of a worker wearing a uniform at work [5]. The primary disruptive factors are (dis)comfort caused by not fitting to the 3D image of a human body (Figure 4) and not suitable compatibility of textile materials in one clothing. Not fitting clothing can be defined as too tight or too baggy and can influence the individual's well-being [6].

Figure 4 Different body shape require different clothing

In the case of protective clothing systems the focus should be in the clothing systems shortcomings which are apparent because of the thermal (dis)comfort and are percepted through thermo receptors in the skin [5]. This requirement should take into account the relation "human-

clothing-the environment" in the process of planning and optimization of combat clothing systems. It should consequently take into account the reactions of people wearing such clothing system while performing their work duties. 30 years ago a German researcher Umbach [7] has connected the requirement "human-clothing-the environment" with the soldiers responses in a comprehensive 5 stage process for the development and evaluation functional protective clothing systems. In the course of several years research work his 5 stage process has been upgraded to the requirements of clothing system for the implementation of specific work activities in specific environments, Figure 5 [3].

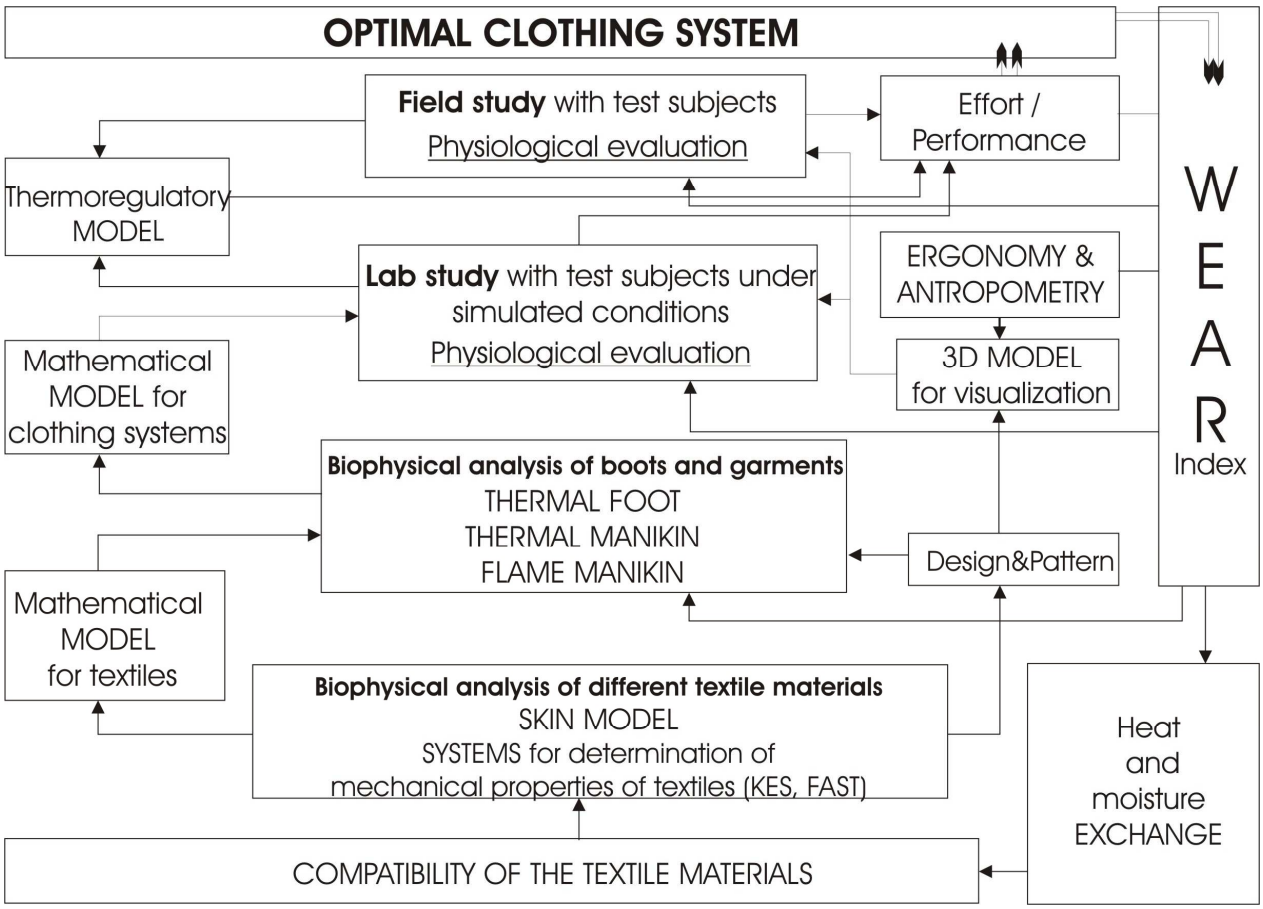


Figure 5 Model of development of the optimal clothing ensemble

The process of development and evaluation of protected clothing following the above model begins with the testing of characteristics of all textile materials incorporated in a specific clothing system. For the achievement of mechanical comfort we are using modern measuring devices for the determination of mechanical and physical characteristics like KES-FB and FAST (placed in research organizations in Slovenia) [6]. The aspect of thermal comfort – thermal isolation and resistance to vapor permeability is researched using skin models. The test is conducted so that the model simulates normal skin temperature in estimated skin humidity for the whole range of environmental conditions for which the protective clothing system was designed.

If other measurement and testing equipment available the testing continues with the help of different modern numeric models. One of the worldwide recognized systems which the author has been using for numerous years is the Wissler thermoregulation model, Figure 6 [8-12].

The model used in this study represents human geometry by 21 cylindrical elements as shown in Fig. 1. In each major element, temperature is computed as a function of time at fifteen points along twelve equally spaced radius vectors. Each cylindrical element is divided into 157 small regions defined by 15 cylindrical shells subdivided into 12 angular sections.

Temperature, and physical and physiological properties, such as density, specific heat, thermal conductivity, rate of metabolic heat generation, and perfusion rate, is computed for each of the 157 small regions. An additional 6 radial shells divided into 12 angular sections are used to define the properties of

clothing on each major element. Longitudinal conduction of heat is neglected, and transport of heat between major elements is affected by arterial and venous blood flow. The model makes allowance for counter-current heat transfer between arterial and venous blood in each element.

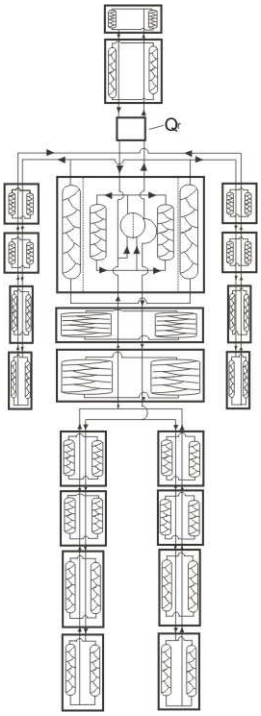


Figure 6 Wissler's human thermal model

An atlas of human anatomy [15] was used to position bone, brain, lung, viscera, muscle, fat, and skin in each section, and information regarding the distribution of bone, muscle, and fat were obtained from the literature, and used to place an appropriate amount of each material in each element. Regional thicknesses of subcutaneous fat are assigned according to gender and the mean skinfold thickness [16, 17].

The use of the abovementioned model is optimal in cases where we have the opportunity to compare the results from the simulations to experimentally measured results. This is not always possible therefore the mathematical model is used when field trials are financially noneligible costs of development. Definitely the knowledge of these modern tools, like mathematical models, requires its time in the development process.

IV. CONCLUSION

In work environments where work is performed under different work conditions (environmental conditions, danger, duration of exposure to certain danger, health risks), the workers from the standpoint of their own safety and protection of health should use personal protection/work equipment a part of which are also suitable clothing systems . The workers are the most familiar with their own specific workplace with all of its demands therefore they should be actively included in the design and choosing personal work and/or protective equipment. The EU includes in its Directive (Council Directive 89/391/EEC of 12 June 1989) a principle that the employers should consult the workers when designing personal protective equipment with the purpose of reducing the risk to their health and/or reduce the exposure time and eliminate the threat is at all possible.

REFERENCES

[1] Linen production (Pridelovanje lanu): Kmetijske in rokodelske novice, 1.1.1876, letnik 34. In Slovene.

[2] Hospital Golnik manuals: How to prevent the spreading of MRSA?, p.1-23, 2008. in Slovene.

[3] http://www.zdravstvena.info/vsznj/preprecevanje-okuzb-in-bolnisnicna-higiena-drugi-del/

[4] www.lenzing.at

[5] Eiken, O. et al.: Termisk belastning hos markstridssoldater vid simulerad patrullering i ökenklimat; effekter av två olika ballistiska kroppsskydd samt av två olika strategier för avkylning, STH Rapport 2010, 0063:8. Internal project report of Royal Institute of Technology.

[6] http://www.human-solutions.com/fashion

[7] D.Z.Pavlinic & I.B.Mekjavic: Potrebe okolja sooblikujejo specifične bojne oblačilne sisteme, Slovenska vojska, Year XVII, No. 16, pp. 29-31, 2009.

[8] Kawabata, S.: The Standardization and Analysis of Hand Evaluation, The hand evaluation and standardization committee, The Textile Machinery Society of Japan, 2nd Edition, Osaka, 1980.

[9] K.H.Umbach: Physiological tests and evaluation models for the optimisation of the performance of protective clothing. In: *Environmental Ergonomics*. Sustaining Human Performance in Harsh Environments. Eds.: Mekjavic I.B., E.W. Banister and J.B. Morrison. Taylor & Francis: Philadelphia, (1987), pp. 139-161.

[10] D.Z.Pavlinic & I.B.Mekjavic: Extreme environmental conditions requires optimal military clothing ensemble, 5th International Textile, Clothing & Design Conference – Magic World of Textiles, October 2010, Dubrovnik.

[11] D.Z.Pavlinic: Delovno oblačilo: Zdravje, dobro počutje ali delovna učinkovitost?, Gospodarski izzivi, Časopis Štajerske gospodarske zbornice, Letnik 5, Junij 2012, p.35.

[12] Wissler E H. Mathematical simulation of human thermal behavior using whole body models. Chapter 13 in Heat Transfer in Medicine and Biology, pp. 325–373, A Shitzer and RC Eberhart (Eds.), Plenum Press, New York, 1985.

[13] Wissler EH. Prediction of survivability of "Lost Bell" accidents under various conditions. Underwater Physiology, VIII, pp. 401-412, AJ Bachrach and MM Matzen (Eds), Undersea Medical Society, Bethesda, 1984.

[14] Pavlinic DZ, Wissler E and Mekjavic IB. Modeling thermophysical responses of military personnel conducting a variety of activities during simulated field operations in a cold environment. NATO 2008 Finnland, Helsinki.

[15] Netter, F.H. (1989). *Atlas of Human Anatomy*. Ciba-Geigy Corp., Summit, New Jerse.

[16] Hayes, P.A, P.J. Soward, A. Belyavin, J.B. Cohen, and F.W. Smith (1988). *Sub-cutaneous fat thickness measured by magnetic resonance imaging, ultrasound, and callipers.* Med. Science in Sports and Exercise, Vol. 20, pp. 303-309.

[17] Hayes, P.A., J.B. Cohen, and P.J. Soward (1987), *Subcutaneous fat distribution of adult males and females measured by nuclear magnetic resonance.* IAM report No. 655, RAF Inst. Aviat. Med., Farnborough, Hampshire, UK.

I-TEHMED Biomedical Technology Platform – Structure and Results

J.Stanič*, J.Jelenc**, U.Stanič***
* Iskra Techno research and development institute, Podnart, Slovenia
**Iskra Medical d.o.o., Ljubljana, Slovenia
*** Biomedical Research and Innovative Society, Ljubljana, Slovenia
janastanic@gmail.com

Abstract - Slovene Biomedical Technology Platform I-TEHMED was established in 2005 at the initiative of leading Slovene industrials, research organisations, universities in the field of biomedicine and biomedical engineering as well as University Clinical Centre and other secondary and tertiary hospitals. I-TEHMED is an open structure, which utilises a bottom up approach to research and develop new generations of biomedical devices, medical procedures, innovative drugs, therapeutic methods, medical and other services in the sphere of biomedicine. The platform provides a basis for constant generation and exchange of knowledge and experiences between its members. It also enables the creation of synergistic effects with other entities both foreign and domestic. Thus the platform is ideal for implementation of research projects funded from either national or EU grants. In this paper we will discover the reasoning behind I-TEHMED, its mission and vision, take a look at its multidisciplinary members and above all focus on four case studies – research projects, which were successfully implemented by its funding members. The results of said projects were the basis for new high added value products and processes.

I. INTRODUCTION

European Technology Platforms focus on strategic issues where achieving Europe's future growth, competitiveness and sustainability depends upon major technological advances. They bring together stakeholders, led by industry, to define medium to long-term research and technological development objectives and lay down markers for achieving them. The achievement of these objectives will significantly improve the daily lives of the European citizen in many areas.

Technology platforms play a key role in better aligning EU research priorities to industry's needs. They cover the whole economic value chain, ensuring that knowledge generated through research is transformed into technologies and processes, and ultimately into marketable products and services.

To remain competitive, European industry needs to specialise more in high-technology areas. Investment in research must be increased, coordination across Europe enhanced and the technological content of industrial activity raised. Technology platforms address these challenges through:

- Shared vision of stakeholders;
- Positive impact on a wide range of policies;
- Reduced fragmentation of research and development efforts;
- Mobilisation of public and private funding sources.

Technology platforms follow a three-stage process of development:

- Stakeholders, led by industry, come together to agree a common vision for the technology.
- Stakeholders define a Strategic Research Agenda setting out the necessary medium- to long-term objectives for the technology.
- Stakeholders implement the Strategic Research Agenda with the mobilisation of significant human and financial resources.

The technology platform I - TEHMED is an open structure, which – in order to be efficient – needs to include all key parties, from the economy to institutes, universities, public institutions and the state itself. The platform induces partnership with other sectors and in this way represents the foundation for political dialogue.

The technology platform support constant growth of adequate knowledge in specific technological areas, as well as usage of new technologies. In the technology platforms, the leading role belongs to the economy, which is the main carrier of research & development (R&D) projects. The main goal of the technology platform is to form a vision for future activities in the technological areas the Technology platform covers and to define a strategic program or research agenda for that area

Strategic Research Agenda of I-TEHMED for 2007-2013 is a fundamental program document of the technology platform which that defines the areas of activities of the technology platform's members and its work programme for a seven-year medium-term period between 2007 -2013.

A. *Vision and strategy of I-TEHMED*

Members of TP I-TEHMED will - through active participation of partners from EU and elsewhere – develop a new generation of innovative drugs, medical procedures, therapeutic methods and other services and medical devices that will enable Slovenian medicine to offer high-quality services with high added value to home and foreign patients. Technology platform I-TEHMED will, in order to create synergistic effects, work in connection with technology platforms in Slovenia and EU, as well as with academic and research organizations, technology networks, the health sector, relevant Ministries and Government Agencies …

The Technology platform I - TEHMED is connected or cooperates with:

- Technology platform Innovative Medicines for Europe
- Technology platform Nanotechnologies for Medical Applications
- Technology umbrella – Technology Park of Slovenia
- Tourist umbrella- STO
- Network of Slovenian hospitals
- National agencies Slovenian Research Agency and Slovenian Technology Agency
- national institute for medical research – (Institute BRIS)
- Integrated information system ”Slovenia Your Cooperation Partner- SYCP ”
- National technology platform Fotonika 21
- Technology platform Photonics 21
- National technology platform eMobilnost
- Centre of excellence "Biotechnology with pharmacy”

B. *Organisation of I-TEHMED*

Members of the TP are companies that develop new technologies, companies that use technologies and Institutions of knowledge (Universities, other private or public research and development organizations. At the moment more than 60 Members from 7 different are Countries joined in Technology Platform I-TEHMED which is active in:

- 1.Workgroup for Biomedical Engineering
- 2.Workgroup for Biopharmacy
- 3.Workgroup eHealth and Telemedicine
- 4.Workgroup Nanotechnologies for medical applications
- 5.Workgroup for Health Tourism
- 6.Workgroup for Space medicine

C. *Potential impact of I-TEHMED technology platform*

Description of health & medical potentials of Slovenia:

- 29 hospitals that equally cover whole territory of Slovenia (2 mio people)
  - 2 two teaching hospitals
  - 4 teaching clinics
  - 11 general
  - 4 psychiatric

Cooperation between the medicine and technology has a long tradition in Slovenia and has brought much success. The technical devices and innovative methods of treatment which were developed to fulfil the specific needs of leading Slovene doctors have markedly increased the quality of health services and at the same time the companies and health institutions (both public and private) could market these products and services on the global market.

The members of I-TEHMED have with its research projects and development enabled the implementation of modern, quality and competitive medical products and services which go beyond the state of the art. The synergies inherent in the platform have stimulated the increase of knowledge and technology transfer in the biomedical sector in Slovenia and beyond. Slovene biomedical companies were and are still able to translate this wealth of knowledge into new products which they could market on the global markets. The synergies within the platform were also beneficial for the increase of quality of research in our research institutions and universities. I-TEHMED is excellent at providing the networking needs that the biomedical community sorely needed.

II. CASE STUDIES

The members of I-TEHMED have conducted a fair number of successful national and European research projects. To shed a light on many areas of interest within the platform some of these ongoing or finished research projects are presented bellow.

A. *InTraMed-C2C*

InTraMed-C2C is a European project, with the participation of 7 countries and 10 project partners.

The project idea was born by the project partner Clusterland Upper Austria. They experienced that clinics have a high potential for innovations in product, process and service development in the overall medical industry, but this innovation potential is insufficiently transferred into results such as product development for companies, Intellectual Property (IP) generation and better healthcare processes.

InTraMed-C2C will initiate, develop and provide tools for strengthening the access of small and medium enterprises (local and European wide) to the innovation potential in the clinics.

One of the reasons for the insufficiently transferred results is that clinics and companies have different missions concerning handling innovative ideas. Clinics are "end-users" of innovations and clinics are also driving forces for bottom-up approaches of innovation transfer into the industry and their products. Companies, especially global players, are highly interested in providing their existing range of products to the clinics, new innovations are interesting only regarding the economic aspects. Moreover, these global players have established adequate access to the key functions in the clinics, supported by their distinctive marketing resources. As a result, global player dominate the supply of products and services in the clinics. Small and Medium Enterprises (SME) and developments initiated by the clinic staff often do not have the resources and chances to transfer their ideas and knowledge into adequate product developments.

InTraMed-C2C develops an innovation transfer system ("framework") as an exchange and coordination mechanism for innovations in the medical/clinical sector between key players of the regional innovation environment. This innovation transfer system will be embedded in the regional innovation framework. With its trans-regional approach InTraMed-C2C will extend this framework beyond the regional approach in order to provide innovative ideas to companies for realizing these ideas within the system developed by the project. It is in the scope of the project to initiate a European network for Innovation transfer as an open exchange structure for innovations and their realizing companies and institutions within the Central Europe area.

B. Planetary Habitat Simulation

The aim of the Framework 7 project "Planetary Habitat Simulation (PlanHab)" project, coordinated by Jozef Stefan Institute, is to investigate the combined effects of hypoxia and sustained recumbency (bedrest), on human physiological systems. The partial pressure of oxygen in the environmental gas inside future planetary habitats will be lower than in atmospheric air. Prolonged exposure to low gravity will result in deconditioning of vital physiological systems, and may consequently constitute a threat to the health of the astronauts. However, it is unknown how prolonged exposure to both reduced gravity and hypoxia will affect health. The new knowledge has also implications for society in general, since chronic hypoxia and bedrest constitutes a model of the basic conditions experienced by patients suffering from respiratory insufficiency restricting them to a physically inactive life style. The challenge of the project lies in the complexity of the experimental interventions where healthy humans are confined to a hypoxic environment during prolonged bedrest. A series of studies will be conducted at the Planica hypoxia facility capable of housing 20 subjects at any simulated altitude. Subjects will remain in horizontal position (bedrest) or be ambulatory, but confined to the facility (ambulation) for 21 days/trial. Each subject will participate in three trials: hypoxic bedrest (simulated altitude 4000m), normoxic bedrest, and hypoxic ambulation. The effects will be investigated in experiments concerning metabolic, cardiorespiratory, musculoskeletal, haematological, immunological and thermoregulatory functions. In addition to the specific objectives, the study will be explorative in the sense that it will collect a broad spectrum of basic data corresponding to that obtained when 21-day bedrest experiments are conducted by ESA/NASA (bedrest core data). Thus, data from the experiments can readily be compared with core data from previous bedrest studies, and hence the added effects of hypoxia should be evident.

C. *PhysiCoDerm: Enhancement of Cell Membrane and Skin Permeability for Transdermal Drug Delivery by Means of Physical Methods*

The PhysiCoDerm project is a part of the Competency Center for Biomedical Engineering (CC BME).

CC BME is an EU regional Competency Center for Biomedical Engineering, coordinated by Laser and Health Academy, and supported by the European Regional Development Fund and Slovenia's Ministry of Higher Education, Science and Technology. The vision of CC BME is to establish Slovenia on the global map as an incubator of highly innovative companies that build the most interesting biomedical devices in the world.

Biomedical engineering is one of the world's most promising and fastest growing industries. Rapid technological developments are being fuelled by breakthroughs in basic sciences and engineering, as well as by increasing global demand for more affordable, effective, and minimally invasive medical treatments.

With a focus on global market leadership, 12 partners, supported by Slovenia's Ministry of Higher Education, Science and Technology, as well as the European Regional Development Fund, established an EU regional competency centre for biomedical engineering (CC BME). The consortium's partners include key institutions of higher education, industrial representatives, and experts in the field of biomedical engineering and medicine that wish to expand upon their existing avenues of cooperation to achieve new levels of synergy. The program of the CC BME is broadly focused on the research and development of products and applications that reduce the invasiveness of medical treatments and increase the reliability and accuracy of diagnostic procedures.

The PhysiCoDerm project is concentrated on the development of new therapeutic devices and methods that are based on the effects of electric fields and ultrasound on cells. The research is aimed towards finding therapeutic methods that will allow the entry of active substances through the skin, into the skin cells and subcutaneous tissue.

D. *Non invasive low frequency ultrasound device for fat reduction*

Iskra Medical in cooperation with Slovenian and international I-Techmed partners developed a non invasive low frequency ultrasound device for fat reduction. Subsequent research and development lead to a novel simultaneous use of low frequency ultrasound and vacuum endodermal treatment, improving clinical results of non-invasive treatment. The clinical study on abdomen of 8 women showed abdomen circumference reduction of 2,938 + 0,623 cm after a single 45 minutes treatment. Corresponding treated local subcutaneous fat thickness was reduced bz 2,96+ 2,01 mm measured using 7.5 MHz diagnostic ultrasonography [7]. The measured results correspond to subsequent visual improvement of treated women. There were no serious side effects observed in the study or vigilance monitoring. This patented technology is now in use 20 different countries with great clinical as well as commercial success [Iskra Medical company data].

III. CONLUSION

Although the state financing for the platforms is no longer available the members of the I-TEHMED technology platform are undeterred in their pursuit of knowledge, research and new products that will satisfy the needs of the medical and biomedical profession worldwide.

I-TEHMED and its members strive to actively contribute to the Slovene strategic maxim:

"Slovenia should continue to be a country, in which it is pleasant to live, learn, work and have fun. Slovenia should be a beautiful and safe country, where good business deals are made; the home of state-of-the-art technology and culture; a place of pure water and healthy food, where citizens and visitors alike shall find rest, good health and regeneration."

ACKNOWLEDGMENT

The authors thank the coordinators and partners of the projects highlighted in the paper for their insight, cooperation and excellent research. Namely we thank Mr Bogo Filipič, prof. Igor Mekjavič, Mr Jože Jelenc, prof. Damijan Miklavčič.

REFERENCES

[1] "Strategic Research Agenda of the I-TEHMED technology platform - Strateški razvojni načrt tehnološke platforme I-TEHMED" Ljubljana, 2007

[2] European Technology platforms - Knowledge for growth, European Commission, 2005

[3] Simona Rataj, "Analysis of the technology platforms in Slovenia - Analiza stanja tehnoloških platform v Sloveniji, GZS, November, 2009

[4] InTraMed – C2C project, CENTRAL EUROPE Programme co-financed by ERDF

[5] "Planetary Habitat Simulation" 7th Framework Programme http://www.planhab.com/

[6] The Competency Center for Biomedical Engineering (CC BME) Introduction http://www.bmecenter.com/en/

[7] J. Smiljanic and V. Persic, "Reduction of Subcutaneous Adipose Tissue Using a Novel Vacuum-Cavitation Technology,"*Acta Dermatovenerologica Albanica*, vol. 7, no. 1, pp. 71–75, 12 2010.

[8] Uroš Stanič & Jana Stanič Regional Innovation Strategy of Slovenia, , 1st Conf.with Int.Participants, 30.Sept.-1.Oct. 2010, pp. 107-114, Udruga Inovatora Hrvatske, Zagreb, Croatia

Multi-Functional Wireless Body Sensor – Analysis of Autonomy

K. Bregar and V.Avbelj
Jožef Stefan Institute, Department of Communication Systems, Ljubljana, Slovenija
Klemen.Bregar@ijs.si, Viktor.Avbelj@ijs.si

A wireless multi-function biosensor that measures a potential difference between two proximal electrodes on the skin, enables monitoring of vital functions - heart activity and respiration. The sensor is designed as a small plaster-like reusable unit that can be easily fixed onto the body surface and is therefore minimally obtrusive for users. It is equipped with a signal acquisition unit, a processor for on-line data analysis and with enough memory for a temporal storage of measured data. Incorporated low-power radio system transmits the measured data to a radio receiver, which is installed either in a dedicated personal terminal, smartphone or in a ward gateway. The sensor is powered by a small coin battery. Visualization, archiving and detailed interpretation of data can be implemented on a remote computer server. The autonomy of the monitoring system, regarding its power consumption, depends significantly on the portion of local signal processing on the sensor, on the frequency of data transmission and on the amount of transmitted data. Different test scenarios have been tested and evaluated regarding the power consumption, reliability and robustness.

I. INTRODUCTION

The benefit of telemedicine and telecare systems has been confirmed in numerous studies [1, 2]. In our previous studies we developed a proof-of-concept system for biosensor data acquisition through wireless body sensors (WBS) [3] and an application with three wireless ECG sensors [4]. To be unobtrusive the WBSs should be small, multi-functional and with long autonomy. These requirements are in opposition, so a careful design of the whole system is needed.

The main biosignal of interest is ECG which also carries embedded information of respiration rate. We can easily extract the respiration rate from observing variation of R-peak heights during time if respiratory related modulation of heart rate is high enough in measured individual [5]. We can also extract heart rate information from time differences between R-peaks in ECG signal.

The most critical part of our multi-function WBS in case of energy efficiency is wireless communication part of circuitry (radio). There are a few good communication standards for battery powered wireless sensors available. All of them are working in 2.4 MHz ISM band and are inexpensive and very energy efficient.

The first candidate to consider of was ZigBee. The main benefits of using ZigBee are self-organizing networks, great energy efficiency, incorporated industry security standard AES-128 and great co-existence with other devices in the same frequency band [6]. The biggest flaw of ZigBee standard is that we can't use it on smartphones without additional special extension modules for mobile phones.

The second candidate for wireless transmission was ANT+ technology. In addition to ZigBee specification it has very compact protocol stack which means that it requires really low level of interaction with application processor and it can be used with newest smartphones with ANT+ technology incorporated [7].

The last and in our case the best communication protocol for wireless transmission of vital functions was Bluetooth low energy (BLE) technology. The main reason for choosing BLE was the accessibility of technology in smartphones nowadays. BLE part of Bluetooth technology is incorporated in all new Bluetooth 4.0 devices. This means that every new phone in the future will have BLE technology incorporated which is not necessarily true for ANT+ technology.

The main advantages of BLE technology are [7, 8]:

- Ultra-low peak, average and idle mode power consumption
- Multi-vendor interoperability
- Enhanced range (depends on individual implementation)
- Great co-existence with other devices in same frequency band.

In the next section Multi-Functional WBS design and battery characteristics are described. A complexity of data analysis algorithms is then presented. Finally, energy consumption measurements are made for all test cases and autonomy approximation is made according to measured energy consumption. The paper concludes with a summary of results and an indication of future work.

II. WIRELESS SENSOR FOR SKIN BIOPOTENTIAL

Our Multi-Functional WBS consists of two self-adhesive ECG electrodes which need to be positioned 5 cm apart, analog front-end (AFE), ultra-low-power (ULP) microcontroller and ULP BLE radio.

AFE is designed to suppress RF frequencies in differential input signal and to filter the differential input signal with band-pass filter (0.15 Hz to 50 Hz, -3dB). The amplification is set to level which is high enough to make the weakest signals strong enough for detection by onboard 10-bit A/D converter and low enough to prevent clipping by the strongest signals.

The microcontroller we used is TI's MSP43F2274. This 16-bit microcontroller is part of TI's family of ULP microcontrollers for use in ULP devices. It has built-in 10-bit A/D converter, 32 kB of FLASH and 1 kB of RAM. The average power in active mode is 270 μA at 1 MHz [10].

The best choice for BLE radio was Nordic Semiconductor's nRF8001 single-chip BLE connectivity IC. It integrates BLE Radio, Link Layer, Host stack and features simple serial interface for use with almost all external application microcontrollers available with peak current of 14.6 mA during RX and average current of down to 9 μA (for 1 s connection interval) [11]. It is designed to operate only in Peripheral (Slave) role.

We have chosen lithium coin battery CR2032 as power source for our Multi-Functional WBS. The nominal cell capacity is 220 mAh at discharge current of 200 μA [12].

The actual (effective) battery capacity is not fixed for all conditions. Wireless application as a contrast to standard coin cell application subjects the battery mainly to very dynamic current profile where the maximum impulse current can be around 80 mA which is much higher current as that the battery capacity is rated at (200 μA) [13].

A lithium coin battery's available capacity is dependent on the rate of discharge. Its high current rate is governed by internal resistance (IR), electrolyte conductivity, ion mobility and total surface area within a battery [13]. High average drain will reduce the battery capacity, but the high current pulsed drain will add a further drop in the battery capacity.

It is important when designing for coin cell operation to choose active devices with as low peak current and minimal voltage of operation as possible [13].

We can expect effective coin cell capacity of around 200 mAh with 1 ms pulse load of 10 mA with pulse period of 25 ms, 180 mAh with 1 ms pulse load of 30 mA with pulse period of 25 ms, 160 mAh with 1 ms pulse load of 30 mA with pulse period of 15 ms and only around 130 mAh with 1 ms pulse load of 30 mA with pulse period of 10 ms [13].

We have to be careful when calculating the autonomy of our device working on CR2032 coin cell battery. As long as application can stay in a mode with a lower current consumption it will be able to operate longer, but when the capacity is going towards end, it may fail entering mode with higher current consumption [13].

III. COMPLEXITY OF DATA ANALYSIS ALGORITHMS

We compared two scenarios for measuring and sending information about heart activity. In first case we capture the ECG samples and store them in a message buffer and then send them in raw format to personal BLE collecting terminal. In second case we capture ECG samples and calculate the R-R interval lengths and also heart rate information and then send them to BLE collecting terminal.

For monitoring of heart activity in both cases we choose sampling frequency of 125 Hz for telemedicine applications and frequency of 1000 Hz for diagnostics purposes [14].

In first case where we send raw measurement data, we gather 10-bit samples in 19 bytes long message buffer in the microcontroller. This is the maximum payload length for nRF8001 radio when used for Heart Rate Service. This means we can put 15 ECG samples in one BLE message and with 125 Hz sampling rate it requires 120 ms maximum BLE connection interval or 8.33 messages per second. For 1000 Hz sampling rate that means 15 ms BLE connection interval or 66.67 messages transferred per second.

In second case we calculate R-R interval and heart rate. The R waves of an ECG signal are parts of QRS complexes. The R waves have normally positive polarization with the duration of less than 50 ms. The R peak time (short R time) is usually interpreted as the time of maximal amplitude of the R wave. There are several approaches for the determination of R time [15,16], e.g. detection of the maximal value of QRS by simple comparison and sorting of amplitudes, best correlation with an R wave template, analysis of derivatives with some time restrictions, etc.

We have used an adaptive and robust algorithm that searches for the consecutive positive and negative fist derivatives in a predefined time interval and maximal value of the amplitude in this interval. The first derivative is obtained by calculating the differences between each two neighboring samples. The detector searches for four consecutive positive and four consecutive negative derivatives above a threshold in a time interval determined by the expected R peak width. If such an event occurs, it is declared as the R-peak. The threshold values for the derivatives are estimated dynamically as 10% of the average from the maximum and minimum calculated derivatives. The search window is 50 ms. The R time is the maximum value of the ECG inside the current interval. If an R peak is found, the search counter is increased by the width of the search window to jump over the detected peak. If we need to detect the time of negative short waves, e.g. Q wave, the same algorithm could be used on the inverted ECG input. A time interval between two R peaks denotes the duration of a heart beat and its reciprocal value, multiplied by 60, is the beat rate. The schematic block diagram of the R peak detector algorithm is shown in Fig. 1.

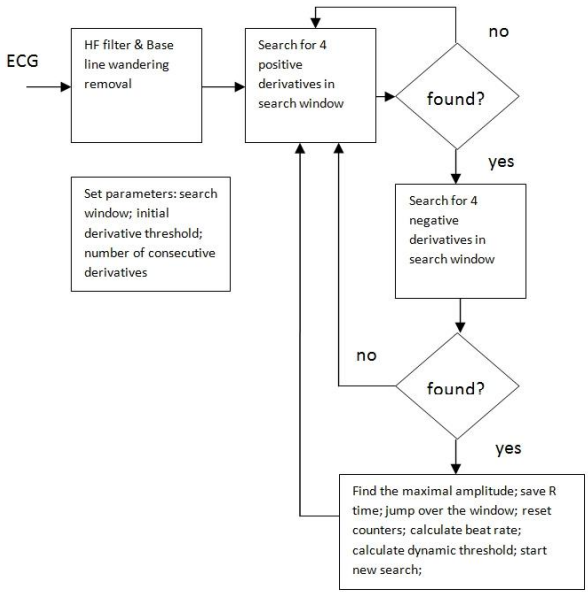


Figure 1. Schematic block diagram of R peak detector algorithm

IV. RESULTS

We can measure average current consumption with very simple method. In series with WBS we connect measuring resistor and then we can measure voltage drop caused by current flow through resistor. The calculation from voltage to current is trivial and is based on Ohm's law:

$$i_n = \frac{u_n}{R}, \qquad (1)$$

where i_n represents the electric current at exact moment, u_n represents voltage drop in the same moment and R represents resistance of measuring resistor. We calculate average current by calculating numerical integral of current profile on WBE over measuring interval.

As a voltage source we used regulated DC power supply CALTEK-PSM3/2A with output voltage set to 3 V and digital storage oscilloscope Tektronix TDS5104B with 1 GHz bandwidth and 5 GS/s sampling rate. The resistance of measurement resistor is 12 Ω.

Measurement of average current for raw data transmission shows, that average current is rising in correlation with sampling rate. The main part of current consumption is caused by transmitting data packets. Radio uses few times more energy for transmission than microcontroller for data analysis. This is clearly seen in Fig. 2 where very small spikes on the bottom of the figure represents active mode of microcontroller and the big wave with peak at around 13 mA represents activity of BLE radio. This means that we have to hold the frequency of transmissions as low as possible and to transmit as much data as possible at once.

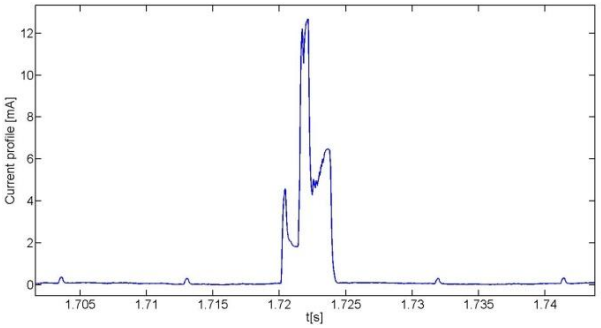


Figure 2. Current profile for raw data transmission with sampling rate of 125 samples per second.

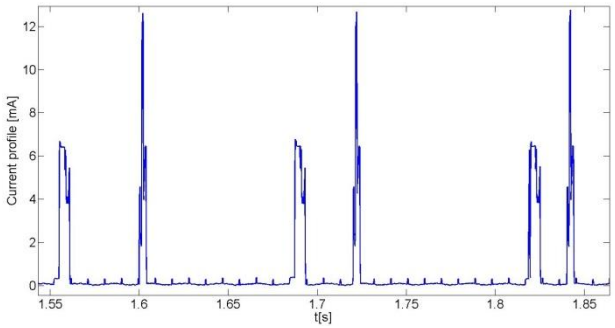


Figure 3. Density of raw ECG transmissions with sampling rate of 125 samples per second .

Measurements for raw ECG transmission mode with sampling rate of 125 samples per second shows us average current consumption of 480 µA. This mode demands 120 µs connection interval or 8.33 transmissions per second. At this rate of average discharge rate and frequency of pulse drain we can expect at least 200 mAh [13] effective capacity of CR2032 battery which means that BLE WBS will run for 17 days on one battery.

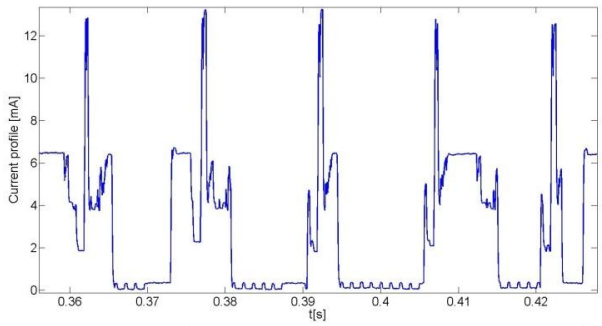


Figure 4. Density of raw ECG transmissions with sampling rate of 1000 samples per second.

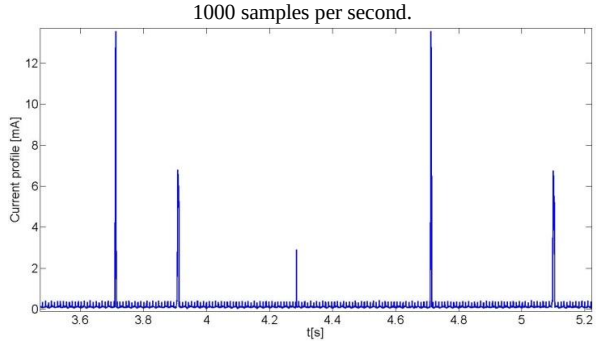


Figure 5. Density of transmissions with sampling rate of 125 samples per second and R-R interval detection algorithm.

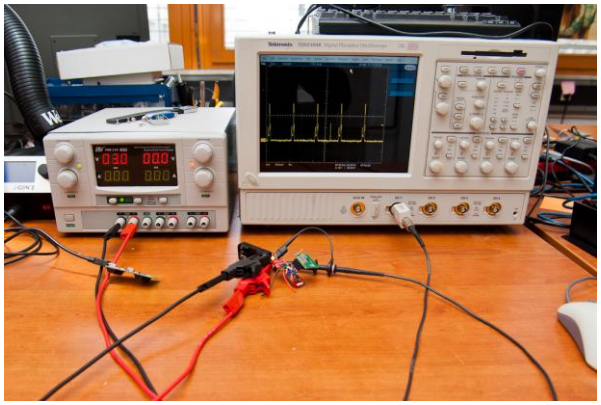


Figure 6. Measuring current consumption of wireless sensor for skin biopotential.

In example with sampling rate of 1000 samples per second the average current is 2.86 mA. It bring 15 ms connection interval and 66.67 data transmissions per second and according to pulse drain battery characteristics we have effective battery capacity of at least 160 mAh [13] which brings 56 hours of autonomy with one CR2032 battery.

Different story is when we use data analysis for heart beat rate extraction. If we capture ECG signal with sampling rate of 125 samples per second and we send only beat rate information with connection interval of 1 s we get 170 µA average current consumption. With the average current which is lower than measuring current for CR2032 battery capacity in datasheet [12] we can expect that battery will give us full capacity of 220 mAh. Simple calculation tells us that system with current consumption of 170 µA will give us 54 days of autonomy.

With data analysis and 1000 samples per second we get just a bit higher average current consumption. Application now takes 210 µA which also allows battery to develop its full capacity. Autonomy with 1000 samples per second and with connection interval of 1 s gives us 44 days of autonomy on CR2032 battery.

V. CONCLUSION

We can see clearly that average current consumption is strongly dependent on sampling frequency and consequently on connection interval of BLE system.

With just an algorithm for calculating R-R intervals and heart beat rate we can seriously reduce power consumption which means longer battery life. Heart beat rate doesn't carry much information about heart activity but just enough that we can detect anomalies in rhythm and in case of anomalies we can immediately start raw data transmission for better insight in current problem with heart activity.

Further we can improve our multi-functional WBS with algorithms for respiration rate extraction from ECG signal. With respiration rate information we can get additional vital sign information without adding new sensors to WBS.

For the future work we are also planning to develop additional algorithms for local on-line ECG signal analysis. The most important here is compression algorithm based on ECG features detection where we transfer only ECG signal features and on the receiver side we can synthesize original signal back. This mode is very useful when multi-functional WBS detects anomaly in beat rate and starts sending raw ECG signal. With that kind of algorithm we can significantly lower the power consumption in comparison to full raw ECG signal transmission and nevertheless we can obtain all information about heart activity.

ACKNOWLEDGEMENT

This work was partially supported by the Slovenian Research Agency through the programme P2-0095 and project J2-4197.

REFERENCES

[1] A.G. Ekeland, A. Bowes and S. Flottorp, “Effectiveness of telemedicine: A systematic review of reviews”, Int. J. Med. Inform. 79, 736-771, 2010.

[2] W. Chaboyer, L. Thalib, M. Foster, C. Ball and B. Richards, “Predictors of Adverse Events in Patients After Discharge From the Intensive Care Unit”, Am. J. Crit. Care, vol. 17, 255-263, 2008.

[3] A. Rashkovska, I. Tomasic, K. Bregar and R. Trobec, “Remote monitoring of vital functions - proof-of-concept system”, MIPRO 2012, 35th International Convention, Opatija, Croatia, May 21-25, 2012, pp. 463-467.

[4] A. Rashkovska, I. Tomasic and R. Trobec, “A telemedicine application : ECG data from wireless body sensors on a smartphone”, MIPRO 2011, 34th International Convention, Opatija, Croatia, May 23-27, 2011, MEET & GVS, vol. 1, pp. 293-296.

[5] R. Trobec, A. Rashkovska, V. Avbelj, “Two Proximal Skin Electrodes — A Respiration Rate Body Sensor”, *Sensors* **2012**, *12*, 13813-13828.

[6] ZigBee Specification Overview, http://www.zigbee.org/Specifications/ZigBee/Overview.aspx

[7] ANT Basics, http://www.thisisant.com/developer/ant/ant-basics/.

[8] Bluetooth 4.0 with low energy technology paves the way for Bluetooth Smart devices, http://www.bluetooth.com/Pages/Low-Energy.aspx.

[9] Adopted Bluetooth Core Specifications: Core Version 4.0, https://www.bluetooth.org/Technical/Speci_cations/adopted.htm , 2010.

[10] MSP430F22x2, MSP430F22x4 Mixed Signal Microcontroller (Rev. G), http://www.ti.com/lit/gpn/msp430f2274

[11] nRF8001 Single-chip Bluetooth low energy solution: Product Specification 1.1, http://www.nordicsemi.com/eng/content/download/2981/38488/fie/nRF8001_PS_v.1.1.pdf

[12] CR2032-Panasonic-datasheet-4810.pdf, http://ebookbrowse.com/cr3032-panasonic-datasheet-4810-pdf-d328952679

[13] P. Furset, K. Homan, “High pulse drain impact on CR2032 coin cell battery capacity,” http://www.eetimes.com/ContentEETimes/Documents/Schweber/C0924/C0924post.pdf , 2011.

[14] R. Trobec, M. Depolli, V. Avbelj, “Wireless network of bipolar body electrodes”, *Seventh International Conference on Wireless On-demand Network Systems and Services (WONS 2010)*, pp.145-150.

[15] J. Pan and W. Tompkins, “A real time QRS detection algorithm,” IEEE Trans. Biomed. Eng., 32(3), 230–236, 1985.

[16] C. De Capua, A. Meduri, R. Morello, “A Smart ECG Measurement System Based on Web-Service-Oriented Architecture for Telemedicine Applications,” IEEE Transactions on Instrumentation and Measurement 59(10), 2530-2538, 2010

Personal health monitoring with Android based mobile devices

M. Kozlovszky*, Bartalis L.*, B. Jókai*, J. Ferenczi*, P. Bogdanov*, Zs. Meixner*, L. Németh*, K. Karóczkai**

* MTA SZTAKI/Laboratory of Parallel and Distributed Computing, Budapest, Hungary
* Óbuda University/John von Neumann Faculty of Informatics, Budapest, Hungary
kozlovszky.miklos@nik.uni-obuda.hu

Abstract - We have developed an Android based mobile data acquisition (DAQ) solution, which collects personalized health information of the end-user, store analyze and visualize it on the smart device and optionally sends it towards to the datacenter for further processing. The smart mobile device is capable to collect information from a large set of various wireless (Bluetooth, and WiFi) and wired (USB) sensors. Embedded sensors of the mobile device provide additional useful status information (such as: user location, magnetic or noise level, acceleration, temperature, etc.). The user interface of our software solution is suitable for different skilled users, highly configurable and provides diary functionality to store information (about sleep problems, can act as a diet log, or even can be used as a pain diary). The software enables correlation analysis between the various sensor data sets. The developed system is tested successfully within our Living Lab facility. Sensor data acquisition on the personal mobile device enables both end-users and care givers to provide better and more effective health monitoring and facilitate prevention. The paper describes the internal architecture of the software solution and its main functionalities.

Keywords

Remote health monitoring, portable sensor data acquisition, mobile device

I. INTRODUCTION

The aging population of industrialized countries grows and this increases also among other things the health care costs. Transparently embedded remote health care can become a new cost effective paradigm, which can solve most of the problems primarily centralized Health Care system's have. Currently, there is a large number of enabling technologies to measure the patient's physiological signals remotely. With handheld and PC devices used as data acquisition (DAQ) systems we are able to collect vital information about the (elderly and demented) patients remotely. Due to the different - in most cases proprietary and incompatible- sensor technologies and solutions, it is a hard task to create generic, user friendly DAQ systems. There are already remote patient monitoring solutions available such as the Android based MyFitnessCompanion [1], which is able to support the following therapy fields: Fitness, Diabetes, Asthma, Obesity, Hypertension, CHD, or the iCare[2] which provides medical guidance, emergency alarm functionality and collects personal health information. Other example is the Microsoft HealthVault [3], which supports care of elderly persons (e.g.: neurodegenerative diseases, stroke etc.), additionally it provides online web interface to manage (process and share) health information. Biotech Lab at Obuda University is involved in AALAMSRK [4] (a national R&D project), specialized both on Android based (we call this MobileHub) portable remote monitoring applications, and normal PC based (we call this HomeHub) remote monitoring solutions (shown in Fig. 1.).

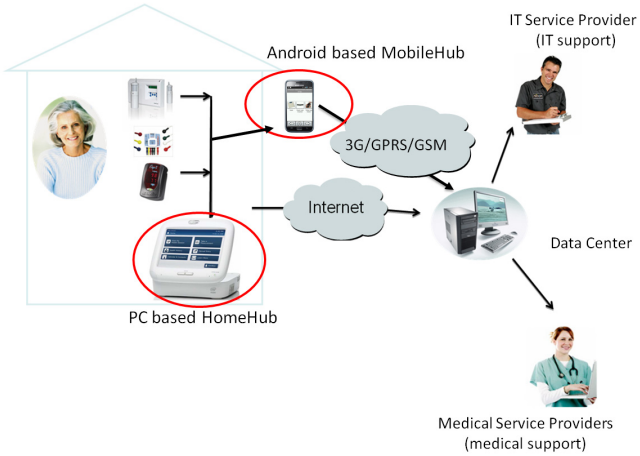


Figure 1. Overview of the remote health monitoring infrastructure (with PC based HomeHub and Android based handheld MobileHub collecting sensor data) [9]

This paper shows how we have built up our remote patient monitoring environment (DAQit)[8] using a client side software and the DrHealth portal. Later on we present how this software environment has been used to do patient's location/sudden event monitoring, remote diabetes and hypertension monitoring in our Living Lab.

Through the co-operation of commercial companies, universities and other non-profit organizations the direct goal of the AALAMSRK project was to develop an integrated, standardized dementia and health monitoring

system supported by innovative, modern measurement and info-communication technologies. By the integration of medical expertise and developing assisted living patterns (ALPs), the realized system offers personalized monitoring solution for monitoring and prevention of elderly people, particularly who suffer from stroke, neurological diseases such as dementia or depression. Our two DAQ solutions (HomeHub and MobileHub) are able to serve overlapping patient categories (shown in Table 1.).

Table 1. Targeted patient categories of the two remote health monitoring solutions (HomeHUB and MobileHUB)

Targeted Patients & persons	PC based HomeHub	Android based MobileHub
Diabetes	Yes	Yes
Hypertension	Yes	Yes
Elderly monitoring	Yes	Yes
Epilepsy	Yes	Yes
Parkinson's disease	Yes	No
Alzheimer	Yes	No
Stroke	Yes	Yes
Movement problems	Yes	No
Obesity	Yes	Yes

A. Motivation

Considering the real social and market demands and the needs of the health care service provider segment [5], the general project aim is to improve the quality and cost effectiveness of health care services by developing service models, methods, tools, products and services. We are doing research and development of a full scale remote telemonitoring system that monitor both activity levels and vital signs such as blood pressure, blood sugar level and heart rate, alerting caregivers about potential health problems or emergency situations.

Service categories of a generic home-based care/remote patient monitoring solution

Remote patient monitoring builds up from the following five evitable service pillars:

- Data acquisition services

DAQ services collect physiological information of a person's condition from deployed sensor infrastructure or from the person directly.

- Store/forward and visualize services

Services to store, process and visualize locally the captured physiological information of a person's condition at the patient's device and to forward these information using ICT towards the central data collector node for further data processing, storing, visualization.

- Activity recognition services

Recognition of psychophysical performance of patients or elderly people for effective therapy intervention (quantitative and qualitative measurement of body movement).

- Behavior monitoring services

Recognition of psychophysical performance of patients or elderly people for effective therapy intervention (quantitative and qualitative measurement of high level living patterns).

- Lifestyle guidance services (knowledge transfer)

Feedback from medical experts/physicians based on the acquired real-time or historical data enables lifestyle guidance, therapy adjustment, early warning/prevention, personalized health care & rehabilitation.

II. LIVING LAB INFRASTRUCTURE

Main novelties of the AALAMSRK project are that it brings into the patient's home medical assistance and lifestyle guidance services and also it supports new potential opportunities to capture insight medical knowledge with its effective non-stop health monitoring methods. The monitoring is done by our clinical trial-ready, standardized distributed monitoring and testing environments (so called Living Labs). Our Living Labs are supporting all the R&D tasks of the medical, engineering and business (marketing) work packages and also provides evaluation and test environment for new hypotheses and results.

The established Living Lab environments are located in three different regions within Hungary (capital city, middle size city and rural area), thus the type of patient environments (living space size, accessibility, communication infrastructure, etc.) are totally inhomogeneous [9]. During the test periods -thanks to the Living Lab studies-, we have learned a lot about the different environment requirements, and revealed many aspects of various problems concerning sustainability, usability of our health monitoring and emergency management solution.

III. FIXED (HOMEHUB) VS. PORTABLE (MOBILEHUB) MONITORING DEVICES

During the AALAMSRK project the PC based HomeHub was firstly developed. The HomeHub framework supports Linux and able to run both on proprietary solutions (like Intel HealthGuide) and on out-of-the-box commodity PCs optionally equipped with touch screen. Beside the HomeHub solution, later on we have developed an Android based MobileHub (shown in Fig. 2.) solution as well, which enables additional usage scenarios due to its different (mobile) characteristics.

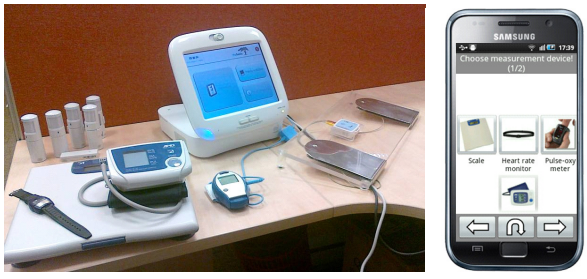


Figure 2. HomeHub (on left) with connected sensors and GUI of the Android based MobileHub (right)

The MobileHub has many attractive features (cheaper price, portable, location awareness, inbuilt touch screen,

etc.), however on the other side it has also significant limitations compared to a full PC hardware (limited CPU power, memory, storage size and external interface connection support, etc.). The MobileHub is targeting different functionalities compared to the HomeHub solution (due to the smaller screen size and fewer hardware interfaces), and it can extend the usability with additional special features, such as mobility, location awareness and small size. Comparison with example hardware configurations is shown in Table 2.

Table 2. Basic parameters of the two DAQ solutions

Parameters	PC based HomeHub	Android based MobileHub
Platform/OS	Intel Health Guide[6]/Linux	Samsung Galaxy S[7]/Android
Size (WxHxD)	280 mm x 90 mm x 270 mm)	64.2mm x 122.4 mm x 9.9 mm
Weight	3820 g	118 g
GUI	10", color touch screen, LCD	4", color touch screen, AMOLED
System help, comm. support	Audio+video	Audio+video
Remote mgmt.	Yes, secure	Yes, but with limited security

We need to note here, that both HomeHub and MobileHub are capable to run on various other hardware configurations. The HomeHub was tested on a large set of (Linux installed and touch screen equipped) PCs with success. According to our tests the MobileHub software is capable to run almost all Bluetooth enabled and Android 1.2+ based Smartphone. We should also note here that even if both solutions working non-stop 7/24, the PC based HomeHub consumes far more power, than the Android based MobileHub. According to our tests the usage of the PC based HomeHub solution requires significant amount of electricity, and the increased electricity costs can become a limiting factor at the number of potential users.

A. Actual sensor set of our health monitoring system

Sensors can provide digital fingerprint of the patient's/person's psychophysical status and performance. We categorize sensors according to their data collection methods: active or passive type sensors. In our solution we are using a wide range of sensors (both passive and active, shown in Table III.) to collect detailed and accurate information about the patient remotely about:

- patient's blood glucose/sugar level
- patient's actual pulse, and oxygenization level
- patient's ECG signals
- patient's blood pressure
- patient movement within the house, with the usage of wall mounted sensors
- patient medication, with sensors of the medicament dispenser
- patient's eating habit, with sensors placed on the refrigerator
- patient activity with so called "Actigraph", which is a watch like sensor on the patient's wrist. Actimetry sensor can measure non-invasively rest/activity cycles, gross motor activity.
- patient's weight
- environmental information, about room temperature, humidity etc.

B. MobileHub basic software features

- Health status visualization.
- Multi-language support (English, German, Italian, and Hungarian).
- Silent sensor DAQ mode (automatic data collection) via Bluetooth.
- Health status visualization (with statistics and data mining facility) at the data center.
- Manual DAQ mode (optionally GUI initiated sensor data collection) only via Bluetooth.
- Limited data compression and encryption during data transmission towards the data center.
- Automatic sensor data pre-evaluation at the MobileHub.
- Full featured location aware mobile emergency alarm.
- Real-time sensor data forwarding to the central data collector server
- Store and forward mode to enable offline data collection

In Table 3. round brackets contain the sensor vendor's name and the connection interface type of the sensors is shown in between slashes (e.g.: BT=Bluetooth).

Table 3. Sensor compatibility of the developed health monitoring infrastructure

Sensors and parameters	PC based HomeHub	Android based MobileHub
Pulse oximeter	No	Yes (NONIN) /BT/
Blood pressure monitor	Yes (AND) /USB + BT/	Yes (AND) /BT/ Yes (MEDITECH) /BT/
Weight	Yes (AND)	Yes (AND) /BT/
Pulse meter		Yes (ZEPHYR) /BT/ Yes (POLAR) /BT/
Activity Sensor/Actigraph	Yes (MEDITECH) /BT/	Yes, in-built sensor, used with external software module
In-house movement sensor	Yes (GE Healthcare - QuietCare) /Zigbee/	No
Medication	Yes (GE Healthcare - QuietCare) /Zigbee/	No
Environment monitoring	Yes (GE Healthcare - QuietCare) /Zigbee/	No
Blood glucose meter	Yes (BAYER) /USB/	Yes (MYGLUCOHEALTH) /BT/
ECG	Yes (MEDITECH) /BT/	No
Cognitive tests	Yes , as software module	No
Diet Log	Yes , as software module	Yes, as external software
Emergency Alarm	Yes , as software module	Yes , as software module
Location aware alarm	Yes , as software module (fix)	Yes , as software module
Two-ways communication	Yes , as software module	Yes , as software module
Location tracking	No	Yes , as software module

The MobileHub and HomeHub software are designed to be suitable for different skilled users. Both user interfaces are highly configurable to support elderly persons (high contrast, huge characters, simple UI). For the HomeHub a large GUI was designed with huge buttons and characters to make touch screen usage easy (shown in Fig. 3.).

Figure 3. User interface of the HomeHub

Based on the test results received from the Living Lab experiments- due to the small touch screen-, on the MobileHub multiple separated GUI have been developed for the differently skilled elderly persons. To overcome all the usage limitation problems, we have identified two user skill sets, which can categorize the hardware/software utilization ability level of the patient:

- GUI for elderly persons without any IT knowledge (oversimplified single button type GUI)
- GUI for normal and expert end-users the panic button is just an icon and all the in-build additional software/hardware functionalities of a normal Android mobile phone are available (menu sets, SMS, dialing, applications, etc.). Patient can manually initiate sensor management and can manually provide measurement inputs on the GUI (shown in Fig. 5. and Fig.6.).

Figure 4. Screens of the MobileHub for normal/expert users

C. *MobileHub additional software functionalities*

1) *Location and sudden event monitoring service*

In a sudden panic situation an alarm can be activated manually (by the patient) or automatically (by e.g. the accelerometer) with the mobile device. When an alarm signal initiated the central dispatcher is able to acquire location information (based on GPS and GSM/GPRS cell information) immediately. The automatically established two way voice communication can help to understand the context of the sudden event. The MobileHub can work as a location independent, always-on personal emergency notification device.

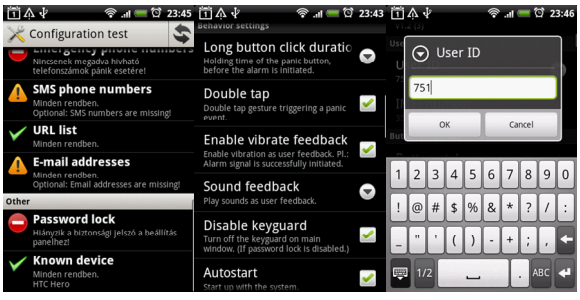


Figure 5. Configuration GUI screens (administrator section)

2) *Diary / reminder services*

Besides pure data collection our MobileHub enables to do correlation analysis of stored data series. Reminders are intended to notify user in pre-defined appointments about data collection needs. This feature enables to track data in certain periods like measuring patient's blood glucose level on a regular basis.

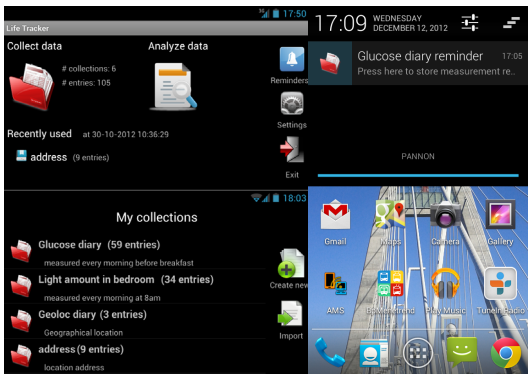


Figure 6. GUIs of the diary and reminder services

IV. CONCLUSIONS

During our almost 4 years long development period both the fixed and portable solutions have been rigorously tested in the LivingLab environment. Beside patient monitoring we had to monitor remotely not only the patient's status, but also some mobile hardware and software specific parameters (such as: battery level of sensors), and we had also to redesigned the whole user interface of the handheld device to support elderly persons with low IT skills. According to the received result both our PC and Android based DAQ solutions are capable to provide seamless remote monitoring of elderly persons not only at home, but with MobileHub also abroad. The developed solutions provide important feedbacks about health status to the patient and to the medical experts.

V. ACKNOWLEDGMENTS

The projects have been supported by the European Union. The authors would like to thank GOP-1.1.1-11-2012-0055- Dialogic „DIALOGIC – Mathematical model based decision support system for diabetes monitoring"" for its financial support. This work makes use some of the results produced by the AALAMSRK project (Research & development of assisted living patterns and their integration into decision support system to promote the quality of life at home), Hungarian National Technology Programme, A1, Life sciences, the "Development of integrated virtual microscopy technologies and reagents for diagnosing, therapeutical prediction and preventive screening of colon cancer "Hungarian National Technology Programme, A1, Life sciences, (3dhist08) project and the ÓE-RH 1104/2-2011 project. Authors would like to thank for their financial support hereby.

REFERENCES

[1] MyFitnessCompanion, http://www.myfitnesscompanion.com/features.html [Online, acc. 02.02.2012].

[2] Ziyu Lv, Feng Xia, Guowei Wu, Lin Yao, Zhikui Chen, iCare: A Mobile Health Monitoring System for the Elderly, The 3rd IEEE/ACM Int Conf on Cyber, Physical and Social Computing (CPSCom), IEEE, Hangzhou, China, December 18-20, 2010, arXiv:1011.3852v1

[3] Microsoft HealthVault, http://www.microsoft.com/en-us/healthvault/ [Online, acc. 01.02.2012].

[4] Proseniis- Innovation for healthier senior Age, http://www.proseniis.hu/ [Online acc. 05.01.2012].

[5] Anders Wimo, Linus Jönsson, Anders Gustavsson; Cost of illness and burden of dementia in Europe - Prognosis to 2030, 27 October 2009, http://www.alzheimer-europe.org/Our-Research/European-Collaboration-on-Dementia/Cost-of-dementia/Prognosis-to-2030

[6] Intel® Health Guide PHS6000, Product Brief, ftp://download.intel.com/healthcare/pdf/Health_Guide_Brief.pdf

[7] Galaxy S specification, http://www.samsung.com/uk/consumer/mobile-devices/smartphones/android/GT-I9000HKDXEU-spec

[8] M. Kozlovszky, Zs. Meixner, G. Windisch, J. Márton, S. Ács, P. Bogdanov, A. Boruzs, P. Kotcauer, J. Ferenczi, V. Kozlovszky; Network and service management and diagnostics solution of a remote patient monitoring system, Lindi 2011 Conference, Budapest, Hungary

[9] M. Kozlovszky, J. Sicz-Mesziár, J. Ferenczi, J. Márton, G. Windisch, V. Kozlovszky, P. Kotcauer, A. Boruzs, P. Bogdanov, Zs. Meixner, K. Karóczkai, and S. Ács,Combined Health Monitoring and Emergency Management through Android Based Mobile Device for Elderly People; K.S. Nikita et al. (Eds.): MobiHealth 2011, LNICST 83, pp. 268–274, 2012. Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2012

Signal processing methods for ST variability assessment in ECG

A. Rashkovska and V. Avbelj
Jožef Stefan Institute, Department of Communication Systems, Ljubljana, Slovenia
Aleksandra.Rashkovska@ijs.si, Viktor.Avbelj@ijs.si

Abstract - The beat-to-beat ST variability in the ECG signal is becoming an important indicator in neurocardiology. However, accurate determination of the ST variability is difficult because of uncertainty in the determination of the end point of the ST interval. T waves change their form because of breathing, heart movements, changing of lead positions etc. If the ST variability is small, as in most neurological patients, their reliable assessment is even more difficult. In this paper, we propose two methods for the assessment of the ST variability: the beat-to-beat variability of the RT interval (the time between the peak of the R wave and the peak of the T wave) and the TTs interval (the time between the peak of the T wave and the T wave point with maximal negative slope). The paper elaborates in details the method for the determination of Ts. The method is analyzed through its noise sensitivity estimation.

I. INTRODUCTION

Electrocardiogram (ECG) is a diagnostic tool in cardiology that monitors the electrical heart activity on the body surface resulting from differences induced by action potentials of the cells in the muscular tissue of the heart. Every phase of heart activity cycle or cardiac cycle (or known as heartbeat) results in a separate waveform in the ECG (Fig. 1a). These waves are result of non-homogeneities of action potentials throughout the heart, related to the start-time of depolarization and the time-course of the action potential in each cell. The waves Q, R and S are manifestation of the depolarization in the two ventricles, while the waves T and U are consequence of the repolarization process in the ventricles. Large differences of action potential courses during repolarization (dispersion of repolarization) can lead to arrhythmias, especially when these differences arise at small distances. Whether the dispersion is present at small distance through the heart muscle wall (transmural dispersion), from base-to-apex, or from one ventricle to the other ventricle, has been a subject of long on-going debate [1, 2].

The repolarization variability in the heart can be measured with the variability of the ST interval defined between the J point (the point at which the QRS complex finishes) and the end of the T wave (Fig. 1a). The ST variability has been assessed with different methods in previous studies [3-7]. As the start of the ST interval (the J point) is not well defined, replacement intervals are often used in assessing the beat-to-beat variability of the repolarization i.e. the interval between the peak of the R wave and the peak of the T wave (RT interval) complemented with the interval from the peak of the T wave to the end of the T wave [3-5]. R waves are used instead of S waves, because the difference between the R and the S peaks is typically small (< 50 ms) and constant. Moreover, R peaks are present on all ECGs with clearly defined peaks. The ST interval can be obtained from the RT interval by subtracting the (S-peak - R-peak) time determined separately. Unfortunately, the end of the T-wave is not a well-defined instant, and, in addition, the presence of a U-wave can further complicate the determination of a ST interval, according to the definition given in [6].

The authors acknowledge the financial support from the state budget by the Slovenian Research Agency under Grant P2-0095.

In our previous study [7] we have presented a method for the determination of the RT variability. In this paper, we present additional method to complement the one in [7] in ST variability assessment. So, the two methods that we propose for the assessment of the ST variability are: the beat-to-beat variability of the RT interval and beat-to-beat variability of the TTs interval (Fig. 1b). The second interval, TTs, is defined as the interval between the peak of the T wave and the T wave point with maximal negative slope. The paper elaborates the method for the determination of the maximal negative slope on the T wave through its noise sensitivity estimation. Additionally, it presents the results for the TTs variability of a measured ECG signal for selected cases of healthy subject and HTX subject (patient with transplanted heart).

II. METHODS

The RT and TTs intervals are defined with three points on the ECG signal: the R peak, the T peak and the maximal negative slope on the T wave. The methods for determination of these points are developed and integrated into a custom software, named NeuroECG, for analysis of biomedical signals (like ECG, breathing). NeuroECG incorporates also a measurement module developed for on-line recording of biomedical signals. There are two types of data streams (channels) in NeuroECG:

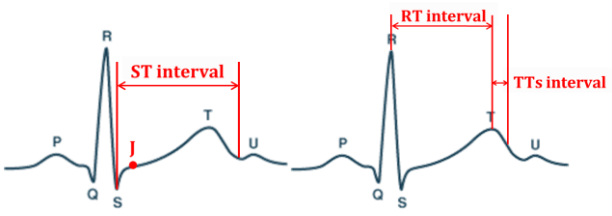


Figure 1. (a) ECG waveforms with marked ST interval. (b) Marked RT and TTs intervals for the assesment of the ST interval.

- Measured (continuous) channels like ECG, blood pressure, breathing, characterized by sampling rate. In particular, the ECG measurements that we work with are with high resolution (1000 – 1800 Hz) and duration of 520 seconds.

- Event channels, an innovative approach of NeuroECG, which enable a more abstract view on the analysis of biomedical signals. An event channel is generated from a measured channel based on different events. The events that we are interested in: the R peak, the T peak and the maximal negative slope on the T wave, are generated from the ECG measured channel. Each event channel is characterized by its time and value. For example, in determination of the R peak, the time of each event in the channel is the time when R peak is determined, and the value is the distance in time to the previous R peak.

What we have added is an integrated procedure build in NeuroECG which determines all important T-wave times automatically, i.e. three event channels with T-time (maximal T-wave amplitude), Ts-time (maximal negative slope), and Te-time (intersection of the maximal negative slope line with the baseline).

The methods for determination of the R peak and the T peak have been described and evaluated in [7]. Following in this section is the description of the method for determination of the maximal negative slope on the T wave.

A. Method for determination of the maximal negative T wave slope

The method for determination of the maximal negative T wave slope is graphically presented in Fig. 2 on a T-wave retrieved with down sampling of a real ECG signal. We show every 10-th sample to obtain a hypothetical sampling rate of 100 Hz for easier description of the proposed method. Before applying the method, a baseline correction and a low pass filtering is performed. These pre-steps are integrated in the procedure and are not analyzed in this paper.

The main approach that we use in the method is to perform linear least squares approximation between a point and its +/- N neighbors, determine the slope of the approximation line in each point, and assign the value of the slope to the point.

The parameters that this procedure takes as an input are:

- The minimal and maximal RT time (default: 0.1 s, 0.45 s),

- The number of neighboring points N for the approximation to determine the T-wave maximal negative slope,

- The frequency of the low pass filter (in Hz).

For demonstration purpose, we have set the number of neighboring points N to +/- 2. So, for each point and its left and right 2 neighbors, we perform a linear least squares approximation and determine the slope, i.e. the derivative of the approximation line as the ratio between delta (y) and delta (t). After we perform the same step for all points, we get the discrete derivative (Fig. 2a). Next, we find the minimal amplitude among the derivative points (approximate Ts point) and its two neighboring points (Fig. 2b). Then, we improve the resolution with quadratic interpolation using these tree points. Finally, we determine the minimum on the interpolation curve and declare that point as a point with the maximal negative slope (Fig. 2c).

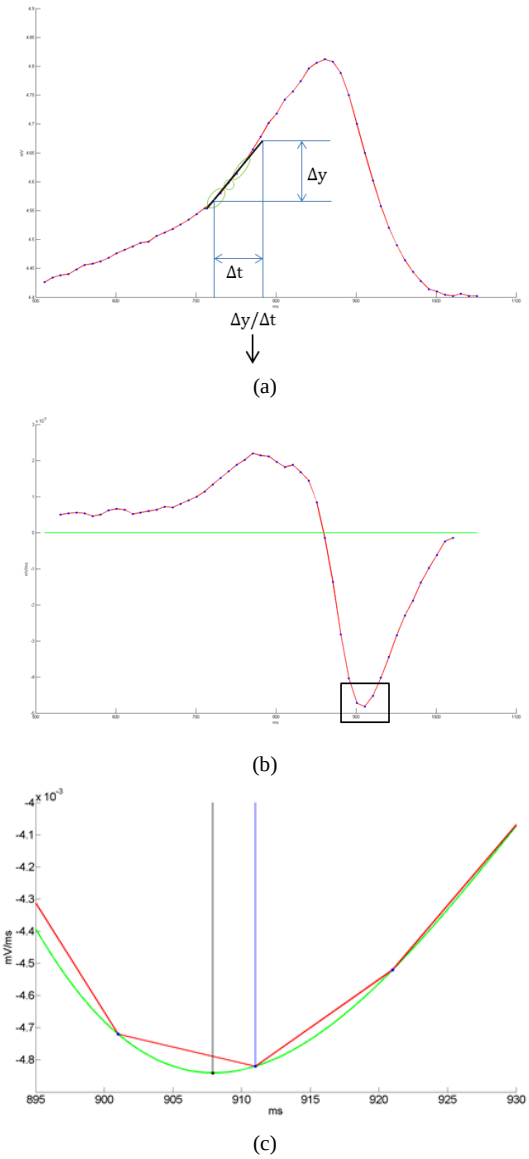


Figure 2. Graphical representation of the method for the determenation of the maximal negative slope on the T wave. (a) Downsampled T-wave from a real ECG (sampling rate 100 Hz) and the least squares approximation procedure for determinating the slope of the approximation line. (b) Discrete derivative curve (red). (c) Approximate Ts point (blue line) and quadratic interpolation (green) of the discrete derivative for improving the resolution of the Ts time (black line).

B. *Noise sensitivity estimation of the methods*

To estimate the noise sensitivity of the method for determination of the maximal negative slope on the T wave, we define an artificial T-wave as an asymmetric slope cosine function [8]. The artificial T-wave is defined as a function of time with the following equation:

$$T=\frac{\cos(\pi(t/a))+1}{2}\quad\begin{array}{lll} a=0.16 & for & -0.16<t<0 \\ a=0.08 & for & 0\le t<0.08 \\ T=0 & for & 0.08\le t\le -0.16 \end{array} \qquad (1)$$

For the artificial simulated T-wave we can analytically determine the Ts points and use it for evaluating the timing error after we add noise to the signal and apply the method.

We tested two types of noise:

- Gaussian noise with signal-to-noise ratio (SNR) ranging from 1 to 150. For each SNR, 100 runs were performed. The SD of the timing errors and its maximal value were calculated.

- Sinusoidal noise with frequency in the range from 0 to 100 Hz, and amplitude 10%, 1% and 0.1% of the one of the simulated T-wave (equivalent to SNR of 20, 40 and 60 dB, respectively). After 100 simulation runs on a single frequency, the SD of the timing errors and its maximal value were calculated. In each simulation run, only the sinusoid phase was changed randomly.

Following our previous analysis, the parameter N (neighboring points for the least squares approximation) was set to 20 and the frequency of the low pass filter to 40 Hz.

III. Results and Discussion

A. *Noise sensitivity estimation results*

The dependence of the timing error on the SNR when the T wave is disturbed with Gaussian noise is shown in Fig. 3. With improved SNR, the timing errors become smaller. For SNR greater than approximately 40 dB, the timing error becomes less than 1 ms. Great disturbances can be observed for lower SNR, indicating instability of the method for low SNR values.

The dependence of the resulting timing error on the frequency of a disturbing sinusoidal noise is shown in Fig. 4. Local minimums can be observed at particular frequencies. This result can be beneficial when selecting the parameters of the method (the N parameter and the frequency of the low pass filter). Large timing errors can be observed for low frequency noises for smaller SNR; however, for larger SNR, the error for low frequency noises becomes smaller (even smaller than 1 ms for SNR of 60 dB). The timing error becomes low for high frequency noises (> 60 Hz).

Compared to the results for the noise sensitivity of the method for determination of the T peak in [7], the method for the determination of the Ts point is more sensitive to noise than the one for T peak.

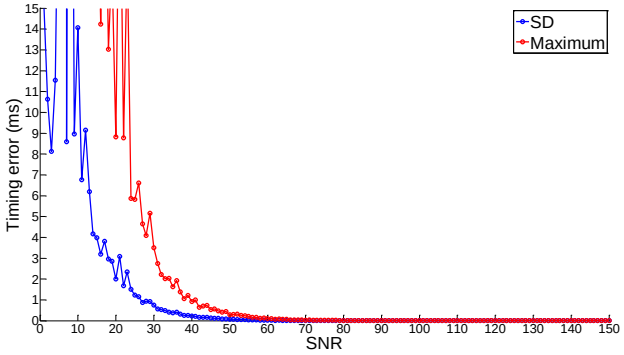
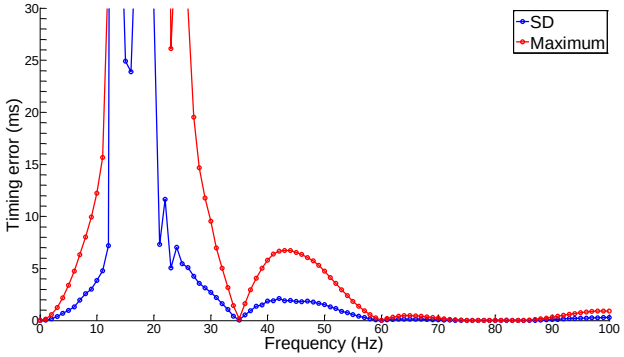
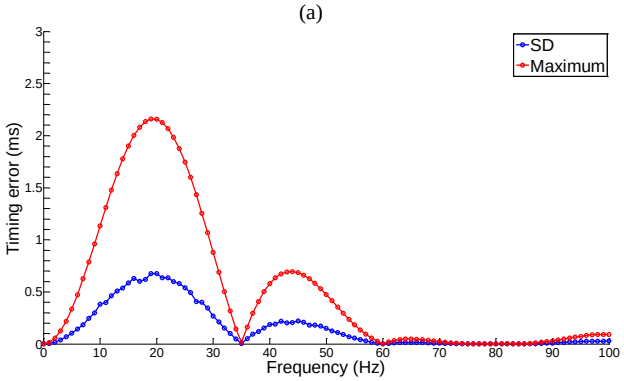


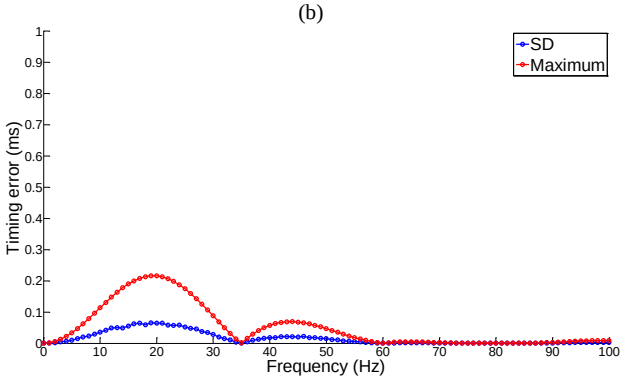
Figure 3. Timing error of the method for the determination of the maximal negative slope on the T wave as a function of the signal-to-noise ratio (SNR) when disturbed with Gaussian noise.



(a)

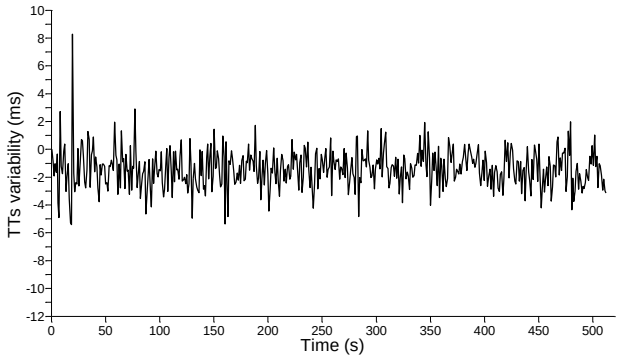


(b)

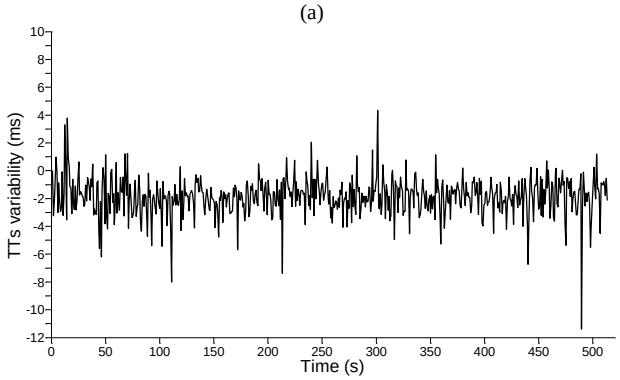


(c)

Figure 4. Timing error of the method for the determination of the maximal negative slope on the T wave as a function of the frequency of a disturbing sinusoidal noise with amplitude (a) 10% of the T wave amplitude, equivalent to SNR = 20 dB. (b) 1% of the T wave amplitude, equivalent to SNR = 40 dB. (c) 0.1% of the T wave amplitude, equivalent to SNR = 60 dB.



(a)



(b)

Figure 5. TTs variability for a selected case of (a) Healthy subject. After linear detrend of the TTs data, the TTs variability is 1.4 ± 1.4 ms. (b) HTX subject (patient with translanted hearth). After linear detrend of the TTs data, the TTs variability is 1.9 ± 1.4 ms.

B. Measured TTs variability

Using the method for determination of the maximal negative slope on the T wave, we calculated the TTs variability for selected cases of a healthy subject and a HTX subject (patient with transplanted heart). The results are shown in Fig. 5. We measured TTs variability of 1.4 ± 1.4 ms for the health subject and TTs variability of 1.9 ± 1.4 ms for the HTX patient. The TTs variability for the case of a healthy subject is in the same range as the QT variability measured in [7].

The SNR for high frequency noise, before assessment of the TTs variability (after filtering), was > 50 dB for both cases. Low frequency noise, after the removal of base line wondering, had frequencies < 1Hz and consequently, according to the noise sensitivity estimation results, the timing errors was small. For both cases, the low frequency noise was < 10%.

IV. CONLUSION

In this study, we have demonstrated a method for determination of a characteristic point on the T wave that can further complement the assessment of the beat-to-beat repolarization variability: the maximal negative slope on the T wave.

We have analyzed the method through its sensitivity on a Gaussian and sinusoidal noise. The results show that for sufficient SNR, the timing errors of the method are small enough. Large timing error is expected in low frequency noises (movement artifacts). The results for the measured variability of the interval between the peak of the T wave and the point on the T wave with the maximal negative slope (TTs interval) in two selected cases of a healthy subject and a HTX patient show that the TTs variability in those cases is in the same range. Clinical discussion of this result would be a subject of further work. Moreover, a study of the TTs variability in a group of health subjects and a group of HTX patients is also planned as future work. The study will assess the TTs variability in these groups, and, furthermore, eventual coherence with breathing/RT interval will be determined.

REFERENCES

[1] C. Antzelevitch et al., "Does Tpeak-Tend Provide an Index of Transmural Dispersion of Repolarization?", Heart Rhythm., vol. 4, pp. 1114–1119, August 2007.

[2] T. Opthof et al., "Dispersion of repolarization in canine ventricle and the electrocardiographic T wave: Tp-e interval does not reflect transmural dispersion", Heart Rhythm., vol. 4, pp. 341–348, March 2007.

[3] A. Porta et al., "Performance assessment of standard algorithms for dynamic R-T interval measurement: comparison between R-T_{apex} and R-T_{end} approach", Med. Biol. Eng. Comput., vol. 36, pp. 35–42, January 1998.

[4] R. Wolk, T. Mazurek, T. Lusawa, W. Wasek, and J. Rezler, "Left ventricular hypertrophy increases transepicardial dispersion of repolarisation in hypertensive patients: a differential effect on QTpeak and QTend dispersion", Eur. J. Clin. Invest., vol. 31, pp. 563-569, July 2001.

[5] V. K. Yeragani et al., "Relationship between beat-to-beat variability of RT-peak and RT-end intervals in normal controls, patients with anxiety, and patients with cardiovascular disease", Ann. Noninvasive Electrocardiol., vol. 12, pp. 203-209, July 2007.

[6] R. S. Macleod, R. L. Lux, M. S. Fuller, and B. Taccardi, "Evaluation of novel measurement methods for detecting heterogeneous repolarization", J. Electrocardiol., vol. 29, pp. 145-153, 1996.

[7] V. Avbelj, R. Trobec, and B. Gersak, "Beat-to-beat repolarisation variability in body surface electrocardiograms", Med. Biol. Eng. Comput., vol. 41, pp. 556-560, September 2003.

[8] P. Langley, D. D. Bernardo, and A. Murray, "Quantification of T-wave shape changes following exercise", Pacing Clin. Electrophysiol., vol. 25, pp. 1230-1234, August 2002.

Coarse grain molecular dynamics study of voids present in the membrane of a lipid vesicle

J. I. Pavlič
Institute Jožef Stefan/Department of Communication Systems, Ljubljana, Slovenia
janez.pavlic@ijs.si

Abstract – By using molecular dynamics (MD), we study the course of behavior and constitution of lipid vesicle in order to be able to recreate the measured results obtained in previous experiments. Simulated lipid vesicles are composed of thousands of DSPC lipid molecules. All preformed vesicles have a mean radius of around nine nanometers and interact with the watery environment. We use coarse grain molecular dynamics (CG-MD) simulation method. The CG-MD simulations are run, on multicore parallel computer, up to 30 ns of simulated time. We show that created vesicles remain stable throughout the whole simulated time, but possess an intrinsic defect of having voids in between the two monolayers, probably due to not taking into account different effective sizes of DSPC lipid molecules in the two monolayers.

I. INTRODUCTION

Today's modern computers enable complex *in silico* studies of biological nanostructures. One of the important application fields in biomedicine is the simulation of proteins, which can help to understand the nature of their interactions with lipid nanostructures. The studies can help to improve and accelerate new drug design [1, 2]. Lipid bilayers have often been used as the simplest model of cell's membrane for the investigation of basic phenomena, such as membrane – protein interactions, mechanical and structural properties or conformational changes of membranes [3, 4, 5, 6]. Lipids in watery environment form bilayers, and bilayers regularly form membrane patches. Membrane patches possess high edge energy. To reduce the membrane's edge energy, the membrane patch prefers to close on itself, thus forming a vesicle, which is the simplest model representing a living cell. The vesicles are composed of lipids, which are arranged in a closed bubble-like shaped bilayer [7].

The lipid molecules are amphiphilic with a polar part and a nonpolar part. The polar part (head) is hydrophilic, i.e., it "likes" water, which is a polar solvent. The non-polar part (tail) is hydrophobic, which means that it does not "like" water and repels it. In our studies we use Distearoyl – Sn – glycerol – PhosphoCholine (DSPC) zwitterionic lipid species, which is shown in Fig. 1. The zwitterionic lipid headgroup is a dipole, thus the lipid itself is electrically neutral.

The author acknowledge the financial support from the state budget by the Slovenian Research Agency under grant P2-0095.

In this paper, the study is focused on preformed lipid vesicles, and how their lipid membranes evolve over the simulation time. We observe voids, which are present in between the two monolayers composing the lipid membrane of vesicles.

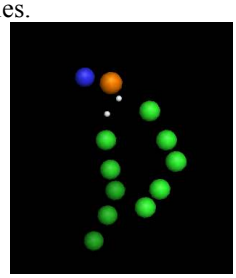


Fig 1: CG model of a single molecule of DSPC zwitterionic lipid. The polar head is represented by blue and orange sphere, and the nonpolar tail is represented by green spheres.

The rest of the paper is organized as follows: In Section II we shortly discuss the background and the methodology of molecular dynamics simulations. Next section is devoted to simulation results of lipid vesicles in watery environment. The paper concludes with a short summary and outlooks for the future work.

II. METHODS AND SYSTEMS

A. Molecular dynamics simulations

The molecular dynamics (MD) simulation [8] is based on calculation of particle – particle force interactions in a multi-particle system. The Newton's equation of motion has to be integrated for every particle of the system in order to obtain new positions and new velocities for each time step. The full atomic simulation is limited to short simulation intervals (few nanoseconds) and small number of particles due to significant computational requirements [9, 10, 11]. In our study, we used the coarse grain (CG) MD simulation [12], which is appropriate for bigger simulation systems and able to perform longer simulation periods (up to a few microseconds). In the CG-MD simulation we apply the well known MARTINI force field (ff) [13]. The CG-MD simulations were performed by using open source software, called GROMACS [14]. All tests were run on our computing cluster at Institute Jožef Stefan (IJS), Slovenia [15]. It consists of 37 computing nodes. All nodes run under Linux operating system and are interconnected by Gigabit Ethernet links.

We ran several simulations on computer nodes in parallel; trying out different parameter sets (e.g. lipid molecule's cross-section size, radius, simulation specific settings, etc.). In the following, we show our typical study of simulation of preformed lipid vesicles in watery environment.

B. *Molecular System*

Molecular systems of interest consist of DSPC lipids and water molecules. The spherical shell of a lipid vesicle is composed of different numbers of DSPC lipids; ranging from approximately 2800 up to 3800, and roughly 100000 water molecules in simulation box of approximate size of 24 nm^3. All the molecules are modeled by the MARTINI ff, which merges four heavy atoms into a single bigger CG particle [11]. In such a way, the number of particles, constituting the molecular system, is greatly reduced. This enables us to simulate larger systems of interest for longer simulation times.

Before we start the MD simulation a starting structure of lipid molecules, usually a vesicle without water molecules, is shaped with a dedicated particle-placing program. After having the initial structure, the energy minimisation (EM) in vacuum is carried out, which adapts initial positions of particles to the desired energy. The minimised lipid structure is already placed in the simulation box. The empty volume of the simulation box is filled up with water molecules. The whole system (lipid and water molecules) is energy minimised in vacuum, again. The whole system is exposed to the final MD simulation run (MD run). All presented MD simulations are performed for 30 ns time interval. In all MD runs, we use Parrinello-Rahman barostat at 1 bar with pressure constant τ_p set at 3 ps and Nose-Hoover thermostat at 323 K with temperature constant τ_t set at 3 ps. Periodic boundary conditions are applied and time step is set to 20 fs.

III. Results and Discussion

While running MD simulations, all pieces of information about atomic coordinates are being stored in a trajectory file. These coordinates can be used for animated visual representations. Variety of plots, with respect to time or coordinates, can be obtained for simulated molecular ensembles (temperature, pressure, radius of gyration, energy, force, root mean square distance, density profile, etc.).

From the beginning of our research, we have started with a random distribution of lipid and water molecules and hoped that they will spontaneously form into lipid bilayers later to membrane patches and finally into vesicles. However, we found out that such a process is very sensitive to initial molecular distributions and velocities, and to other simulation parameters, e.g. pressure, interaction models, etc. Up to now, we have not been able to form a lipid vesicle from a starting structure with randomly distributed lipids, probably because of relatively short simulated intervals. Instead, we place lipids into desired shapes (membrane, cylindrically shaped or vesicle) with our custom program. Such a procedure provides desired shapes of starting structures, and allows us to spend much less simulation time for the formation of the final structures. The GROMACS software tools do not offer an option for the designing and manipulating initial custom shapes.

Our final goal is to simulate more complex cases such as interactions among vesicles of different sizes and of different compositions, foreseen some future applications, e.g. in the circulating blood where many various kinds and sizes of spherical objects are present, such as leukocytes, platelets, microvesicles, etc.

A. *Lipid vesicle*

The simulation results of lipid vesicle composed of DSPC lipid molecules is shown in Fig. 1. The mean radius of the lipid vesicle is approximately 9,20 nm, and bilayer thickness is approximately 4,45 nm, which is slightly more than the twice of the length of DSPC lipid molecule. The starting structure of the vesicle is composed of 3311 DSPC lipids and 98147 waters (not shown for better visibility of lipids) as shown in Fig 1A. The pore on one of the poles closes after EM in vacuum with no water present in the simulation box as shown in Fig 1B. EM in vacuum, when waters are present, the spherical structure relaxes even more as shown in Fig 1C (again without water molecules). Through the whole MD run the vesicle's shape stays more or less spherical. Spherical membrane does not fall apart, and pores through the bilayer do not appear after 30 ns simulation time as can be seen in Fig 1D, again without water molecules.

In all of the cases we have studied, voids remain present in membranes of our vesicles. Our hypothesis about possible reasons are; first, the system could be over energy minimised (maximal force is set below 100 to 200 kJ mol^{-1} nm^{-1}) or/and second, we do not take into account different effective sizes of lipid molecules in inner and outer lipid monolayer due to high curvature at small radius [16, 17, 18]. During the EM, the lipids are pushed apart to achieve low potential energy due to smaller interaction forces among neighboring lipid molecules. Later on, when water is added, it also occupies the empty spaces (voids) in between the monolayers. To take into account the different effective size of lipid molecules, the starting structure should have few intentionally present pores, where the lipids can exchange between the two leaflets during MD run [17, 18] to release the lateral pressure in the membrane.

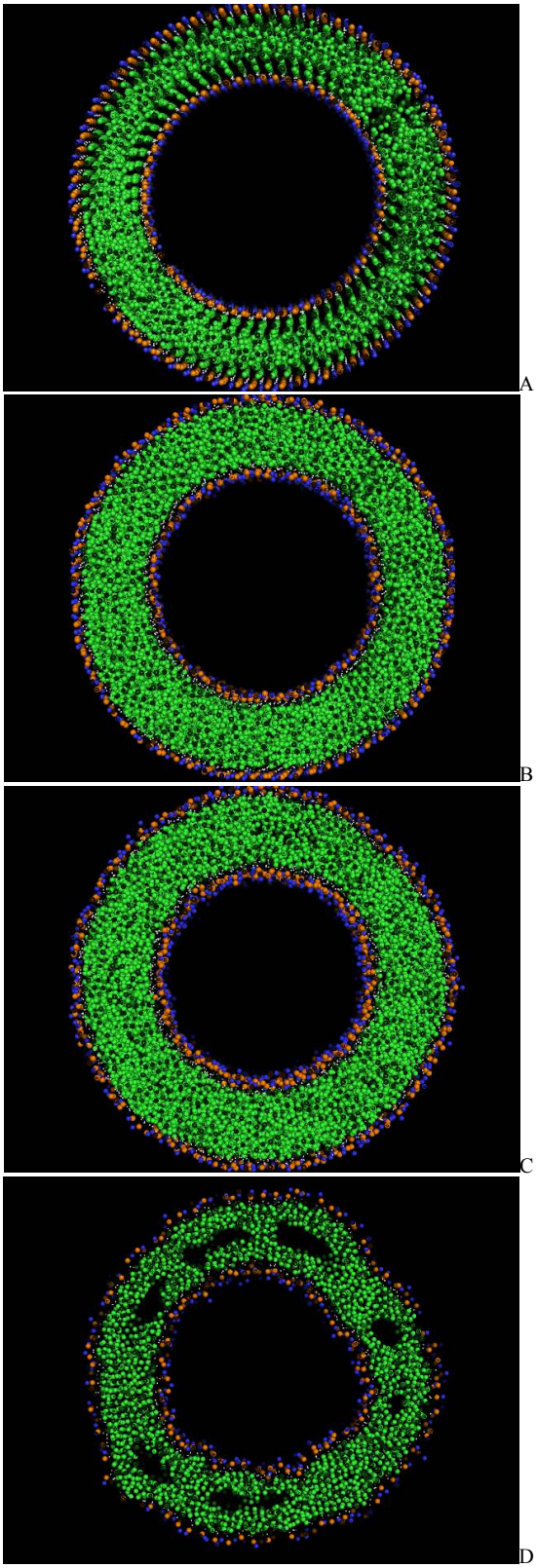
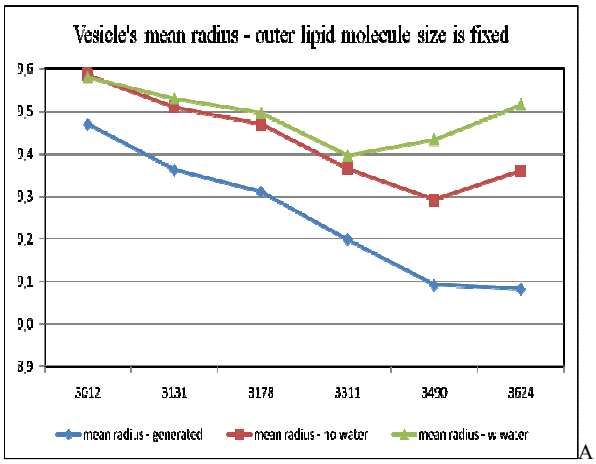


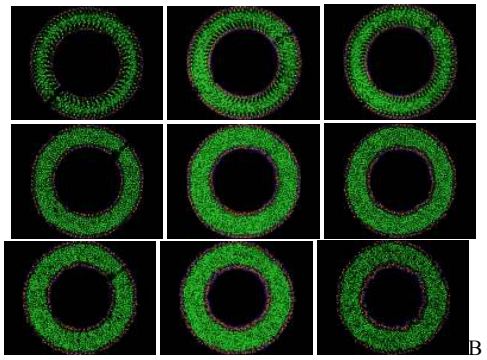
Fig 1: A) Starting structure of the lipid vesicle composed of 3311 DSPC lipids. The pore on the pole is clearly visible. B) EM of DSPC lipid vesicle in vacuum. The pore closes up completely. C) EM of DSPC lipid vesicle and water. Water molecules are not shown for better visibility of vesicle's cross-section. D) Vesicle after 30 ns of MD simulation. Voids in between the two monolayers can be clearly seen. The voids persist throughout the MD simulation; even after 100 ns.

Fig 2 and Fig 3 show how the mean radius of the vesicle changes with respect to different sizes between inner and outer DSPC lipids. Fig 2A shows the case when the size of outer lipids is fixed ($0,9A_0$; A_0=0,654 nm^2), while we change the size of inner lipids [$1,1A_0$, $1,05A_0$, A_0, $0,95A_0$, $0,88A_0$, $0,85A_0$]. Note, that we change the effective size of lipids by changing their density, i.e. their number in the inner or outer monolayer. The total number of DSPC lipids constituting vesicle's membrane is related to lipid size and is as follows {3012, 3131, 3178, 3311, 3490, 3624}. Fig 2B shows visual representation of vesicle's cross-section during the workflow for performing MD simulations.

Fig 3A shows the case when the size of inner lipids is fixed (0,9·A0), while the outer lipid molecule size is being varied {$1,1\cdot A_0$, A_0, $0,95\cdot A_0$, $0,88\cdot A_0$, $0,85\cdot A_0$}. Total number of lipids composing a vesicle is as follows [2811, 3052, 3240, 3606]. Fig 3B shows visual representation of vesicle's cross-section during the preparation for performing MD simulations.

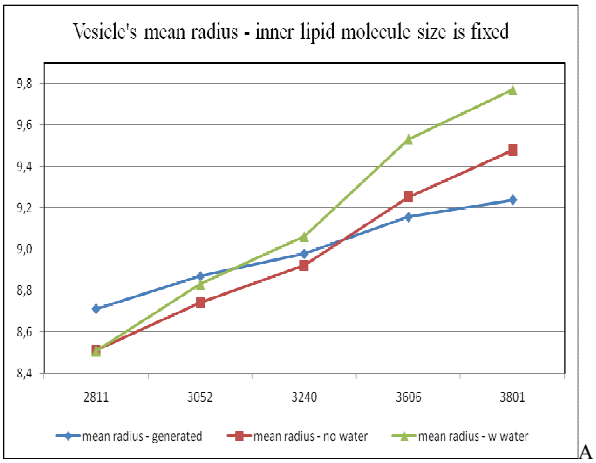


A



B

Fig 2: A) Graph shows vesicle's mean radius with respect to number of DSPC lipid molecules composing the vesicle when outer lipid molecules are fixed. B) Pictures show from left to right cross-sections of vesicles for cases of 3012, 3311 and 3624 lipids present in the membrane, respectively. The pictures from top to bottom show: starting structure, EM of lipids in vacuum, EM of lipids and water in vacuum.



A

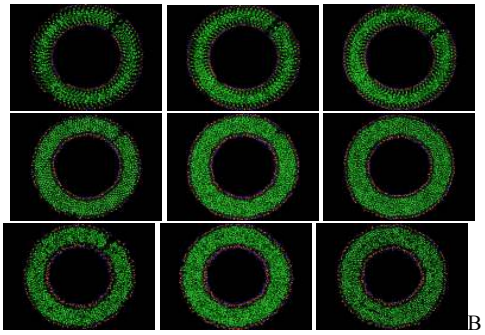

B

Fig 3 A) Graph shows vesicle's mean radius with respect to number of DSPC lipid molecules composing the vesicel when inner lipid molecules are fixed. B) Pictures show from left to right the cross-section of vesicles for cases of 2811, 3240 and 3801 lipids present in the membrane, respectively. The pictures from top to bottom show: starting structure, EM of lipids in vacuumm, EM of lipids and water in vacuum.

We hypotesise for the results shown in Fig. 2A, related to the mean radius that decreases with respect to increasing number of lipids present in innner monolayer. This could be due to wrinkling of inner monolayer, since the outer one does not allow expansion. For the case of increasing number of lipids in outer monolayer the mean radius is increasing, which could be due to billow up of the outer monolayer, and the inner one just simply follows it, due to hydrophobic interactions among the lipid tails.

B. Summary and Future Work

We confirmed, with performing CG-MD simulations on a multicore parallel computer, the stability of vesicles from ensembles of a few thousand of lipid molecules and tens of thousands of water molecules, simulated in a rectangular box for a time period of 30 ns. A typical computational time needed for such simulations on four hyper-threaded computing cluster cores is in the range of hours (8 to 10) to days (2 to 5), depending on the number of lipids and the length of simulation. However, we can run multiple experiments in parallel, which could significantly broaden and speedup simulation studies.

The future work will be concentrated in detailed analysis of simulated results, with a focus to evaluate and validate the results with already measured experimental values of mechanical properties, such as bending elasticity obtained via thermal fluctuations of lipid membranes [19], vesicle fusion and fission, etc.

IV. CONCLUSIONS

We have simulated spherically shaped structures such as lipid vesicles, composed of DSPC lipids surrounded with water molecules in a rectangular simulation box. It was demonstrated that the density of lipids has a significant impact on vesicle shapes, in particular by vesicles with dimensions in the same range as bilayer thickness. All our coarse grain molecular dynamics simulations were run on a multicore parallel computer with GROMACS software package using MARTINI force field model. The above mentioned structure is our building block for future studies of mechanical and physical properties of lipid structures, interactions between proteins and lipid membranes, and effects of chemical parameters on their interactions.

REFERENCES

[1] E. Marco, and F. Gago, “Overcoming the Inadequacies or Limitations of Experimental Structures as Drug Targets by Using Computational Modeling Tools and Molecular Dynamics Simulations,” *Chem.Med.Chem.*, 2, 1388–1401, 2007.

[2] H. Gonzalez-Diaz, S. Vilar, L. Santana, and E. Uriarte, “Medicinal Chemistry and Bioinformatics–Current Trends in Drugs Discovery with Networks Topological Indices,” *Curr. Top. Med. Chem.*, 7, 1015–1029, 2007.

[3] K. V. Damodaran, Kenneth M. Merz, Jr., and Bruce Paul Gaber, “Interaction of Small Peptides with Lipid Bilayers,” *Biophys. J.*, 69, 1299–1308, 1995.

[4] F. J.–M. de Meyer, J. M. Rodgers, T. F. Willems, and B. Smit, “Molecular Simulation of the Effect of Cholesterol on Lipid–Mediated Protein–Protein Interactions,” *Biophys. J.*, 99, 3629–3638, 2010.

[5] T. Murtola, T. A. Vuorela, M. T. Hyvonen, S.–J. Marrink, M. Karttunen, and I. Vattulainen, “Low density lipoprotein: structure, dynamics, and interactions of apoB–100 with lipids,” *Soft Matter*, 7, 8135–8141, 2011.

[6] U. Seifert, K. Berndl, and R. Lipowsky, "Shape transformations of vesicles: Phase diagram for spontaneouscurvature and bilayer-coupling models," *Phys. Rev. A*, 44, 1182–1202, 1991.

[7] M. I. Angelova, and D. S. Dimiter, “Liposome electroformation,” *Faraday Discuss. Chem. Soc.*, 81, 303–311, 1986.

[8] R. Trobec, I. Jerebic, and D. Janežič, “Parallel algorithm for molecular dynamics integration ,” *Parallel Comput.*, 19, 1029–1039, 1993.

[9] R. Trobec, F. Merzel, and D. Janežič, “The complexity of parallel symplectic molecular dynamics algorithms,” *J. Chem. Inf. Comput. Sci.*, 37, 1055–1062, 1997.

[10] M. Praprotnik, L. Delle Site, and K. Kremer, “Adaptive resolution scheme for efficient hybrid atomistic-mesoscale molecular dynamics simulations of dense liquids,” *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 73, 66701-12, 2006.

[11] A. Velikonja, Š. Perutkova, E. Gongadze, P. Kramar, A. Polak, A. Maček-Lebar, and A. Iglič, " Monovalent ions and water dipoles in contact with dipolar zwitterionic lipid headgroups-theory and MD simulations," *Int. J. Mol. Sci.*, 14, 2846–2861, 2013.

[12] J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein, “A Coarse Grain Model for Phospholipid Simulations,” *J. Phys. Chem. B*, 105, 4464–4470, 2001.

[13] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, “The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations,” *J. Phys. Chem. B*, 111, 7812–7824, 2007.

[14] B. Hess, “GROMACS 4: Algorithms for Highly Efficient Load–Balanced, and Scalable Molecular Simulation,” *J. Chem. Theory Comput.*, 4, 435–447, 2008.

[15] J. I. Pavlič, and R. Trobec, "Coarse grain simulation and visualisation of lipid and water molecules ensembles," MIPRO, 2012 Proceedings of the 35th Internatonal Convention, 462–466, 2012.

[16] S. J. Marrink, and A. E. Mark, "Molecular Dynamics Simulation of the Formation, Structure, and Dynamics of Small Phospholipid Vesicles," *J. Am. Chem. Soc.*, 125, 15233–15242, 2003.

[17] A. H. de Vries, A. E. Mark, and S. J. Marrink, "Molecular Dynamics Simulation of the Spontaneous Formation of a Small DPPC Vesicle in Water in Atomistic Detail," *J. Am. Chem. Soc.*, 126, 4488–4489, 2004.

[18] H. J. Risselada, and S. J. Marrink, "Curvature effects on lipid packing and dynamics in liposomes revealed by coarse grained molecular dynamics simulations," *Phys. Chem. Chem. Phys.*, 11, 2056–2067, 2009.

[19] J. Genova J, A. Zheliaskova, and M. D. Mitov, "The influence of sucrose on the elasticity of SOPC lipid membrane studied by the analysis of thermally induced shape fluctuations," *Coll. and Surf. A*, 282-283 420–422, 2006

Low-frequency ultrasound *in vitro*: experimental system and ultrasound-induced changes of cell morphology

J. Jelenc*, J. Jelenc *, D. Miklavčič** and A. Maček Lebar**
* Iskra Medical d.o.o., Slovenia
** University of Ljubljana, Faculty of Electrical Engineering, Slovenia
*Jure@iskramedical.eu

Abstract - Ultrasound can temporary or permanently increase the membrane's permeability for molecules that would otherwise be unable to enter the cell. Sonoporation that causes temporary changes on cell membrane can be used to introduce foreign material to the cell interior. Permanent increase of cell membrane permeability however causes cell death that can be used for water purification, waste water treatment, ballast water treatment in shipping, food processing and other purposes.

To study these effects in an in vitro setting, we have built a custom low-frequency ultrasound experimental system based on an ultrasound transducer submerged in a waterbath. Using a hydrophone we have evaluated the ultrasound intensity in the waterbath with or without an ultrasound absorbing material lining the bath walls. In this system we evaluated effects of low frequency ultrasound on B16-F1 mouse melanoma cells.

I. INTRODUCTION

Sonoporation is a phenomenon where ultrasound increases cell membrane permeability. As a result, molecules that are otherwise unable to pass the cell membrane can be transported across it. This becomes possible for molecules with both small and large molecular masses [1]–[3]. If the cell remains capable of repairing the damage to the membrane and re-establishing its normal state, the phenomenon is called reversible sonoporation. If the cell dies as a consequence to ultrasound exposure, the sonoporation is irreversible. Irreversible sonoporation can be used for water purification, waste water treatment, ballast water treatment in shipping in food production process and other purposes.

To study effects of ultrasound in an in vitro setting, we have built a custom experimental low-frequency ultrasound system based on an ultrasound transducer submerged in a waterbath. Ultrasound pressure is one of the most important parameters in such a system. We used a hydrophone to evaluate the ultrasound pressure in the waterbath with or without an ultrasound-absorbing material lining the bath walls. The developed and evaluated system allows sonoporation experiments to be performed under known and verified ultrasound exposure conditions. Some ultrasound-induced changes of cell morphology are presented at the end of this article.

II. MEHTODS

In our study we use a waterbath with a submerged transducer that was previously described [4]. In brief, the waterbath with a length of 68 cm, width of 38 cm and height of 34 cm is filled with distilled water up to a height of 24 cm (Figure 1). Ultrasound is generated using a prototype center bolt (Langevin type) piezoelectric ultrasound transducer [5], [6] with an operating frequency of 29.6 kHz (Iskra Medical, Slovenia). The transducer is submerged in the waterbath at a depth of 12 cm. Ultrasound intensity is calculated from ultrasound pressure measured with a piezoelectric hydrophone (8103 hydrophone, Brüel & Kjær, Denmark) located on the central axis of the ultrasound transducer. The hydrophone positioning system allows the distance between the hydrophone and transducer to be adjusted. In the experiments with cells, the hydrophone was replaced by a cell container. The walls of the bath are made from Plexiglas®. In part of hydrophone measurements and sonoporation experiments the walls of the waterbath are lined with the SA-J35 ultrasound absorber (Hangzhou Applied Acoustics Institute, China).

A. Measuring the ultrasound intesity

In order to study ultrasound reflection, we generated 8 periods of an ultrasound wave (29.6 kHz, power set to 50%) and calculated instantaneous acoustic intensity I as defined in IEC 61101:

$$I = p^2 / (\rho \cdot c) \quad , \qquad (1)$$

where ρ is the density of the propagating medium, and c is the velocity of sound in the propagating medium. For water at room temperature ρ has a value of 1000 kg/m^3 and c has a value of 1480 m/s. Ultrasound pressure p was measured with a hydrophone at a distance of 3.5 cm from the piezoelectric ultrasound transducer. After measuring the ultrasound intensity over the first 0.6 ms, we measured the reflected intensity from 0.6 to 10 ms. The measurement procedure was first performed in a bath with no absorbable lining. The walls were then lined with an absorbable lining and the procedure was repeated. In both cases, the maximum intensity of the signal was estimated.

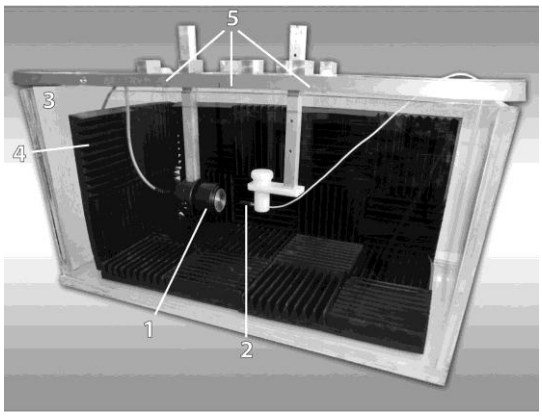


Figure 1. Experimental system consisting of: an ultrasound transducer (1), a hydrophone (2), waterbath walls (3), ultrasound-absorbing lining (4) and a positioning system (5).

B. Exposure of cells to ultrasound

The effect of ultrasound was evaluated on mouse melanoma B16-F1 cells. B16-F1 cells were cultivated in DMEM (Sigma-Aldrich Chemie GmbH, Germany) cell growth medium as previously described by Usaj [7]. After cell detachment, a cell suspension with concentration of 10^6 cells/ml was introduced into a 0.2 ml PCR tube (Invitrogen, USA). The cell suspension was vigorously mixed before ultrasound exposure in order to introduce gas bubbles (acting as cavitation nuclei) into the cells suspension.

The cell dish was positioned at the axial center of the transducer, 2.5 cm from the ultrasound transducer face. Cells were exposed to 300 seconds of continuous-wave 21.1 W/cm^2 ultrasound intensity [8]. The waterbath walls were lined with the ultrasound-absorbing material. A sham exposure with no applied ultrasound was conducted with the same procedures.

Just before the experiment, 5 µl of Propidium Iodide (PI) (Molecular Probes, USA) were mixed into the cell suspension. In normal conditions, a cell membrane is impermeable to PI. Damage caused to the cell membrane by ultrasound enables PI to enter the cytoplasm, where it binds to the nucleus. Characteristic PI fluorescence can be used to identify cells with increased cell membrane permeability.

Immediately after the exposure, cells were transferred onto a Petri dish, which was placed under an inverted fluorescence microscope (Zeiss AxioVert 200, Zeiss, Germany) with 200x magnification. Pictures were acquired by a cooled CCD camera (VisiCam 1280, Visitron, Germany) using Metamorph 5.0 (Molecular Devices Corporation, PA, ZDA) software. Morphological changes of the exposed cells and PI intake-induced fluorescence were analyzed in the acquired pictures.

III. RESULTS

A. Measuring the ultrasound pressure

The measured pulsed signal can be divided into two parts. The first 0.6 ms represent eight pulses of incident ultrasound with maximum intensity 12.3 W/cm^2. Reflections of pulsed ultrasound from boundaries are shown in Figure 2. Reflections with no absorbable acoustic lining are shown in Figure 2a while reflections using the absorbable acoustic lining on the boundaries of the waterbath are shown in Figure 2b. In both cases, the measurement causes measurement saturations. We have therefore omitted the first 0.6 ms from Figure 2. Figure 2 thus only shows reflections. If no absorbable lining is used, the maximum intesity of the reflected wave is 1.19 W/cm^2 or 10% of the incident signal. The maximum value of reflection pressure using absorber is 0.41 W/cm^2 or only 0.3% of the incident signal.

B. Exposure of cells to ultrasound

To analyze effects of ultrasound on cells, we observed morphological changes induced by ultrasound using phase contrast microscopy (Figure 3A) and Propidium Iodide (PI) intake observed using fluorescence microscopy (Figure 3B).

Phase contrast microscopy showed heterogeneous effects of ultrasound within a single sample. The result on cells varied from cells with no visible morphological change to cells torn into several parts. A number of cells disintegrated into small cell debris, which were clearly distinguished inside the sample. We additionally observed blebs on some cell membranes after ultrasound exposure.

Merging membrane permeability and cell morphology information enabled us to classify the heterogeneous effect on cells shown in Figure 3. Morphologically intact cells were classified into cells on which ultrasound had no effect (Figure 3-1a) and reversibly sonoporated cells which are morphologically intact cells with observed PI intake (Figure 3-1b). Ultrasound-induced damage on some of the cells was so intense that we believe they are unlikely to survive the ultrasound exposure. These cells were grouped into necrotic (Figure 3-2) and cell debris-exibiting (Figure 3-3a) and without exibiting PI fluorescence (Figure 3-3b). Furthermore we noted some cells with blebs on the cell membrane (Figure 3-1c).

The majority of sham-exposed cells were morphologically intact, with only a small fraction, less than 5 % of cells, exhibiting PI intake-related fluorescence (data not shown). The sham exposure group had only a limited number of cell debris and no blebs were noted on cell membranes.

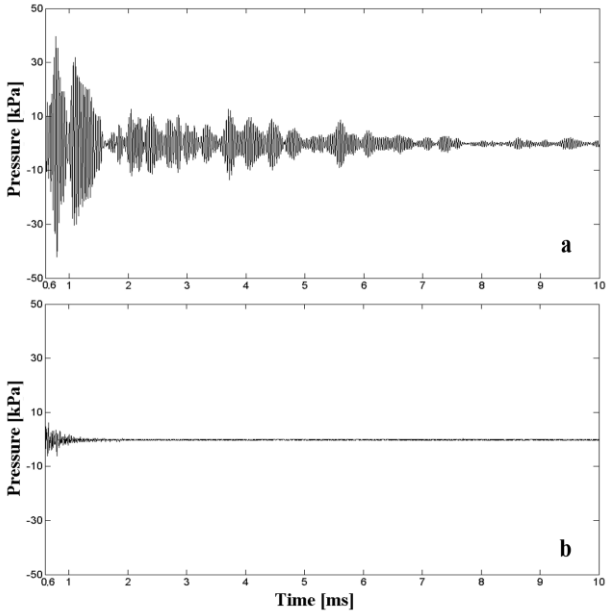


Figure 2. Reflection of eight periods of the ultrasound signal from the boundaries; the reflections are shown 0.6 ms after the signal was generated, so the incident ultrasound has already passed the hydrophone; signal a) without absorptive lining of the waterbath walls and b) with the absorptive lining.

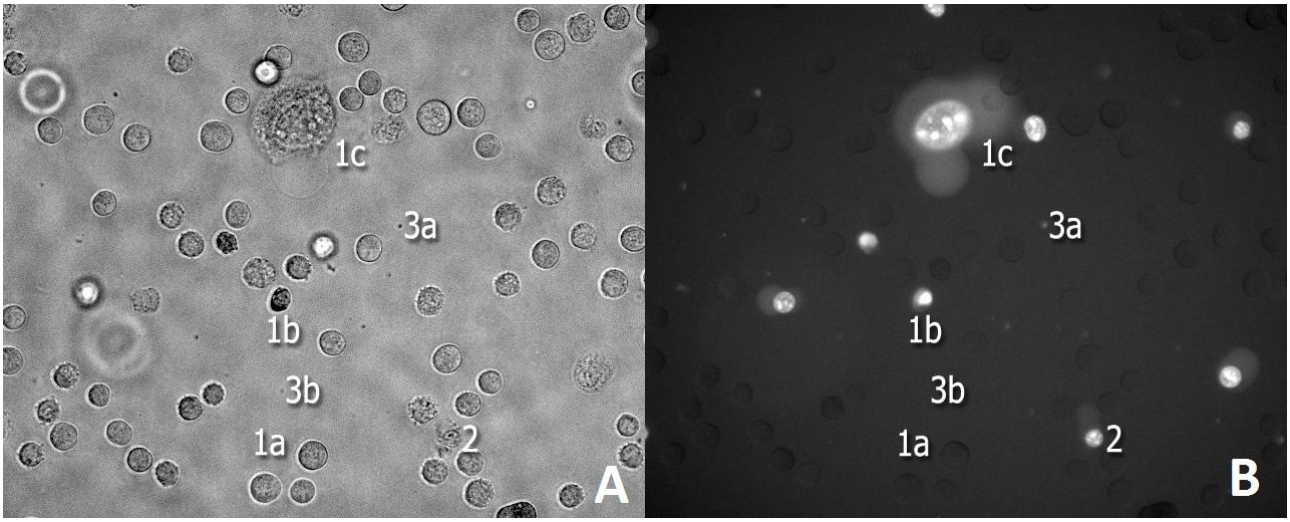


Figure 3. Image of B16-F1 cells at 200x magnitude after ultrasound exposure: A) phase contrast microscopy and B) fluorescent microscopy. Typical representatives of ultrasound effect are marked: 1a. intact cell, 1b. reversibly sonoporated cell, 1c. cell with blebs on the cell membrane, 2. necrotic cell, 3a. cell debris exibiting Propidium Iodide (PI) fluorescence, 3b. cell debris without exibiting PI fluorescence.

IV. DISCUSION

The results of our study show that reflected waves are present in a low-frequency experimental system of finite dimensions. Reflected waves are especially strong if no absorbable material is used. Though reflected waves are weaker than the waves coming directly from the ultrasound transducer, the intensity of the reflected waves in system with absorbable lining represents up to 10% of the ultrasound intensity of the waves coming directly from the transducer. Reflected ultrasound waves are present in the waterbath up to 2 ms after the ultrasound transducer has stopped emitting waves. In this time, ultrasound can travel across the longest dimension of the waterbath four times. It is thus very likely that reflections of reflected waves are also present in the waterbath. All these reflections are superimposed onto the incident wave, preventing accurate measurements of the incident ultrasound. If the system is working continuously (cw), stationary waves may also occur and further reduce the accuracy of the incident ultrasound measurements. In our system, reflections were successfully reduced by lining the walls of the waterbath with a special material absorbing ultrasound. We were therefore able to do experiments with continuous-wave ultrasound exposure.

In our low-frequency ultrasound experiment, we observed heterogeneous effects on cells within the same sample ranging from no effects to total cell destruction as small cell debris was formed. This is ubiquitous for ultrasound since cavitation shockwaves are believed to be responsible for plasma membrane opening. Shockwaves from cavitation activity are emitted from point sources and dissipate away from the bubble during propagation. Thus cells close to a bubble experience a stronger shockwave compared with cells located farther away [9]. We did not attempt to quantify the number of cells in groups by their effects since their number varied significantly between samples. This was mainly because we were unable to efficiently control the number of cavitation bubbles introduced by mixing of the cell suspension. To increase reproducibility of the results, an alternative way of introducing cavitation bubbles will have to be used.

An increase in membrane permeability of eukaryotic cells due to low-frequency ultrasound has been previously shown using calcein and/or sulhordamine dye [10]–[13]. Immediate cell viability after ultrasound exposure was determined by the absence of Propidium Iodide (PI) fluorescence. By staining cell plasma membrane using wheat germ agglutinin - WGA and intracellular membrane using fluorescence marker FM1-43 Schlicher et al were able to define intracellular origin of observed membrane blebs that were formed at sites of plasma membrane disruption [11] after the ultrasound exposure. These effects and their quantification were studied in low-frequency standing wave like ultrasound filed [14]. The ultrasound exposure conditions therefore differ significantly since use of efficient acoustic lining provides a progressive ultrasound wave and inhibits standing wave formation. This system design difference is important since a number of investigators have shown that cavitation, as the main mechanism of sonoporation, can be more easily induced by a standing wave than by a progressive wave [15], [16]. However, a progressive ultrasound wave is more common than a standing wave when translating to clinical conditions.

Wei et al utilized progressive waves with lower exposure pressure than ours, but longer exposure durations to transfect 35% of cells with Green Fluorescent Protein at 90% cell viability [17]. That study did not report the described morphological changes, but the study performed by Schlicher et al [11] did. During our experiments, we observed all typical morphological changes in cells described by Schlicher et al, but for now have not been able to obtain the experimental reproducibility that is mostly described in literature [10]–[13], [17]. In future work, considerable effort will be made to better control the cavitation bubbles and thus increase reproducibility of the results.

V. Conclusion

We successfully constructed an experimental ultrasound exposure system and lined it with absorbable lining to reduce reverberation from waterbath walls. Using this system, we have observed an effect of low-frequency ultrasound on eukaryotic cells comparable to described in the literature.

Acknowledgment

Operation is part financed by the European Regional Development Fund (Biomedical Engineering Competence Center, Slovenia) and European Social Fund.

References

[1] H.-D. Liang, J. Tang, and M. Halliwell, “Sonoporation, drug delivery, and gene therapy.,” *Proc Inst Mech Eng H*, vol. 224, no. 2, pp. 343–361, 2010.

[2] R. Karshafian, P. D. Bevan, R. Williams, S. Samac, and P. N. Burns, “Sonoporation by ultrasound-activated microbubble contrast agents: effect of acoustic exposure parameters on cell membrane permeability and cell viability.,” *Ultrasound Med Biol*, vol. 35, no. 5, pp. 847–860, May 2009.

[3] R. Karshafian, S. Samac, P. D. Bevan, and P. N. Burns, “Microbubble mediated sonoporation of cells in suspension: clonogenic viability and influence of molecular size on uptake.,” *Ultrasonics*, vol. 50, no. 7, pp. 691–697, Jun. 2010.

[4] J. Jelenc, J. Jelenc, D. Miklavčič, and A. Maček Lebar, “Low-Frequency Sonoporation in vitro: Experimental System Evaluation,” *Strojniški vestnik – Journal of Mechanical Engineering*, vol. 58, no. 5, pp. 319–326, May 2012.

[5] M. Đ. Radmanović and D. D. Mančić, *Design and modeling of the power ultrasonic transducers*, vol. 2004. Nis: University of Nis, Faculty of Electronic Engineering.

[6] E. Moreno, P. Acevedo, M. Fuentes, A. Sotomayor, L. Borroto, M. E. Villafuerte, and L. Leija, “Design and construction of a bolt-clamped Langevin transducer,” in *Electrical and Electronics Engineering, 2005 2nd International Conference on*, 2005, pp. 393–395.

[7] M. Usaj, K. Trontelj, D. Miklavcic, and M. Kanduser, “Cell-cell electrofusion: optimization of electric field amplitude and hypotonic treatment for mouse melanoma (B16-F1) and Chinese Hamster ovary (CHO) cells.,” *J Membr Biol*, vol. 236, no. 1, pp. 107–116, Jul. 2010.

[8] J. Jelenc, J. Jelenc, D. Miklavčič, and A. Maček Lebar, “Low frequency ultrasound in vitro: effects on B16-F1 cells (Vplivi nizkofrekvenčnega ultrazvoka na celice B16-F1 v in vitro pogojih),” in *Zbornik dvajsete mednarodne Elektrotehniške in računalniške konference ERK 2011, 19.-21. september 2011,*, Portorož, 2011, vol. 2, pp. 363–366.

[9] Y. Liu, J. Yan, and M. R. Prausnitz, “Can Ultrasound Enable Efficient Intracellular Uptake of Molecules? A Retrospective Literature Review and

Analysis," *Ultrasound in medicine & biology*, vol. 38, no. 5, pp. 876–888, May 2012.

[10] J. Sundaram, B. R. Mellein, and S. Mitragotri, "An experimental and theoretical analysis of ultrasound-induced permeabilization of cell membranes.," *Biophys J*, vol. 84, no. 5, pp. 3087–3101, May 2003.

[11] R. K. Schlicher, J. D. Hutcheson, H. Radhakrishna, R. P. Apkarian, and M. R. Prausnitz, "Changes in Cell Morphology Due to Plasma Membrane Wounding by Acoustic Cavitation," *Ultrasound in Medicine & Biology*, vol. 36, no. 4, pp. 677–692, Apr. 2010.

[12] K. Keyhani, H. R. Guzmán, A. Parsons, T. N. Lewis, and M. R. Prausnitz, "Intracellular drug delivery using low-frequency ultrasound: quantification of molecular uptake and cell viability.," *Pharm Res*, vol. 18, no. 11, pp. 1514–1520, Nov. 2001.

[13] J. D. Hutcheson, R. K. Schlicher, H. K. Hicks, and M. R. Prausnitz, "Saving Cells from Ultrasound-Induced Apoptosis: Quantification of Cell Death and Uptake Following Sonication and Effects of Targeted Calcium Chelation," *Ultrasound in Medicine & Biology*, vol. 36, no. 6, pp. 1008–1021, Jun. 2010.

[14] J. Liu, T. N. Lewis, and M. R. Prausnitz, "Non-invasive assessment and control of ultrasound-mediated membrane permeabilization.," *Pharm Res*, vol. 15, no. 6, pp. 918–924, Jun. 1998.

[15] M. Kinoshita and K. Hynynen, "Key factors that affect sonoporation efficiency in in vitro settings: the importance of standing wave in sonoporation.," *Biochem Biophys Res Commun*, vol. 359, no. 4, pp. 860–865, 2007.

[16] A. H. Barati, M. Mokhtari-Dizaji, H. Mozdarani, Z. Bathaie, and Z. M. Hassan, "Effect of exposure parameters on cavitation induced by low-level dual-frequency ultrasound," *Ultrasonics Sonochemistry*, vol. 14, no. 6, pp. 783–789, Sep. 2007.

[17] W. Wei, B. Zheng-zhong, W. Yong-jie, Z. Qing-wu, and M. Ya-lin, "Bioeffects of Low-Frequency Ultrasonic Gene Delivery and Safety on Cell Membrane Permeability Control," *J Ultrasound Med*, vol. 23, no. 12, pp. 1569–1582, Dec. 2004.

Using Hadoop MapReduce in a Multicluster Environment

I. Tomašić, A. Rashkovska and M. Depolli
Jožef Stefan Institute/Department of Communication Systems, Ljubljana, Slovenia
ivan.tomasic@ijs.si, aleksandra.rashkovska@ijs.si, matjaz.depolli@ijs.si

Abstract - Hadoop MapReduce has become one of the most popular tools for data processing. Hadoop is normally installed on a cluster of computers. When the cluster becomes undersized, it can be scaled by adding new computers and storage devices, but it can also be extended by real or virtual resources from another computer cluster. We present a utilization of the MapReduce paradigm on a Hadoop installation extended across two clusters connected over the Internet. We measured execution times of Map and Reduce tasks in a multicluster environment, and compared them to the corresponding times obtained while only computers from a single cluster are used. The results show that there might be a decrease in MapReduce performance depending on: the concrete data analyses application, the ratio of the number of local and remote computers, and connection bandwidth to remote computers. Additionally, the investigation suggests an upgrade to the Apache Hadoop MapReduce, making it more adjusted to the multicluster environment.

I. INTRODUCTION

Apache Hadoop [1] is a highly popular set of open source modules for distributed computing. The key Hadoop components are the Hadoop Distributed File System (HDFS) and the MapReduce - a data-processing component. Since 2004, when it was introduced [2], the MapReduce paradigm has become one of the most popular tools for batch-processing large datasets, mostly because it allows users to build complex distributed programs using a simple model.

Hadoop has proved its ability to store and analyze huge datasets often referred to as the BigData [3]. It has been used by Yahoo and Facebook [4]. If a Hadoop cluster becomes undersized, a commonly used approach is to scale the cluster by adding new computers and storage devices. Other possibility is to resort for resources on another computer cluster. The additional resources can be virtual or real computers.

In our previous work [5], we presented the steps and configurations needed for extending a HDFS installation with computers from another cluster accessed over the Internet. The presented benchmarks results, for one additional computer, showed a drop in read and write operations by approximately one order of magnitude. In the present study, we investigate the applicability and efficiency of the Hadoop MapReduce on the same hardware and network topology (see [5] for details). For the present study, the Cloudera Apache Hadoop distribution CDH 4.1.0 has been used [6].

For evaluating the MapReduce performance in a multicluster environment, we have measured Map and Reduce tasks execution times, and compared them to the corresponding times obtained on a single cluster. The multicluster environment is achieved by connecting an additional server using a 100Mb/s link (note that the network bandwidth inside the cluster is 1 Gb/s). We developed MapReduce jobs for analyzing tabular data coming from heat transfer simulation in a biomedical application, in particular, cooling of a human knee after surgery [7]. Similar data sources can be found also in other scientific areas related to multi-parametric simulations [8], environmental data analysis [9], high energy physics [10], etc., whereas some special problems may benefit from specific interfaces to the Hadoop [11].

A. MapReduce Paradigm

The MapReduce user specifies two functions called Map and Reduce, which operate on data arranged in key/value pairs. The first step is data splitting which is done by the MapReduce frameworks. The splits are than processed by Map functions which are commonly applied on each line of every split. Each Map function invocation outputs a list of key/value pairs. In the case of a word count, the Map function would output words as the keys and numbers of instances of each word in the line as the values. The MapReduce framework groups together all intermediate values associated with the same intermediate key and passes them to the Reduce function, which aggregates or merges them together to form a new, possibly smaller, set of values. In the word-count example, the Reduce function sums and outputs total number of instances for each word. The executions of the Map and Reduce functions are referred to as Map and Reduce tasks, or shortly Mappers and Reducers. A set of tasks executed for one application are referred to as a MapReduce job.

B. Apache Hadoop MapReduce Implementation

The Hadoop splits are fixed-size, whereas a separate Map task is created for each split (Fig. 1). The default split size is the same as the default size of an HDFS block, which is 64 MB. The Hadoop performs data locality optimization by running the Map task on the node where the input data resides in the HDFS.

The Map tasks write their outputs to their local disks, and partition their outputs, creating one partition for each

The research was funded in part by the European Union, European Social Fund, Operational Programme for Human Resources, Development for the Period 2007-2013.

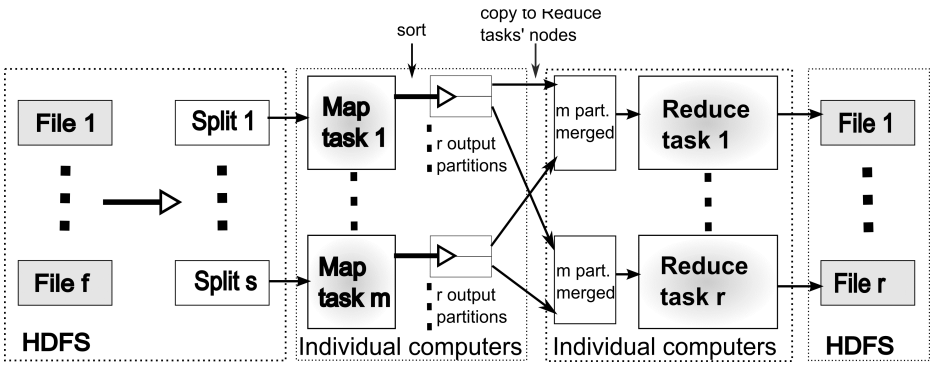


Figure 1. Schematic representation of MapReduce data flow (Combiner functions are not presented).

Reduce task. Each partition may contain various keys and associated values. All records sharing the same key are processed by the same Reduce task. This is achieved by using the so-called Partitioner function. The default Hadoop MapReduce Partitioner employs a hash function on the keys from the Maps' outputs. Modulo function by the number of reducers is subsequently applied to the hash values resulting with the Reduce task indexes for each key. Fig. 1 shows schematically the MapReduce data flow.

Reduce tasks, unlike Map tasks, cannot convey on data locality because the input to a single Reduce task is generally formed from outputs of multiple Map tasks (Fig.1). On the Reduce task's node, the sorted map outputs are merged before being passed to the Reduce task. The data flow between Map and Reduce tasks is known as "shuffle", whereas the merging is logged as a part of the Reduce task. The number of Reduce tasks is specified independently for a given job. Each Reduce task outputs a single file, which is usually stored in the HDFS.

Hadoop allows a user to specify an additional function, called Combiner, which receives all the data emitted by the Map tasks on the same node as the input, executes on that node, and forms an output that becomes an input to a Reduce function. The Combiner function is used to achieve data reduction that consequently minimizes data transfer over the network and reduces the impact of the limited communication bandwidth on the performances of a MapReduce job.

C. MapReduce limitations

The main limitation of the MapReduce paradigm is that each Map and Reduce task must not depend on any data generated in other Map or Reduce tasks of the current job, as user cannot control the order in which they execute. Although applicable to a wide variety of problems, there are problems to which the MapReduce is not directly applicable. These are recursive computations, and algorithms that depend on shared global state. A classic example of recursive computation is the Fibonacci series where each value is summation of the previous two values: $F(i) = F(i-1) + F(i-2)$. Examples of the algorithms that require global synchronization are Online learning and Monte Carlo simulations [12].

II. METHODS

A. System Arhitecture

The utilized system hardware and networking architecture has been described in our previous study [5], where we used the original Apache Hadoop distribution [1]. For the current study, we have used the Cloudera distribution [6] that was installed on a cluster of six computing nodes. The nodes are connected with Gigabit Ethernet, whereas the additional external node is connected through a 100Mb/s channel.

One of the nodes is designated as the namenode, while others are the datanodes. The namenode also hosts the jobtracker. All machines in the cluster run an instance of a datanode and a tasktracker. For a description of the HDFS and MapReduce nodes refer to [13], [14]. After joining the additional node, the obtained Hadop multicluster has been rebalanced by executing Cloudera's HDFS balancer [15], to allow the data blocks to be equally distributed to the additional node.

B. Test Dataset

A computer simulation of two hours cooling of a human knee after surgery was performed for 10 different knee sizes, 10 different initial temperature states before cooling, and 10 different temperatures of the cooling pad. This resulted in 1000 simulation cases. The results of those simulation cases were gathered in 100 files. Each file contained 71970 rows or approximately 44 MB of data. Each data row was composed of the following parameters, i.e. columns: RT, D, IS, CT, T1, T2, … , T85, where are: RT - relative time in a simulation case, D - knee size, IS –initial state, CT – cooling temperature,

TABLE I. LIST OF TEST CASES

Case	Parameters
1	T1
2	T1-T5
3	T1,T6,T11,T16,T21
4	T1-T21
5	T1,T6,T11,T16,T21,T46,T51,T56,T61
6	T1-T21,T46-T61
7	T1,T6,T11,T16,T21,T26,T31,T36,T41,T46,T51, T56,T61,T66,T71,T76,T81
8	T1-T85

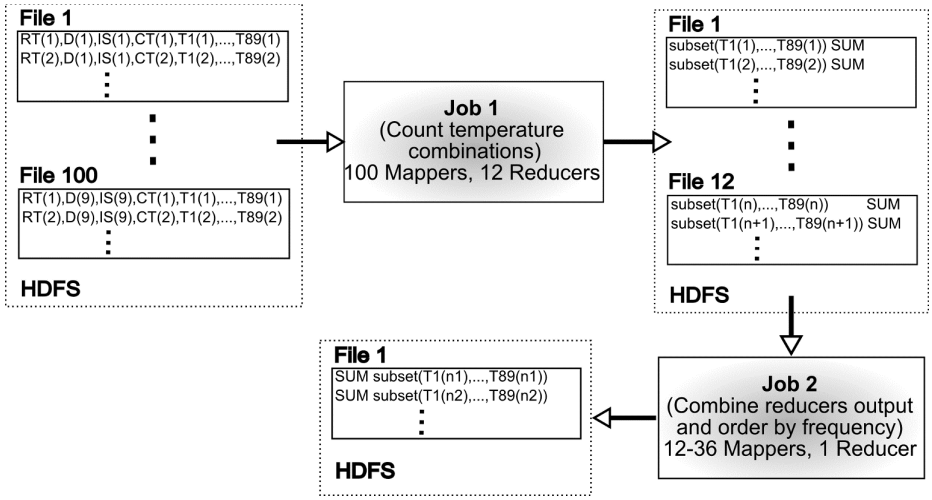


Figure 2. MapReduce jobs pipeline.

T1-T85 – inner and outer knee temperatures, i.e., temperatures at a particular location in the knee center, 8 locations on the knee skin and 8 respective locations under the cooling pad, all taken in the current and in previous time steps. In order to assess the periodicities in the knee simulation results, we assigned the MapReduce to count the occurrences of the same value arrays for a subset of knee temperatures T; more precisely, to count the occurrences of identical rows after having projected only columns of T that are of interest. We will refer to the number of occurrences of identical rows as temperatures' frequencies.

We defined and examined 8 cases with different sets of T. The cases are given in Table I.

C. MapReduce jobs

The MapReduce jobs pipeline for solving our test cases is illustrated in Fig. 2. Job 1 counts the occurrences of the same value arrays for a subset of temperatures, whereas Job2 combines outputs from Job 1 and orders the combined rows by temperature frequency.

The sizes of the input files are smaller than the HDFS block size (in our case: 64 MB). Hence, the number of input Map tasks in Job 1 is equal to the number of input files [16], i.e., each input file is processed by a different Map task. The number of Reduce tasks is set to 12 for Job 1, hence the output of Job 1 now consists of 12 files. Each file contains a unique combination of temperatures and the number of their occurrences. Job 2 combines Reduce tasks' outputs from Job 1 into a single file (in Job 2, the number of Reduce tasks is set to 1). It also sorts the input columns in the output file by temperatures' frequencies. The number of Map tasks in Job 2 depends on the test case (Table I) and varies between 12 for Case 1 and 36 for Case 8. Job 1 has the same Combiner and Reduce functions, whereas the Combiner function for the Job 2 is not specified, since there are no multiple rows for the same key in the Map tasks output data. The implementation of both Jobs is given in Fig. 3.

In the Map function of Job 1, from each input row, only the relevant columns (see Table I) are extracted. For example, in test case 2, only the columns belonging to T1-T5 will be extracted in the SearchString variable. Reduce functions sum, i.e., count the number of occurrences of each combination of temperatures (the key) and outputs it as the new value for the current key. Because all the values for one key are processed by a single Reduce task, it is evident that the output of Job 1 consists of unique combinations of temperatures and the number of their occurrences.

```
//Job 1
public void map(LongWritable key,Text value,
OutputCollector<Text,IntWritable> output, Reporter reporter)
throws IOException{
        String line = value.toString();
        String[] lineElements  = line.split(",");
        String SearchString = null
        //depending on a case (Table I) concatenate different
        lineElements in //SearchString
        …
        word.set(SearchString);
        output.collect(word, new IntWritable(1));
}
public void reduce(Text key, Iterator<IntWritable> values,
OutputCollector<Text, IntWritable> output, Reporter reporter)
throws IOException{
        int sum = 0;
        while (values.hasNext()){
                sum += values.next().get();
        }
        output.collect(key, new IntWritable(sum));
}
```

```
//Job 2
public void map(LongWritable key, Text value,
OutputCollector<IntWritable,Text> output, Reporter reporter)
throws IOException{
        String line = value.toString();
        //\t is the default delimiter used by a reducer
        String[] lineElements  = line.split("\t");
        output.collect(new
        IntWritable(Integer.parseInt(lineElements[1])),
                                        new Text(lineElements[0]));
}
public void reduce(IntWritable key, Iterator<Text> values,
OutputCollector<IntWritable, Text> output, Reporter reporter)
throws IOException{
        //there is only one value
        output.collect(key, values.next());
}
```

Figure 3. Java code segments of Map and Reduce tasks for Job 1 and Job 2.

TABLE II. MAPREDUCE TASKS EXECUTION TIME DIFFERENCES BETWEEN MULTICLUSTER AND SINGLE CLUSTER CASE.

	Job 1								Job 2							
Case:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Total time spent by all maps (s)	96	77	53	43	25	69	71	52	1	0	-4	-1	1	11	6	38
Total time spent by all reduces (s)	5	-12	56	97	131	308	29	2082	0	1	5	83	15	2	0	284
Map tasks avg. time (s)	1	1	0	1	0	1	0	0	0	0	-1	0	0	1	1	1
Worse performing map task (s)	1	1	1	-1	0	-1	3	0	0	0	0	0	0	0	0	4
Shuffle avg. time (s)	0	-1	0	1	1	2	3	14	0	0	0	14	0	13	0	12
Worse performing Shuffle (s)	1	-2	0	0	0	4	6	50	0	0	0	14	0	13	0	12
Reduce tasks avg. time (s)	0	0	4	7	10	22	0	160	0	0	5	69	15	-11	-1	272
Worse performing reduce task (s)	0	0	5	38	14	34	-1	255	0	0	5	69	15	-11	-1	272
CPU time spent (s)	-2	-8	-2	-2	0	-8	-6	4	0	1	-3	4	4	-2	1	16
Total duration (s)	**0**	**-1**	**8**	**42**	**16**	**37**	**6**	**264**	**0**	**0**	**2**	**83**	**14**	**7**	**0**	**284**
No. Map tasks on remote node	**4**	**8**	**5**	**3**	**2**	**3**	**9**	**6**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**2**
No. Reduce tasks on remote node	**2**	**2**	**2**	**2**	**2**	**2**	**0**	**2**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**0**

The last two rows show the number of Map/Reduce tasks executed on the remote node. A negative value indicates that a particular time value vas smaller for the multicluster.

In Job 2, the Map function inverts its key/value pairs, making temperature occurrences the keys, and emits them to the Reduce function that outputs the received pairs. The sorting by the occurrence is done by the framework as explained in Section I.B.

Note that default MapReduce configuration parameters have been used. For the specification please refer to the Cloudera documentation [6].

III. RESULTS

Table II shows MapReduce tasks' execution times for Job 1 and Job 2, for each test case. The presented time values are the differences between the execution times for multicluster and single cluster runs. The table also shows the total CPU time differences and total jobs' durations differences. The last two rows, relevant only for the multicluster run, indicate the number of Map and Reduce tasks executed on the remote computer.

Fig. 4 shows total durations of the MapReduce jobs in the single cluster and in the multicluster run, for each test case.

IV. DISCUSSION

It is evident from Fig. 4 that the introduction of a remote node may introduce a significant slowdown in MapReduce performance. The slowdown depends on two parameters. One is the amount of data coming out from Map tasks, whereas the other is the number of tasks executed on a remote node (the last two rows in Table II). The amount of data coming from Map tasks increases with the test case, as the number of columns in the data increases (see Table I). The jobs' execution times are similar until Case 4, regardless whether there are tasks' executions on the remote node or not.

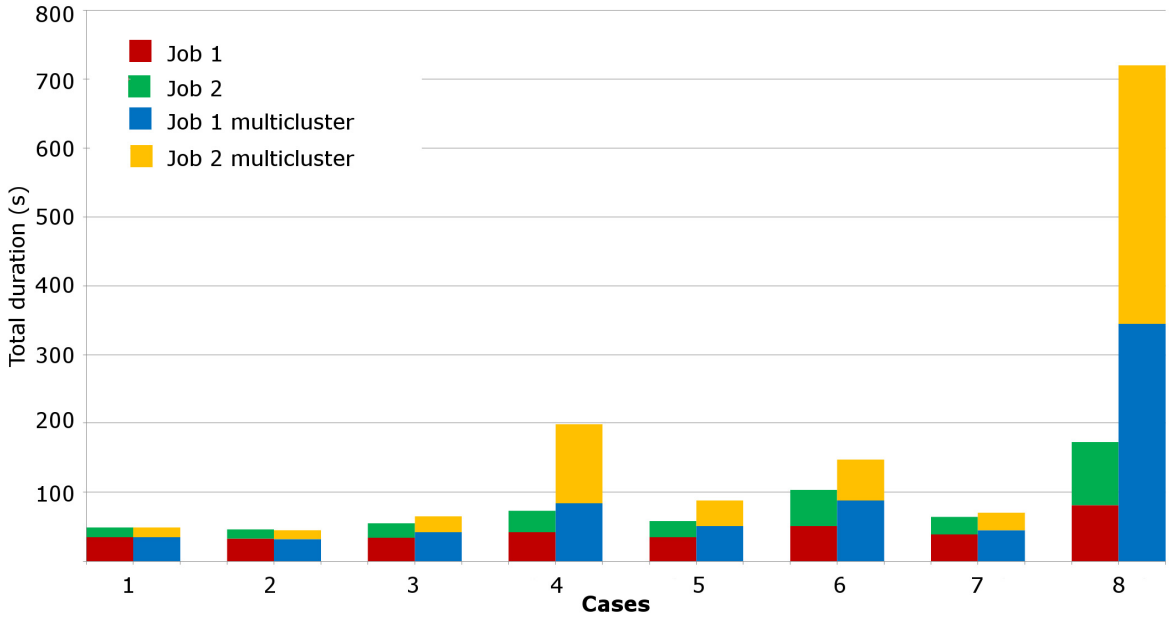


Figure 4. Jobs' total execution times for the single cluster (left columns) and for the multicluster (right columns)

The data amount significantly increases in Case 4, which evidently influences the execution times of the Shuffle and Reduce tasks (Table II). This result was expected, because in this case the remote node was allocated for Map and Reduce tasks, which imposed a heavy data communication between local and remote nodes (Fig. 1) with significantly slower communication than in the local cluster.

The same pattern can be observed for all the succeeding cases. Particularly interesting is Case 7 with a low increase in execution time for Job 1, and no increase in execution time for Job 2. Table II reveals that there were no Reduce tasks for Job 1 on the remote node, which restricted the necessary communication with the remote node only in one direction – from the remote node to the local nodes. That obviously decreased the performance but not as significantly as if the remote node was also used for Reduce tasks. For Job 2, all processing was local, so there was no decrease in performance.

For Case 8, the amount of the data is the highest and consequently is the slowdown. For Job 2, there were no Reduces on the remote node, but the worst Reduce task's execution time is relatively very long compared to the increase in the Shuffle tasks' execution times. We suspect that it is probably because of possible instabilities in the Reduce tasks' merging procedures, influenced by the introduction of the remote node.

A comparison of the MapReduce performance in our multicluster, to the performances previously investigated for the HDFS [5], reveals that the MapReduce performance is not as decreased as the HDFS performance. Even in the Case 8, which is the worst case, it is not decreased by a level of magnitude as the HDFS performance approximately is. This result was to be expected, because Hadoop does its best to run MapReduce jobs locally on the computers where the data resides. Furthermore, the communication to and from the distant node is consisted of the data coming out of Map tasks that can be, depending of a job at hand, significantly smaller than the source data. Additionally, the amount of data sent to Reduce tasks may be decreased by using Combiner functions.

The addition of the remote node in the presented use case means increase of the cluster storage resources for 20%, since the initial number of computes is five. The small number of the computers used, means that the chances for the remote computer to be allocated for a Map or Reduce task are higher than in the case when there are more computers in the cluster. Hence, the MapReduce performances should be higher as the ratio of the number of local and remote computers increases.

V. Conclusion

Adding remote nodes to a Hadoop cluster is a viable option for increasing storage resources. The MapReduce performances may or may not decrease, which depends on the concrete data analyses application, the ratio of the number of local and remote computers, and on connection bandwidth to remote computer or computers.

We have observed that the Reduce tasks execution times may increase even without obvious justifications in communication latencies. We suspect this result is due to possible instabilities in the merging processes on Reduce nodes. This issue needs to be further explored.

The Apache Hadoop MapReduce implementation may be upgraded for a multicluster environment with a decision algorithm that would prefer local computers to the remote. The algorithm should take into account the amount of data expected to come from Map or Combiner tasks into, which depends on the concrete application, and consider the connection bandwidth to the remote nodes.

Acknowledgment

The authors are grateful to Turboinštitut d.d. and CHS d.o.o., both from Ljubljana, Slovenia, for granting a full access to their storage and computational resources.

References

[1] "Welcome to Apache™ Hadoop®!," Oct., 2012; http://hadoop.apache.org/.

[2] J. Dean, and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in OSDI'04, 2004.

[3] B. Franks, "What is big data and why does it matter?," *Taming the big data tidal wave: finding opportunities in huge data streams with advanced analytics*, pp. 3-29, Hoboken, New Jersey: John Wiley & Sons, Inc., 2010.

[4] D. Borthakur *et al.*, "Apache hadoop goes realtime at Facebook," in ACM SIGMOD International Conference on Management of Data, 2011, pp. 1071-1080.

[5] I. Tomasic, J. Ugovsek, A. Rashkovska, and R. Trobec, "Multicluster Hadoop Distributed File System," in MIPRO, 2012 Proceedings of the 35th International Convention, 2012, pp. 314-318.

[6] I. Cloudera. "CDH Proven, enterprise-ready Hadoop distribution – 100% open source," Oct, 2012; http://www.cloudera.com/hadoop/.

[7] R. Trobec, M. Šterk, S. Almawed, and V. M., "Computer simulation of topical knee cooling," *Comput. biol. med,* vol. 38, pp. 1076-1083, 2008.

[8] G. Kosec, Šarler, Božidar, "Solution of a low Prandtl number natural convection benchmark by a local meshless method.," *International journal of numerical methods for heat & fluid flow,* vol. 23, pp. 22, 2013.

[9] U. Stepišnik, and G. Kosec, "Modelling of slope processes on karst," *Acta Carsologica,* vol. 40, no. 2, pp. 267-273, 2011.

[10] L. Wang *et al.*, "G-Hadoop: MapReduce across distributed data centers for data-intensive computing," *Future Generation Computer Systems,* vol. 29, no. 3, pp. 739-750, 2013.

[11] M. Niemenmaa *et al.*, "Hadoop-BAM: Directly manipulating next generation sequencing data in

the cloud," *Bioinformatics,* vol. 28, no. 6, pp. 876-877, 2012.

[12] J. Lin, and C. Dyer, "Limitations of MapReduce," *Data-Intensive Text Processing with MapReduce*, Synthesis Lectures on Human Language Technologies, pp. 143-145: Morgan & Claypool Publishers, 2010.

[13] T. White, "MapReduce," *Hadoop: The Definitive Guide*, pp. 15-40, Gravenstein Highway North, Sebastopol: O'Reilly Media, Inc., 2010.

[14] T. White, "The Hadoop Distributed Filesystem," *Hadoop: The Definitive Guide*, pp. 41-73, Gravenstein Highway North, Sebastopol: O'Reilly Media, Inc., 2010.

[15] I. Cloudera, "Cloudera Manager Free Edition User Guide," *https://ccp.cloudera.com/download/attachments/22151829/CM-4.1-free-user-guide.pdf?version=1&modificationDate=1354668306000*, Dec, 2012.

[16] T. White, "MapReduce Types and Formats," *Hadoop: The Definitive Guide*, pp. 189-224, Gravenstein Highway North, Sebastopol: O'Reilly Media, Inc., 2010.

Approaches to DNA *de novo* assembly

Ivan Sović*, Karolj Skala*, Mile Šikić**
* Ruđer Bošković Institute/Centre for Informatics and Computing,
Bijenička cesta 54, 10 000 Zagreb
** University of Zagreb Faculty of Electrical Engineering and Computing/Department of Electronic
Systems and Information Processing, Unska 3, 10 000 Zagreb
ivan.sovic@irb.hr, skala@irb.hr, mile.sikic@fer.hr

***Abstract*— DNA is the basic building block of all known life, accounting for all the diversities in nature. Determining the DNA of an individual organism is performed through a process called DNA *sequencing*. Although several different sequencing technologies do exist, they are limited and are able to acquire relatively short sequence reads. One of the approaches to sequencing involves randomly breaking a long DNA molecule into small fragments and sequencing only those fragments. Due to the random positioning of fragments on the source DNA, majority of them overlap, and provide the necessary information to combine them back together. The process of reconstructing the original DNA sequence from fragment reads is called DNA *assembly*. Assembly is a computationally very intensive process that may take days, or even weeks to produce the sequence of a more complex organism. Reconstructing a DNA sequence in the absence of a previously reconstructed reference sequence from a similar organism is called *de novo* assembly. De novo assembly methods currently provide the only means to discover new, previously unknown sequences, and are currently indispensable in biological research.**

In this paper, short descriptions of the sequencing process and the current sequencing platforms are given. DNA assembly process is thoroughly described, and the analysis of several *de novo* approaches used for assembly are presented. Overview and description of existing software tools are given, including some parallel implementations. As a conclusion, aspects of possible future development of DNA assembly are considered.

***Index Terms*—DNA sequencing, assembly, overlap, layout, consensus, de Bruijn, parallel**

I. Introduction

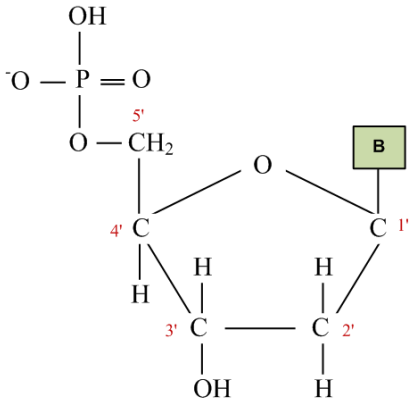
An extremely important research subject in biology is the determination of the sequence of naturally occurring *deoxyribonucleic acid* (DNA) molecules. The DNA is a long molecule consisting of a large number of simple components called *nucleotides* or *bases*, that comprise a long chain (sequence). The process of resolving the structure of the DNA molecule is called *DNA sequencing*. Currently, several methodologically different sequencing techniques exist: *first generation* techniques mostly based on Sanger's method of sequencing, and *next generation* sequencing techniques (NGS), most prominent of them being Roche's 454 Life Sciences, Illumina Solexa, Applied Biosystems' SOLiD systems [1] and Ion Proton from Life Technologies. General differences amongst the first and the next generation sequencers are the lower cost and higher throughput of next generation sequencers, but with disadvantages of shorter read lengths and higher error rates [2] [3].

Commonly, DNA molecules have the length from a few million up to a few billion base-pairs (bp), dependant on the species (i.e. *Escherichia coli*'s DNA has the length of 4.6 million base-pairs, while *Homo sapiens*' DNA is the size of 3.3 billion base-pairs). Limited by the current sequencing technology, the entire DNA can not be read at once. Instead, multiple smaller reads of different parts of the DNA molecule are produced, each about 30-400 bp long, representing fragments of the original DNA molecule. The process of reconstruction of the original DNA sequence from read fragments is called DNA sequence *assembly*. The assembly process is generally based on finding overlaps among reads, and joining them into *contiguous sequences* (contigs). The explanation for such an approach is based on the fact that a large number of copies of the same DNA molecule have been randomly broken into fragments, thus causing the fragments of different DNA copies to partially overlap.

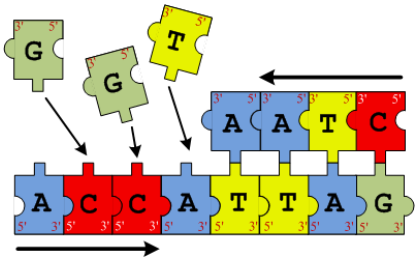
Computational DNA assembly is currently the only means of obtaining the sequence of a DNA molecule. This makes the research in the field of assembly of crucial importance to the development of biology and furthering the understanding of life.

II. DNA sequencing

The molecule of DNA is a long, unbranched, paired *polymer* chain, formed of the same four types of *monomers* called *nucleotides*, connected in a long linear sequence that encodes the genetic information [4]. Monomers are small molecules, such as amino acids, nucleotides and glucose, which may bind to other monomers in order to form larger molecules called polymers. Nucleotides are the basic building blocks of DNA, consisting of a sugar (deoxyribose) with a phosphate group attached to it, and a *base* [4]. The base of a nucleotide can be one of the following: *adenine* (A), *cytosine* (C), *thymine* (T) and *guanine* (G). In graphical or textual representations of DNA molecules, nucleotides are commonly denoted through the abbreviations of their base names A, T, C and G. In this case, the DNA can be written as the sequence of these four letters. Nucleotide A from one strand of the DNA always bonds to a T nucleotide of the other strand of the DNA, and vice versa. For this reason A and T nucleotides are



(a) Generic nucleotide structure



(b) Structure of the DNA chain

Fig. 1: a) A generic molecular structure of a single nucleotide, consisting of a sugar backbone, a phosphate group and a base (denoted by B). The base of a nucleotide in the DNA can be one of the four: A, C, T and G. b) Depiction of the double-stranded DNA molecule, composed of nucleotides.

called *complements*. The same applies for the C and the G nucleotides.

Figure 1a shows the molecular structure of a nucleotide, while Figure 1b depicts the double-stranded nature of DNA consisting of A, C, T and G nucleotides. It is important to note that the carbon atoms on Figure 1a are marked $1' - 5'$. Their order in the molecule is not coincidental, and is of great importance for the formulation of DNA. Bonds between nucleotides are always formed between their 3' and 5' ends via the phosphate group, creating a polymer chain composed of a repetitive sugar-based backbone with a series of bases protruding from it [4], that represents a single strand of the DNA. During synthesis, DNA is always extended in the "direction" from its $5'$ end to $3'$ end - in other words, the $5'$ carbon atom of a nucleotide that is to be added to an existing chain is always bonded with the last $3'$ carbon atom of that chain through covalent bonds between the phosphate group of one nucleotide, and the $3'$ carbon in the deoxyribose ring of the other. By convention, a DNA sequence is always read from its $5'$ end towards the $3'$ end.

Significant effort has been put into the development of methods for determining the exact sequence of nucleotides comprising a DNA molecule. The process of "reading" the sequence is commonly referred to as *DNA sequencing*. Methods and technology for performing DNA sequencing are broadly divided into two groups: the first generation sequencing based on Sanger's method, and the next (or second) generation sequencing (NGS) based on several different approaches. However, although these techniques can read the sequence of nucleotides, the maximum length that can currently be read at once is limited to at most 1000 nucleotides. This is an extremely important limitation, since the DNA can be up to a few billion nucleotides long. All current solutions of this problem are based on the same principle: multiple copies of the DNA molecule are randomly broken into small fragments of such size that can be read by modern technology. The fragments are then multiplied into large number of copies if/as required by the sequencing platform. The method of sequencing the entire genome by dividing it in smaller fragments is called *whole genome shotgun* (WGS) sequencing. Given large enough number of DNA samples, randomly created fragments will start to overlap with each other, which gives the only available information about putting them together. The average number of sequences that independently contain a certain nucleotide is called the *depth of coverage* (often only *coverage*) [5]. Combining the fragments back into a single chain of nucleotides is called DNA *assembly*, while the programs that perform this process are called genome *assemblers* [6].

Sequencing of a DNA fragment can be performed in two ways: sequencing only one strand of DNA from the $5'$ end, called the *single-end* (SE) reads, or sequencing the fragment in both strands from their $5'$ ends, called the *paired-end* (PE) reads. Sequences of a PE read are also often called *mate-pairs*. The method of sequencing an entire genome and producing PE reads is called the *double barrelled whole genome shotgun* (DBWGS) sequencing.

To help gain a better understanding of DNA assembly methods, a short description of current sequencing technologies and their limitations is required. The first generation sequencing, or Sanger sequencing, is based on the modification of the process of DNA synthesis. During this process, the two chains of DNA are separated, followed by the addition of nucleotides that are complements to those contained within the chains. Sanger's method contains modified nucleotides in addition to the normal ones, where each modified nucleotide is missing an oxygen on the $3'$ end, and has also been fluorescently marked [1]. Integrating these nucleotides into a DNA chain causes a halt (or termination) in the elongation process. Using capillary electrophoresis, sequences are separated by their length (mass) and the termination base is read. This process of sequencing can deliver read lengths up to 1000 bases, high raw accuracy, and allow for 384 samples to be sequenced in parallel, generating 24 bases per instrument second [1]. Great shortcomings of this methods are its high price of about \$10 per 10000 bases, and long sequencing time. Examples of machines used for first generation sequencing include the *Applied Biosystems Prism 3730* and the *Molecular Dynamics MegaBACE* [7].

Commercial NGS DNA sequencing platforms today include the *Genome Sequencer* from *Roche 454 Life Sciences*, the

Solexa Genome Analyzer from *Illumina*, the *SOLiD System* from *Applied Biosystems*, *Ion Storm* and *Ion Proton* from *Life Technologies*, the *Heliscope* from *Helicos* and the *Polonator* [7]. There are several different methods that these technologies use for sequencing: *hybridization to tiling arrays*, parallelized *pyrosequencing* (454), *reverse termination* (Solexa), *ligating degenerated probes* (SOLiD) and *single molecule sequencing* (Heliscope) [1]. Read lengths obtained from the NGS reads are in the 400bp range (454), the 100bp range (Solexa and SOLiD), or less [7].

Compared to Sanger sequencing, next generation sequencers have characteristic error profiles, different in case of each of the technologies [7]. These error profiles can include inaccurate determination of simple sequence repeats, enrichments of base call error toward the $3'$ ends of reads, insertion or deletion errors during homopolymer runs and other [7] [8]. Error profiles of some of the platforms have been published and are available [7].

During the process of determining nucleotide types (commonly referred to as *base-calling*), information about the quality of a certain base-call is often available. Software tools such as Phred [9] read chromatogram files and analyse peaks in order to call bases, while assigning them a quality score that is proportional to the probability that the base is called wrong. Quality scores are used by some assembly tools to enhance efficiency; and also for comparison of different sequencing technologies.

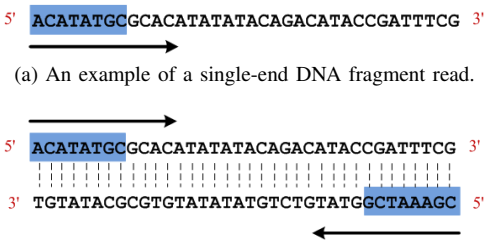
III. Fragment assembly

All sequencing platforms produce observations of the target DNA molecule in the form of reads - sequences of single-letter base calls with a numeric quality value [7]. An example of a single-end read and a paired-end read is shown in Figure 2a and Figure 2b, respectively. Genome assembly is commonly divided into two groups:

- *De novo* - assembly of sequence reads into longer contiguous sequences called *contigs*, followed by the process of correctly ordering contigs into *scaffolds*, in the absence of a reference genome sequence [8] [10] [11].
- *Mapping/reference* - when available, a pre-existing reference genome sequence can be used to *align* the reads of a newly obtained genome, thus avoiding the process of constructing complex data structures as with de novo assembly [8] [10].

As a result of the de novo assembly stage, a set of scaffolded contigs is available. Besides ordered contigs, scaffolds also contain gaps of imprecise length separating the contigs [12]. Scaffolds are also often called *supercontigs*. Low risk assemblies consistent with almost all of the detectable pairwise read overlaps are called *unitigs* [12].

Sequence *alignment* is the procedure of comparing two or more sequences by searching for a series of individual characters (in this case nucleotides) or character patterns that are in the same order in all sequences [13]. In most cases, the sequences cannot be aligned perfectly due to their mismatch. In an optimal case, non-identical characters and gaps are



(a) An example of a single-end DNA fragment read.

(b) An example of a paired-end DNA fragment read.

Fig. 2: Examples of DNA fragments read from a) only one end and b) from both ends.

placed, in order to match together as many identical or similar characters as possible. Popular tools that perform sequence alignment include BLAST [14], BLAT [15] and BLASTZ [16].

The first sequence fragment assembly algorithms, developed in 1980s, were focused on creating multiple overlapping alignments of the reads to produce a layout assembly of the data [17]. The reads were typically sequences obtained using Sanger's method. The alignment is used to read the consensus sequence, from which the DNA sequence is inferred. The described method was called the *overlap-layout-consensus* approach (OLC). Currently, the OLC approach is one of the two mainstream approaches of the de novo DNA assembly, and is especially used during the assembly of long, high-quality reads [17]. Examples of OLC assembly tools include Celera Assembler [18], Arachne [19] and Newbler [20].

Another very important, and currently most utilized approach in combination with NGS data, is based on the de Bruijn graph. This approach was first presented by Pevzner et al. in 2001 [21], and is currently the prevalent representation in short-read assemblers. Unlike the computationally expensive step of determining overlap amongst reads, the de Bruijn graph is constructed from fixed-length subsequences (of length k) that were consecutive in larger sequences, and that have a perfect $(k-1)$-suffix to $(k-1)$-prefix overlap. Examples of assemblers based on the de Bruijn approach include Euler [21], AllPaths [22] and Velvet [23].

Both the OLC and the de Bruijn approaches are graph based. On the other hand, different approaches also do exist. They include greedy string-based methods i.e. implemented in SSAKE [24], and hybrid methods such as Taipan [25].

Recently, several parallelized versions of assemblers have appeared. This step in the development of assembly tools is of great importance, since it can provide rapid execution of otherwise highly computationally intensive task. Examples include ABySS [3]; a method by Kundeti et al. [26]; and a parallelization of the Euler assembler [27].

Errors in assembly can occur for two main reasons: incomplete or incorrect information provided to the assembler and the limitations of the assembly algorithm [6]. The real-world WGS data that contains errors can induce the following problems in overlap and de Bruijn graphs [7]:

1) *Spurs* - short dead-end branches (divergences) of the main path (Figure 3a). Possible causes include sequenc-

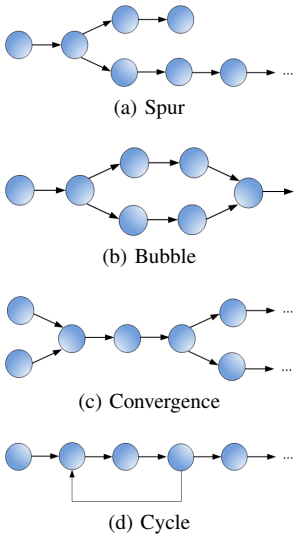


Fig. 3: Problems that can appear during the process of building graph representations for DNA de novo assembly.

ing errors toward one end of a read, and low coverage.

2) *Bubbles* - divergence of a path into two branches that afterwards join together again into one path (Figure 3b). Possible causes include sequencing errors toward the middle of a read, and by polymorphisms in the target. It is important to note that efficient bubble detection is not trivial.
3) *Converging and diverging paths* - inverse definition than for the bubbles, two paths converge into one, that later diverges again into two separate paths (Figure 3c). Possible causes are repeats in the target genome.
4) *Cycles* - paths that converge on themselves (Figure 3d). Possible causes are repeats in the target genome.

Although sequencing technology has greatly improved, still no sequencing platform can produce sufficient data to assemble a complete genome from a single experiment [28]. One of the key problems in shotgun sequencing is caused by repeats in genome sequences [29]. Repeats, in cases when they are longer than fragment reads, induce problems during the overlap phase of assembly. As a result, assemblies result in fragmented contigs, separated by gaps. Repeat caused fragmentation is more pronounced in the NGS technology, since reads are generally of smaller size [28]. Other than repeats, gaps can also be caused by other, technology-specific reasons. Assemblies containing gaps are called *draft assemblies*, while the process of filling the gaps is called *finishing*. Finishing involves obtaining missing sequences, improving low quality regions, resolving misassemblies and ordering scaffolds; and often accounts for the majority of labour and cost of genome projects [28]. In order to fill gaps, further sequencing is usually required, either by amplifying and sequencing fragments spanning gaps, or by resequencing the entire DNA. Tools that perform gap closure and finishing do exist, such as IMAGE [30] and Consed [31], although some hybrid assemblers perform gap closure indirectly by incorporating reads from multiple sequencing platforms (i.e. Velvet [23], CABOG [32]).

In case paired-end sequencing was performed, mate-pair information can be used to provide additional constraints to the placement of reads during the assembly. These constraints limit the possible traversals of the assembly graph, allowing longer segments of the genome to be unambiguously reconstructed [33].

In the next subsections, the following notation is used to formally define the problem of DNA fragment assembly. Let Σ be the alphabet consisting of four elements $\Sigma = \{A, C, T, G\}$. Then, Σ^k is a set of all strings of length k over the alphabet Σ. Also, let $v \in \Sigma^k$. The length of a string v is denoted by $|v|$. Additionally, given $1 \leq i \leq j \leq |v| \mid i, j, |v| \in \mathbb{N}$, the i-th element of a string v is denoted with $v[i]$, while a substring of a string v is denoted by $v[i, j]$, where i is the starting position of the substring within v, and j the end position, inclusive. The string of length k is called a *k-mer*, while the *k-spectrum* of v is the set of all k-mers that are substrings of v. A *k-molecule* is a pair of k-mers which are reverse complements of each other [34]. If v and w are strings, and there exists a maximal length non-empty string z that is a prefix of v and a suffix of w, then it is said that v overlaps w, with the length of overlap $ov(w, v) = |z|$. This definition is not symmetric [34]. If v does not overlap w, then $ov(w, v) = 0$.

A. Overlap-Layout-Consensus approach (OLC)

Let $S = \{s_1, s_2..., s_n\}$ be a set of non-empty strings over an alphabet Σ. An overlap graph of S is a complete weighted directed graph where each string in S is a vertex and the length of an edge between vertices x and y, $x \to y$, is $|y| - ov(x, y)$ [34].

An example of assembly using the overlap graph is depicted in Figure 4. The construction of the overlap graph is the first step of the OLC approach. General process of this approach consists of three main phases [7] [8] [35], from which it derives its name:

1) *Overlap* - reads are compared to each other in a pairwise manner to construct an overlap graph.
2) *Layout* - the overlap graph is analysed and simplified with the application of graph algorithms to identify the appropriate paths traversing through the graph and produce an approximate layout of the reads along the genome. The ultimate goal is a single path that traverses each node in the overlap graph exactly once [6].
3) *Consensus* - multiple sequence alignment of all reads covering the genome is performed, and the original sequence of the genome being assembled is inferred through the consensus of the aligned reads.

A very important observation is that the identification of a path that traverses every node in the graph only once is a computationally difficult *Hamiltonian path* problem. The problem is also known as the Travelling Salesman Problem, and is *NP-complete* [36].

Discussion

The three-phase definition of the OLC approach causes that the algorithms utilizing this method are naturally implemented through a modular design, which allows for simpler modification and optimization of distinct assembly steps. Another advantage of OLC assemblers is that the overlaps among reads may vary in length, which equips them to handle the data either from Sanger or NGS platforms.

On the other hand, the processing cost of the overlap phase is very high, as this phase is very time consuming for determining the overlap of every pair of reads in the data set. Although OLC approach is capable of handling NGS data, this data commonly consists of an order of magnitude more reads than were commonly generated in Sanger-based projects [6] (for which the OLC was initially designed), causing a significant increase of the overlap calculation time (quadratic complexity) [37]. However, some OLC assemblers such as Edena [38] and Shorty [39] have been well optimized to handle such data, and even outperform some other, NGS specialized assemblers [6]. Also, finding the Hamiltonian path in the layout step is still not solvable in polynomial time, which makes OLC additionally difficult and dependant on various heuristics to produce reliable results [21].

Implementations

Celera [18] is an assembler that was first designed to handle data from Sanger sequencers, but was later revised and optimized for the Roche's 454 NGS data. Its revised pipeline, called CABOG, is robust to homopolymer run length uncertainty, high read coverage and heterogeneous read lengths. It also uses mate pair information to merge sets of unitigs into larger structures.

Newbler [20] is a well known and widely used assembly tool, distributed with Roche 454 Life Science sequencers. Since newer revisions it provides scaffold building from paired-end data, exploits coverage to handle base calling errors, uses instrument metrics to overcome inaccurate calls of the number of bases in homopolymer runs, and implements OLC twice - once to generate unitigs from reads, and second time to generate larger contigs from unitigs.

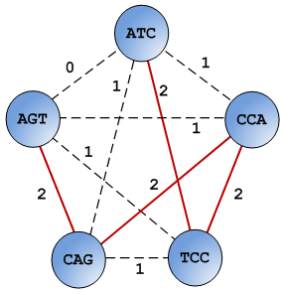


Fig. 4: An example of assembly using the overlap graph by finding a Hamiltonian path. In this simple example, the set of input fragments consists of five reads of equal length, $\{ATC$, CCA, CAG, TCC, $AGT\}$. The result of assembly is the reconstruction of the sequence $ATCCAGT$.

Edena [38] and Shorty [39] are OLC assemblers that are targeted for the short reads generated by Solexa and SOLiD sequencing platforms. Edena discards duplicate reads, finds all perfect error free overlaps and removes individual overlaps that are redundant with pairs of other overlaps. Shorty is designed for a special case when a small number of long reads are available, and uses them as seeds to recruit short reads and their mate pairs, and iteratively repeats this process to create larger contigs from previously constructed contigs.

Parallelization of the OLC approach can be easily achieved by distributing the overlap computation on different processing units [6]. However, OLC based parallel assembly tools are rare and hard to find, as most recent implementations focus on the de Bruijn approach. PASQUAL [40] is an example of a parallel OLC assembler, designed for shared memory parallelism using *OpenMP*. Authors claim to have achieved better performance and results than any other tested assembly tool they have ran. In [41] authors have implemented a parallel OLC and a parallel de Bruijn assembler, in order to analyse the differences in their scalability and efficiency.

B. The de Bruijn approach

As in the case of overlap graphs, let $S = \{s_1, s_2..., s_n\}$ be a set of non-empty strings over an alphabet Σ. Given a positive integer parameter k, the de Bruijn graph $B^k(S)$ is a directed graph with vertices defined by $\{d \in \Sigma^k \mid \exists i \ni d \subseteq s_i\}$, and edges defined by $\{d[1..k] \rightarrow d[2..k+1] \mid d \in \Sigma^{k+1}, \exists i \ni d \subseteq s_i\}$ [34]. A vertex of the de Bruijn graph is commonly identified through its associated k-mer.

An example of the assembly using the de Bruijn graph is depicted in Figure 5. The process of constructing a de Bruijn graph consists of the following steps [42]:

1) *Construction of k-spectrum* - reads are divided into overlapping subsequences of length k.
2) *Graph node creation* - a node is created for every $(k-1)$-spectrum of each unique k-mer.
3) *Edge creation* - a directed edge is created from node a to node b iff there exists a k-mer such that its prefix is equal to a and its suffix to b.

The first application of de Bruijn graphs for genome assembly was proposed by Pevzner et al. in 2001 [21]. By converting the set of reads into edges of the de Bruijn graph, the assembly problem becomes equivalent to finding an Eulerian path in the graph - a path that uses every edge in the graph - for which efficient algorithms do exist [6]. In this case, assembly is a by-product of the graph construction, which proceeds quickly using a constant-time hash table lookup for the existence of each k-mer in the k-spectrum [7]. However, there can be an exponential number of distinct Eulerian paths in a graph, while only one can be deemed to be the correct assembly of the genome. To reduce the complexity of the problem, heuristics are usually applied to the constructed graph. The graph is filtered of erroneous occurrences such as bubbles, spurs, cycles and convergences/divergences, and nodes that are unambiguously connected by an edge are merged together [3].

Discussion

A great advantage of the de Bruijn approach is that explicit computations of pairwise overlaps are not required, unlike the case of the OLC approach. Finding pairwise overlaps is a computationally very expensive process, and thus the de Bruijn approach provides great performance when very large data sets are given (such as NGS data) [42]. Finding an Eulerian path through the edges of a graph is known to be a polynomial problem, and therefore can be executed more efficiently than finding a Hamiltonian path.

Since there can be very many Eulerian paths in a graph, constraints have to be added in order to find the path that represents the actual genome. These constraints can make the assembly much more difficult, and transform this polynomial problem to a NP-hard problem, as shown in [34]. Also, de Bruijn graphs are very sensitive to sequencing errors and repeats, as they lead to new k-mers, adding to the graph complexity [42]. Additionally, although de Bruijn approach can handle both Sanger and NGS data, by dividing long reads from Sanger sequencing into short k-mers, there is an effective loss of long range connectivity information implied by each read [6].

Implementations

Euler [21] was the first assembler based on the de Bruijn approach, developed for Sanger reads, and later modified for short Roche 454 reads, very short unpaired Illumina/Solexa reads and paired-end Solexa reads. It first filters input reads by detecting erroneous base calls through noting low-frequency k-mers, and then constructs two k-mer graphs at different values of k, and compares them.

Velvet [37] is a collection of methods for assembly using de Bruijn graphs. It consists of two parts: first, called Tour Bus, removes sequencing errors and handles polymorphisms, and second aimed to resolve repeats based on the available information from low coverage long reads or paired shotgun reads. It applies a series of heuristics based on local graph topology, coverage, sequence identity and paired-end

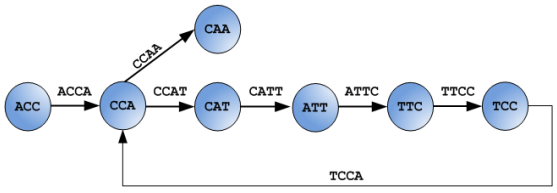


Fig. 5: An example of assembly using the de Bruijn graph by finding a Eulerian path. A sample sequence $ACCATTCCA$ was fragmented into two reads, $\{ACCATTC, ATTCCAA\}$. Reads were used to obtain the k-mer spectrum for $k = 4$, $\{ACCA, CCAT, CATT, ATTC, TTCC, TCCA, CCAA\}$, and the $(k-1)$-mer spectrum, $\{ACC, CCA, CAT, ATT, TTC, TCC, CAA\}$, required to construct the graph. The original sample sequence is fully reconstructed by traversing the graph.

constraints, in order to reduce graph complexity.

AllPaths [22] was initially designed for large genome assembly using paired-end short reads obtained from Solexa sequencers, but now works both on small and large (mammalian size) genomes. AllPaths partitions the graph to resolve repeats by assembling the regions that are locally non-repetitive, and then glues the local graphs where they have overlapping structure.

Unlike the case of the OLC approach, many parallel assembly tools exist that use the de Bruijn approach, some being specially designed for such purpose, while others incorporated support for multi-threading in later versions. ABySS [3] was among first de Bruijn assemblers to utilize the parallel paradigm. ABySS distributes the de Bruijn graph over multiple computers by placing k-mers over available nodes, with the location of a k-mer determined by a simple hash function [3]. ABySS uses MPI (*Message Passing Interface*) protocol for communication between nodes. Ray [43] also distributes the graph across multiple computers through MPI, but also allows simultaneous assembly of reads from a mix of high-throughput sequencing technologies. Contrail presents a cloud computing solution to assembly of large genomes, and relies on *Hadoop* to iteratively transform an on-disk representation of the assembly graph [44]. Contrail is currently being extended to support the OLC approach, that will be of great use for long reads. As mentioned in the previous subsection, implementations and comparison of a de Bruijn and a OLC assemblers is given in [41]. An out-of-core algorithm for constructing large bi-directed de Bruijn graphs in $\Theta(n/p)$, and with $\Theta(n)$ message complexity, where p is the number of processors, and n is a constant degree polynomial in p, is presented in [26].

C. Other methods

Among described approaches, two main other methods exist: greedy and hybrid. Greedy algorithms were common in genome assemblers for Sanger data [6], and were part of the first NGS assembly packages [7]. They operate in simplest and most intuitive fashion: reads are joined together into contigs iteratively, starting with the reads that have best overlaps; the process is repeated until there are no more reads or contigs that can be joined. This approach may not lead to a globally optimal solution, for instance, in a case where a contig at hand takes on reads that, combined with another contig, might result with that contig to grow even larger. Examples of assemblers that use greedy approach include SSAKE [24] that aggressively assembles short nucleotide sequences by progressively searching for the longest possible overlap between sequences through a prefix tree; VCAKE [45] that implements an iterative extension algorithm that, unlike its predecessors, can incorporate imperfect matches during contig extension; and SHARCGS [46] that adds pre- and postprocessing to the SSAKE assembler, where at preprocessing it filters the raw read set three times with different settings, constructs three assemblies, and in the postprocessing step merges them.

Hybrid assembly has twofold meaning. For one, it is an approach where data obtained from several different se-

quencing platforms is simultaneously used for the assembly of one genome [47]. The data is commonly acquired both from: platforms with long reads, *base-space* and moderate throughput (i.e. Roche 454), and platforms with short reads, base- or *color-space* and high throughput (i.e. Solexa and SOLiD). Here, base-space and color space stand for notations of DNA sequences outputted from DNA sequencers. In base-space, a DNA sequence is given plainly through the four types of nucleotides (A, C, T, G), while in color-space each base is coded by two values called colors, with each color used in coding two neighbouring bases. Color-space coding provides additional error checking during the sequencing process. Example of an assembler that uses this type of the hybrid approach is HybridVelvet [47] which presents a one-step pipeline incorporating Velvet, that can take any long base-space sequences and short color-space reads as inputs and export assembled base-space contigs and scaffolds.

The second meaning of hybrid assembly is an approach that combines both greedy and graph based assembly methods [25]. An example of such an assembler is Taipan [25]. It uses greedy extensions for contig construction, but at each step also constructs enough of the corresponding read graph in order to make better decisions of the continuation of assembly.

Additionally, some assemblers use a suffix trie approach as in [48], and an approach using combinatorial algorithms as in [49]. These and other different approaches are seldom found in assembly tools.

D. Comparison of implementations

The efficiency and performance of assemblers is generally assessed through the size and the accuracy of assembled contigs and scaffolds [8], and resource consumption [35]. The resource consumption of a certain assembler includes the total processing time, and RAM occupancy. The common size measurements include:

- *N50 length* - the longest length such that at least 50% of all base pairs are contained in contigs of this length or larger [35]. Provides a standard measure of assembly connectivity, higher N50 lengths indicate better performance of an assembler.
- *Minimum contig length*
- *Maximum contig length*

On the other hand, the accuracy of assemblies is generally difficult to measure, since these estimates can only be given in comparison to a known correct (benchmark) sequence, such as a reference genome or an artificially created data set. Accuracy measurements include:

- *Sequence coverage* - percentage of the benchmark sequence covered by output contigs.
- *Assembly error rate* - output contigs are aligned to the benchmark sequence, and the number of mismatched bases are calculated. Assembly error rate is the percentage of mismatched bases of the total bases in the aligned contigs in the reference sequence.

It is important to note that N50 length can only be comparable between different assemblers when each is measured with the same combined length value [8]. In [35], authors compared seven different assembly tools, and have discovered an interesting pattern in assembly connectivity measured by N50 length and compared against the depths of coverage. Although N50 initially followed the increase of the depth of coverage, N50 reached a plateau (converged) when the depth of coverage exceeded a certain threshold [35]. To compare the N50 values among assemblers, authors in [35] have used the value of N50 at the same depth of coverage where N50s for all assemblers have converged.

Qualitative comparison of several assembly tools is given in Table I. More detailed description of benchmarking procedures and results of various assembly tools can be found in comparative and review articles such as [6] [7] [8] [35] [50] [51].

IV. Discussion about future prospects

With all the presented aspects of DNA assembly taken into account, there are several possible directions in which the future research can progress beyond the state-of-the-art. One such direction is the exploration of new assembly algorithms and data types that can be utilized, reducing computational complexity and increasing assembly accuracy. Also, assembly tools have to become capable of handling and processing large amounts of data more rapidly, while reducing memory consumption. From the implementation point of view, two possible optimizations can be observed: for one, compression techniques can be applied on the sequence data thus reducing storage and operational size, and two, the computationally intensive task of assembly can be parallelized and deployed on *cloud* platforms. Clouds can provide scalable and on-demand resources to researchers, reusable workflows and reusability of results. *CloudMan* is an excellent example of an easy to use cloud manager platform that addresses the stated possibilities and has already been used for bioinformatics purposes [53]. To address the constatation about compression, an assembler that uses this approach has recently been published [52]. On assembly of a *C. elegans* data set, SGA used only 4.5GB of RAM, whereas AbySS required 14.1GB, Velvet $23.0GB$ and SOAPdenovo 38.8GB [52]. However, SGA resulted in longer execution time than other stated assemblers.

On a side note - recently, an interesting announcement regarding sequencing technologies has been made public. A company called *Life Technologies* has released a press statement claiming that they have created a sequencing machine *Benchtop Ion Proton^{TM} Sequencer* capable of sequencing the entire genome under a day, for the price of $1000, and have already started taking orders for the device [54]. A $1000 dollar genome barrier has been a long goal for the developers of genome sequencing technologies, as it is supposed to bring genome sequencing to masses through hospitals, medical institutions and research facilities.

V. Conclusion

While the length of sequences that can be read at once is still limited, the de novo assembly methods are irreplaceable. Cur-

TABLE I: Comparison of performance and quality of several assembly tools on sequences obtained from a *C. elegans* data set. Results were taken from [52].

	SGA	Velvet	ABySS	SOAPdenovo
Scaffold N50 size	26.3 kbp	31.3 kbp	23.8 kbp	31.1 kbp
Aligned contig N50 size	16.8 kbp	13.6 kbp	18.4 kbp	16.0 kbp
Mean aligned contig size	4.9 kbp	5.3 kbp	6.0 kbp	5.6 kbp
Reference bases covered	96.2 Mbp	94.8 Mbp	95.9 Mbp	95.1 Mbp
Mismatch rate at all assembled bases	1 per 21,545 bp	1 per 8,786 bp	1 per 5,577 bp	1 per 26,585 bp
Contigs with split/bad alignment	458 (4.4 Mbp)	787 (7.2 Mbp)	638 (9.1 Mbp)	483 (4.4 Mbp)
Total CPU time	41 h	2 h	5 h	13 h
Maximum memory usage	4.5 GB	23.0 GB	14.1 GB	38.8 GB

rently, they provide the only means to discover new, previously unknown sequences, which is essential for characterization of the biological diversity of our world [8]. However, although genome assembly is the only way to obtain the information contained in the genetic code of an organism, complexity and accuracy of current assembly methods is still greatly influenced by read errors and imprecise algorithms, as well as the number and size of the reads obtained from sequencing.

One can say that the areas of sequencing, assembly and bioinformatics in general have never been as vivid and interesting as now, while also providing great opportunities for research that will find real-life applications, and ultimately, effect our lives for better.

ACKNOWLEDGEMENTS

The author would like to acknowledge the support of the scientific research project "Methods of scientific visualization" $(098 - 098\ 2562 - 2567)$, funded by the *Ministry of Science, Education and Sports* of the *Republic of Croatia*. Additionally, author would like to thank prof.dr.sc. Karolj Skala and doc.dr.sc. Mile Šikić for their support and guidance during the work on this project.

REFERENCES

[1] E. Pettersson, J. Lundeberg, and A. Ahmadian, "Generations of sequencing technologies," *Genomics*, vol. 93, no. 2, pp. 105–111, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.ygeno.2008.10.003

[2] J. R. Miller, S. Koren, and G. Sutton, "Assembly algorithms for next-generation sequencing data," *Genomics*, vol. 11, no. 6, pp. 315–327, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20211242#

[3] J. T. Simpson, K. Wong, S. D. Jackman, and et al., "Abyss: A parallel assembler for short read sequence data," *Genome research*, vol. 19, no. 6, pp. 1117–1123, Jun. 2009. [Online]. Available: http://dx.doi.org/10.1101/gr.089532.108

[4] B. Alberts, A. Johnson, J. Lewis, M. Raff, and K. Roberts, *Molecular biology of the cell*. 978-0-8153-4r05, 2008.

[5] S. Taudien, I. Ebersberger, G. Glöckner, and M. Platzer, "Should the draft chimpanzee sequence be finished?" *Trends in Genetics*, vol. 22, no. 3, pp. 122–125, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16430990

[6] M. Pop, "Genome assembly reborn: recent computational challenges," *Briefings in Bioinformatics*, vol. 10, no. 4, pp. 354–366, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19482960

[7] J. R. Miller, S. Koren, and G. Sutton, "Assembly algorithms for next-generation sequencing data," *Genomics*, vol. 95, no. 6, pp. 315–327, Jun. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.ygeno.2010.03.001

[8] S. Bao, R. Jiang, W. Kwan, X. Ma, and Y.-Q. Song, "Evaluation of next-generation sequencing software in mapping and assembly," *J Hum Genet*, vol. 56, no. 6, pp. 406–414, Jun. 2011. [Online]. Available: http://dx.doi.org/10.1038/jhg.2011.43

[9] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, "Base-Calling of automated sequencer traces using phred. i. accuracy assessment," *Genome Research*, vol. 8: 175-185, 1998.

[10] Applied Biosystems, "Applied Biosystems SOLiD 3 Plus System: De Novo Assembly Protocol," Life Technologies Corporation, Tech. Rep., 2010.

[11] L. Jorde, *Encyclopedia of genetics, genomics, proteomics, and bioinformatics: 8 volume set*, ser. Encyclopedia of Genetics, Genomics, Proteomics, and Bioinformatics: 8 Volume Set. John Wiley & Sons Ltd., 2005, no. s. 8. [Online]. Available: http://books.google.hr/books?id=fytFAQAAIAAJ

[12] S. Koren, J. Miller, B. Walenz, and G. Sutton, "An algorithm for automated closure during assembly," *BMC Bioinformatics*, vol. 11, no. 1, p. 457, 2010. [Online]. Available: http://www.biomedcentral.com/1471-2105/11/457

[13] *Bioinformatics: Sequence and Genome Analysis, Second Edition*, 2nd ed. Cold Spring Harbor Laboratory Press, Jul. 2004.

[14] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, pp. 403–410, 1990.

[15] W. J. Kent, "BLAT—the BLAST-like alignment tool," *Genome research*, vol. 12, no. 4, pp. 656–664, Apr. 2002.

[16] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller, "Human-mouse alignments with BLASTZ," *Genome research*, vol. 13, no. 1, pp. 103–7+, 2003.

[17] D. Edwards, J. Stajich, and D. Hansen, *Bioinformatics: tools and applications*. Springer Science+Business Media, 2009.

[18] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter, "A whole-genome assembly of Drosophila." *Science (New York, N.Y.)*, vol. 287, no. 5461, pp. 2196–2204, Mar. 2000. [Online]. Available: http://dx.doi.org/10.1126/science.287.5461.2196

[19] "Arachne: A whole-genome shotgun assembler," vol. 12.

[20] Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, no. 7057, pp. 376–380, Sep.

[21] P. A. Pevzner, H. Tang, and M. S. Waterman, "An eulerian path approach to dna fragment assembly," *Proceedings of the National Academy of Sciences*, vol. 98, no. 17, pp. 9748–9753, Aug. 2001. [Online]. Available: http://dx.doi.org/10.1073/pnas.171285098

[22] J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, and D. B. Jaffe, "Allpaths: De novo assembly of whole-genome shotgun microreads," *Genome Research*, vol. 18, no. 5, pp. 810–820, 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18340039

[23] D. R. Zerbino and E. Birney, "Velvet: Algorithms for de novo

short read assembly using de bruijn graphs," *Genome Research*, vol. 18, no. 5, pp. 821–829, May 2008. [Online]. Available: http://dx.doi.org/10.1101/gr.074492.107

[24] "Assembling millions of short DNA sequences using SSAKE," *Bioinformatics*, vol. 23, no. 4, pp. 500–501, Feb. 2007. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btl629

[25] B. Schmidt, R. Sinha, B. Beresford-Smith, and S. J. Puglisi, "A fast hybrid short read fragment assembly algorithm," *Bioinformatics*, vol. 25, pp. 2279–2280, September 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1596446.1596453

[26] V. K. Kundeti, S. Rajasekaran, H. Dinh, M. Vaughn, and V. Thapar, "Efficient parallel and out of core algorithms for constructing large bi-directed de bruijn graphs," *BMC Bioinformatics*, vol. 11, no. 1, pp. 560+, 2010. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-11-560

[27] W. Shi and W. Zhou, "A parallel euler approach for large-scale biological sequence assembly," *Information Technology and Applications, International Conference on*, vol. 1, pp. 437–441, 2005.

[28] X. Yang, D. Medvin, G. Narasimhan, D. Yoder-Himes, and S. Lory, "Clog: A pipeline for closing gaps in a draft assembly using short reads," in *Proceedings of the 2011 IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences*, ser. ICCABS '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 202–207. [Online]. Available: http://dx.doi.org/10.1109/ICCABS.2011.5729881

[29] E. Arner, *Solving repeat problems in shotgun sequencing*, 2006. [Online]. Available: http://books.google.hr/books?id=eBMINAAACAAJ

[30] I. Tsai, T. Otto, and M. Berriman, "Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps," *Genome Biology*, vol. 11, no. 4, p. R41, 2010. [Online]. Available: http://genomebiology.com/2010/11/4/R41

[31] D. G. C. Gordon and P. Green, "Consed: A graphical tool for sequence finishing," *Genome Research*, vol. 8:195-202, 1998.

[32] J. R. Miller, A. L. Delcher, S. Koren, E. Venter, B. P. Walenz, A. Brownley, J. Johnson, K. Li, C. Mobarry, and G. Sutton, "Aggressive assembly of pyrosequencing reads with mates," *Bioinformatics*, vol. 24, pp. 2818–2824, December 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1520805.1520818

[33] J. Wetzel, C. Kingsford, and M. Pop, "Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies," *BMC Bioinformatics*, vol. 12, no. 1, p. 95, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/95

[34] P. Medvedev, K. Georgiou, G. Myers, and M. Brudno, "Computability of models for sequence assembly," in *In WABI*, 2007, pp. 289–301.

[35] Y. Lin, J. Li, H. Shen, L. Zhang, C. J. Papasian, and H.-W. Deng, "Comparative studies of de novo assembly tools for next-generation sequencing technologies," *Bioinformatics*, vol. 27, no. 15, pp. 2031–2037, Jun. 2011. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btr319

[36] D. e. Davendra, *Traveling salesman problem, theory and applications.* Rijeka: InTech, 2010.

[37] D. R. Zerbino, "Genome assembly and comparison using de bruijn graphs," 2009.

[38] D. Hernandez, P. François, L. Farinelli, M. Østerås, and J. Schrenzel, "De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer," *Genome Research*, vol. 18, no. 5, pp. 802–809, 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18332092

[39] M. S. Hossain, N. Azimi, and S. Skiena, "Crystallizing short-read assemblies around seeds." *BMC Bioinformatics*, no. S-1.

[40] X. Liu, P. R. Pande, H. Meyerhenke, and D. A. Bader, "Pasqual: A parallel de novo assembler for next generation genome sequencing." [Online]. Available: http://www.cc.gatech.edu/pasqual/index.html

[41] M. Ahmed, I. Ahmad, and S. U. Khan, "A comparative analysis of parallel computing approaches for genome assembly," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 3, no. 1, pp. 57–63, Mar. 2011. [Online]. Available: http://dx.doi.org/10.1007/s12539-011-0062-0

[42] K. Löwe, "Mapping of wdlps neochromosomes," Otto-von-Guericke-Universität Magdeburg, Tech. Rep., Nov. 2010.

[43] S. Boisvert, F. Laviolette, and J. Corbeil, "Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies," *Journal of computational biology a journal of computational molecular cell biology*, vol. 17, no. 11, pp. 1519–1533, 2010. [Online]. Available: http://www.liebertonline.com/doi/abs/10.1089/cmb.2009.0238

[44] [Online]. Available: http://sourceforge.net/apps/mediawiki/contrail-bio/index.php?title=Contrail

[45] "Extending assembly of short DNA sequences to handle error." *Bioinformatics (Oxford, England)*, vol. 23, no. 21, pp. 2942–2944, Nov. 2007. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btm451

[46] "SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing." *Genome research*, vol. 17, no. 11, pp. 1697–1706, Nov. 2007. [Online]. Available: http://dx.doi.org/10.1101/gr.6435207

[47] F. M. You, Y. Wang, Y. Q. Gu, M.-C. Luo, N. Huo, K. R. Deal, J. Lee, G. R. Lazo, and J. D. nad Olin D. Anderson, "Hybridvelvet: de novo hybrid assembly of next generation sequencing data with velvet," *Plant & Animal Genomes XIX Conference*, Jan. 2011.

[48] C. Barnes, G. Dalgleish, G. Jared, K. Swift, and S. Tate, *Trie-based data structures for sequence assembly*. Springer, 1997, no. August, p. 206–223. [Online]. Available: http://www.springerlink.com/index/31617V66L054T702.pdf

[49] J. Kececioglu and E. Myers, "Combinatorial algorithms for DNA sequence assembly," *Algorithmica*, vol. 13, no. 1, pp. 7–51, Feb. 1995. [Online]. Available: http://dx.doi.org/10.1007/BF01188580

[50] M. Ruffalo, T. LaFramboise, and M. Koyutürk, "Comparative analysis of algorithms for next-generation sequencing read alignment," *Bioinformatics*, vol. 27, no. 20, pp. 2790–2796, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btr477

[51] W. Zhang, J. Chen, Y. Yang, Y. Tang, J. Shang, and B. Shen, "A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies," vol. 6, no. 3, 2011.

[52] J. T. Simpson and R. Durbin, "Efficient de novo assembly of large genomes using compressed data structures," *Genome Research*, Dec. 2011. [Online]. Available: http://dx.doi.org/10.1101/gr.126953.111

[53] E. Afgan, D. Baker, N. Coraor, H. Goto, I. M. Paul, K. D. Makova, A. Nekrutenko, and J. Taylor, "Harnessing cloud computing with galaxy cloud," *Nat Biotech*, vol. 29, no. 11, pp. 972–974, Nov. 2011. [Online]. Available: http://dx.doi.org/10.1038/nbt.2028

[54] [Online]. Available: http://www.lifetechnologies.com/global/en/home/about-us/news-gallery/press-releases/2012/life-techologies-itroduces-the-bechtop-io-proto.html

Ski Injury Triggers of Tibial Plateau Compression Fracture

R. Trobec[1], I. Belehar, J. Polajnar[2], M. Veselko[3]
[1]Jožef Stefan Institute, Department of Communication Systems, SI-100 Ljubljana, Slovenia
[2]Slovenian Environment Agency, Vojkova 1b, SI-1000 Ljubljana, Slovenia
[3]University Medical Centre Ljubljana, Zaloška 7, SI-1000 Ljubljana, Slovenia
roman.trobec@ijs.si, matjaz.veselko@kclj.si, janez.polajnar@gov.si

Abstract - **Fractures of the tibial plateau occur in 3.4% of skiing-related fractures and this number is significantly increasing. Despite experimental results, which suggest that significant forces (over 30 kN) are needed for a compression fracture of the tibial plateau, the fracture often occurs during an accident with no serious fall or collision. We show that a possible trigger for the injury is a sudden pull of a ski, towards the centre of the ski turn, due to a momentary reduction in the turn radius while passing a snow brim or hollow depression. A sudden deceleration of the skier's leg can initiate the tibial rotation, leg extension and its swing to backwards. The swing of the leg can end with a sudden stop because of limited movability in hip and knee joints. The rotational energy of the ski and lower leg produces an extensive torque in the knee joint and compression force on the tibial plateau that can cause a serious injury.**

I. INTRODUCTION

The number of skiing injuries with damages in lower extremities slightly decrease over the last decade [1],[2]. The tibial plateau fractures account for up to 0.52% of all ski injuries and 3.4% of all fractures [3]. This study is motivated by a hypothesis that several of non-documented less serious injuries have the same injury mechanism as the tibial fracture, however, with forces not strong enough for the compression fracture, but still sufficient for evident bone, muscle or ligament damages. The number of tibial fractures has increased over time; however, this change coincides with an aging of the population at risk, so that the trend over time is associated with an increase in the mean age of the skiing population [4],[5].

According to published experimental studies [6],[7], the typical compression fracture with crushing of the subchondral bone is the result of a predominantly compressive force, alone or in combination with a valgus force. It follows from experimental results in cadavers that compression forces above 30 kN can provoke tibial plateau fractures. In extension, the fracture site is anterior, but with flexion the fracture site moves posterior. Hyperextension and valgus displacement were suggested to be the predominant and characteristic mechanism of the compression fracture of tibial plateau in a conventional skier with 44 % of cases [8]. Other reported causes of tibial plateau fractures are: falls from a height with improper landing, accidental falls or twists, and the direct application of a force to the lateral aspect of the knee when hitting by another skier [9],[10]. No specific mechanisms were described in the literature to the fracture of tibial plateau on shaped skies.

Recently, a new injury mechanism was proposed [11] that is based on the swing of the inner leg and ski. The proposed injury mechanism could potentially produce high forces in a knee, even in cases with no accidental fall. It was shown that with typical skiing velocities of 40 km/h the compression forces can reach up to 70 kN. The proposed mechanism could explain also more frequent sport injuries that result in ligament injuries and less serious injuries of bones, also in a connection with other sport activities.

We show that a momentary reduction in the turn radius, while passing a snow brim or hollow depression, could produce a sudden pull of the inner ski, away from the body and towards the center of the ski turn. Consequently, a lateral rotation of the lower leg around the antero-posterior axis is initiated, together with a significant mass of the ski. The leg rotation is momentarily stopped because of limitations in knee and hip joints movability, which produce an almost instant increase in the compression force to the lateral condyle of the tibia. Large axial forces on the tibial plateau cannot result just from vertical forces produced during skiing but only from the axial forces resulting from an instant stop of the swinging leg. The potential injury is an unfortunate consequence of the interplay of different body parts and skiing equipment.

The injury can be triggered by a "sudden arrest of the ski", e.g. hitting an obstacle on the skiing surface, like snowball, snow heap, slalom pole, etc., however, the most dangerous consequences can be expected if the inner ski edge grabs the snow surface, which suddenly decelerates the inner ski, while the skier is still moving forward, initiating the leg and ski rotation. In this paper we analyze in details possible triggers for the pull-in of the inner ski based on the momentary reduction of the ski turn radius on a snow brim or a hollow depression.

II. MATERIALS AND METHOD

A. Model of skier position

The trigger for the compression fracture of the tibial plateau, even at low or medium velocities during the ski turn, is often provoked by a "sudden arrest of the ski". It can also be interpreted as a sudden deceleration of the

The authors acknowledge the financial support from the state budget by the Slovenian Research Agency under Grant P2-0095.

inner or outer ski, produced either by an obstacle on the skiing surface or by an unexpected reduction of the ski-turn radius. The first reason is often treated as obvious and is frequently declared as the "cause" of injury, while the second reason is less obvious; however, it can explain a few reports of injured skiers claiming that "nothing unusual happened" before the injury. We have focused on the second reason, because of increased popularity of carving techniques; however, the model also explains some other causes of tibial compression fractures.

The dynamic force $\mathbf{F}_t$ that acts on a skier with mass m and velocity v during a carving turn with radius r is shown in

Figure 1. It is a vector sum of the body-weight force $\mathbf{F}_g = m\mathbf{g}$, $\mathbf{g}$ representing the acceleration due to gravity, and the centrifugal force $\mathbf{F}_c = mv^2 / r$. For example, for r = 12 m and v = 11 m/s (≈ 40 km/h), $\mathbf{F}_c$ will be approximately equal to $\mathbf{F}_g$. The skier inclination angle α is equal 45° to reach the equilibrium and should be approximately the same as the inclination of skies. The skier will feel under his feet, a force $\mathbf{F}_t$ that is 40% greater than if he just stood on the surface. The necessary α for a general case can be calculated from v and r:

$$\alpha = \operatorname{atan}\frac{v^2}{rg}. \qquad (1)$$

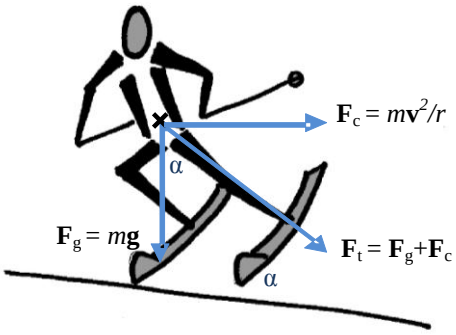


Figure 1: Main forces on the skier during a carving turn with radius r and velocity v. A skier inclination α is needed for the equilibrium.

B. *Model of ski turn radius*

To describe the initial conditions just before the fracture we list some basic mechanical characteristics of the carving turn. Shaped skies have curved edges of length L with a side-cut h, and a curvature is defined as a section of a circle with the radius R. Supposing that the ski is minimally inclined against the flat horizontal surface and loaded in its centre by a force directed perpendicular to the flat of the ski, it will bend, due to the curved edge. The curvature of the bent ski will adapt to a circle defined by the radius R. In the case of continuous pressure during skiing, without ski sliding, the ski would cut a circular trace into the snow surface. The trace corresponds to a circle with the radius R denoted by solid line in Figure 2. The radius of ski R can be calculated using the Intersecting Chord theorem:

$$R = \frac{\left(L^2 + 4h^2\right)}{8h}. \qquad (2)$$

For a carving ski of length L = 165 cm and side-cut h = 22 mm the theoretical ski radius, at very small non-zero inclinations α, is R = 12 m.

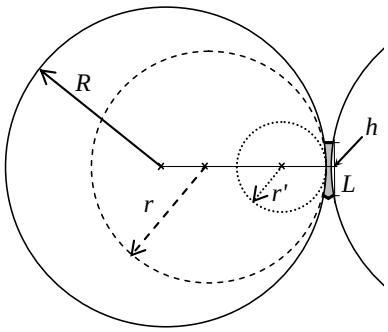


Figure 2: Carving ski (grey) with radii of carving turns R = 12, r = 8.5 and r' = 2.9 m.

During skiing, the force $\mathbf{F}_t$ that acts on the ski is directed perpendicular to the flat of the ski. In Figure 3, a cross-section of a ski (bold lines) with the inclination towards the flat horizontal surface α = 45°, is shown in unbent (above) and bent (below) position. Thinner lines of the unbent position (upper inclined line with bold and thin segments) denote the widest rear part of the ski. Assuming there is no sliding of the ski edge on the snow surface, the ski bends due to force $\mathbf{F}_t$ for a distance d, until the ski edge leans to the snow's surface (lower bold line). For example in Figure 3, for $d = h$ and α = 45°, the apparent "side-cut" is H and the turn radius r is reduced to 8.5 m (dashed circle in Figure 2).

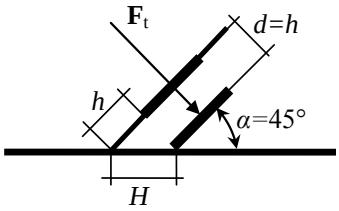


Figure 3: Bending of a ski inclined for α = 45°. Bold lines represent the cross-section of the unbent ski (above) and the bent ski (below), loaded with force $\mathbf{F}_t$.

If the inclination is still greater, the ski bends further. The apparent "side-cut" H is:

$$H = \frac{h}{\cos(\alpha)}. \qquad (3)$$

Replacing h in Eq. (2) with H, we obtain the radius of the carving turn r as a function of ski shape and its inclination:

$$r = \frac{L^2 + 4\left(\frac{h}{\cos(\alpha)}\right)^2}{8h}\cos(\alpha). \qquad (4)$$

C. *Model of ski pull*

If v and r are constant during the ski turn, the forces and inclination do not change throughout the curve. But, if on an uneven ground the ski slope suddenly changes for an angle φ, the inclination of the ski towards the surface

becomes α + φ. In Figure 4, a cross-section of uneven surface and both skies are shown in unbent and bent position. It is evident that the left inner ski experiences a sudden multiple-fold additional bending *H'* that reduces the radius of the ski turn for the same amount (Eq. (4)). The final reflection of the above circumstances, as felt from the skier, is a sudden pull of the inner ski away from the outer ski towards the centre of the turn radius.

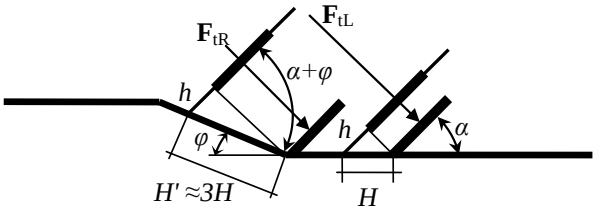


Figure 4. Cross-section of an uneven surface (brim) with a change of the slope φ = 31°. The skis are inclined by α = 45°. $\mathbf{F}_{tR}$ and $\mathbf{F}_{tL}$ are forces on the right and the left ski. The effective inclination of the inner (right) ski is 76°.

The same effect occurs in a spherical depression sketched in Figure 5, however, now in the same extent to both skies. For example, in a depression with the radius, which is the same as the radius of the ski turn *r*, and with the inclination α = 45°, $h=h_d$, $d \approx 2.4h$ and $H \approx 2.6h$. Note, that for a ski with length L, *h* can be obtained from Eq. (2) if R is replaced with the radius of depression r_d.

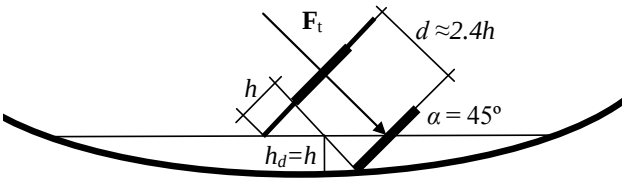


Figure 5: Increased bending of the ski, inclined for 45°, in a spherical depression.

With no inclination the ski fits the shape of the depression, i.e. bends to the radius of the depression. We see from Figure 6 that with higher ski inclinations the turn radius in the depression can be further decreased.

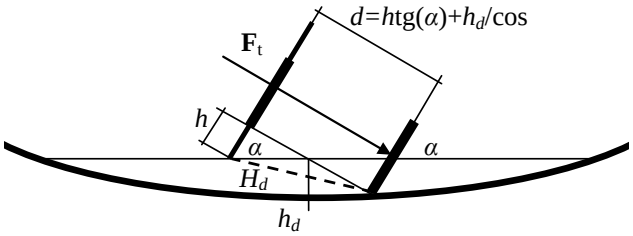


Figure 6. Ski bending in a spherical depression with increased inclination α.

The ski bending in the depression can be evaluated by:

$$H_d = \sqrt{\left(h^2 + \left(\mathrm{htg}(\alpha) + h_d / \cos(\alpha)\right)^2\right)}, \qquad (5)$$

which can cause again a rapid and unexpected reduction of the turn radius and pull-in of both skis.

The presented geometrical causes of the turn radius reduction apply to whichever ski approaches the brim/depression first, regardless of whether it is the inner or the outer ski. The timing is crucial for an eventual subsequent injury. When approaching to a brim laterally, with the brim and skies almost parallel, the inner ski usually reaches the brim a moment earlier as the other ski, which could be enough to trigger an accident. The reports of injured skiers [12] confirm that the inner leg is more exposed, while it was injured in almost all cases.

Beside crossing the obstacle first, another weakness of the inner leg is also that it is not fully loaded and therefore prone to uncontrolled changes of its route. Finally, we have shown that more loaded and inclined outer ski has smaller turn radius, on the other hand, the inner ski turn radius should be slightly smaller because of geometrical reasons. These facts inherently dictates different ski turn radii, but both skies are usually identical in their shape. Several initiatives from the ski production industry have promoted asymmetric skies [13] that could increase the effectiveness of skiing and possibly its safety, if the inner ski edge in a turn grabs less aggressive as the outer ski edge [14].

III. RESULTS

We evaluated the proposed models by custom Matlab programs.

A. *Turn radius*

The ski turn radius is in a relation with ski parameters and skier inclination, which depends on the velocity (see Eqs. (2), (3), (4)). In Figure 7, the ski turn radius is shown as a function of inclination (black) and as a function of velocity (blue).

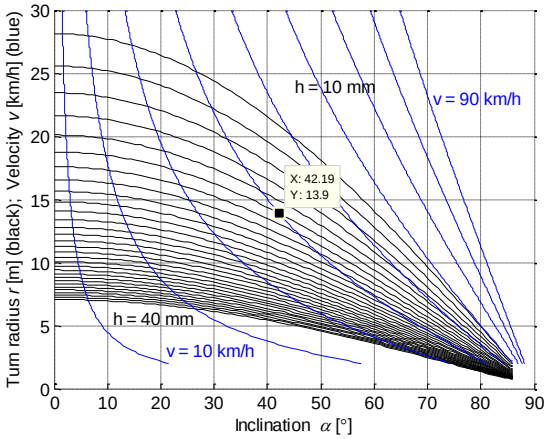


Figure 7: Theoretical ski turn radii and skier velocities as a function of ski inclination for side-cuts h = 10 to h = 40 mm.

Note, that at zero inclination ski turn radii *r* are equal to ski radii *R*. We can see that with a specific velocity different turn radii with corresponding inclinations are possible. Usually, the choice depends on the skier's ability. However, higher velocities imply larger ski radii and skier inclinations. For example, at 40 km/h, the ski radius 18.5

m reduces to the ski turn radius of 13.9 m at skier inclination of α = 42°. The ski radius 13.2 m reduces to r = 5 m and α = 69°. We can see that velocities higher than 60 km/h are unmanageable with practical values of ski turn radii. Therefore, the sliding of both ends of ski should also be considered.

We have measured ski turn radii and ski inclinations for 10 turns of 4 expert skiers with two different ski radii: R = 13.2 m and R = 18.5 m on the 27 % slope and at approximate velocities of 40 km/h (see Figure 8).

Figure 8: Measurement of a ski turn radius and ski inclination.

The statistics of measured results expressed as (mean ± SD) is shown in Table I.

Table I: Measured results

	r [m]	α [°]
R=13.2 m	10.7 ± 3.1	55.1 ± 5.3
R=18.5 m	16.3 ± 3.7	48.2 ± 4.1

The reductions of the turn radii are smaller than theoretical values while ski inclinations are in the same range. We have estimated from the recorded videos that the inclination of skier's centre of gravity is approximately 10° smaller as the ski inclination. The difference in inclinations depends on the skier's technique.

We further modelled the sliding as approximately proportional to the reduction of the turn radius $k(R-r)$, where R is the ski radius, r is the ski turn radius and k is an experimentally determined constant, which depends on the skier's ability and technique and should be personally determined. k = 0.4 was selected in this study. The corrected ski turn radii shown in Figure 9 are obtained by adding up the calculated ski turn radius from Eq. (3) and the correction for sliding $0.4(R-r)$.

If we compare the results from Figure 7 and Figure 9, we see that at 40 km/h, the ski radius 18.5 m results in the ski turn radius of 16.5 m at the skier inclination of 37.4°. For R = 13.2 m we get r = 10.7 m and α = 50°, which is in a good agreement with measurements.

B. *Ski pull on brim*

While crossing a brim, the inner leg hits the brim first. It then experiences a sudden decrease of the inner leg ski turn radius, deceleration of the leg, followed by its rotation and swing in an opposite direction to that of the skiing direction.

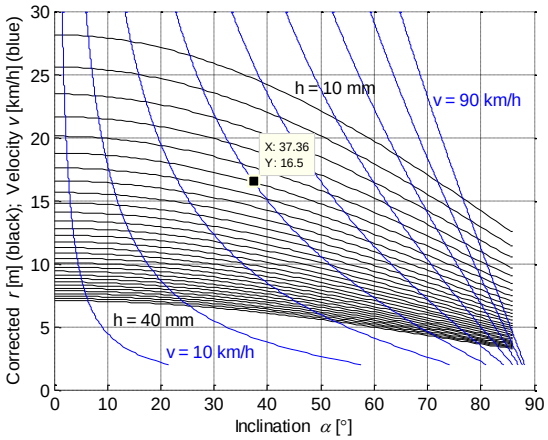


Figure 9: Corrected ski turn radii because of ski sliding as a function of ski inclination and skier velocity.

In Figure 10, a skier position is shown just before the injury on an uneven snow surface.

Figure 10: Sudden pull-in of the inner ski due to a decrease of the ski turn radius on an uneven snow surface.

The evaluation of the theoretical ski track with ski inclination α = 45°, over a 0.5 m wide and 0.3 m high brim (ski slope of the brim inclined for φ = 31º), and with the assumption of no-sliding conditions is shown in Figure 11. The skiing direction is marked with arrows. Compared to the ski turn radius on the lower flat surface before the brim, which is r = 8.5 m, the ski turn radius over the brim becomes threefold smaller r' = 2.9 m (dotted circle in Figure 2).

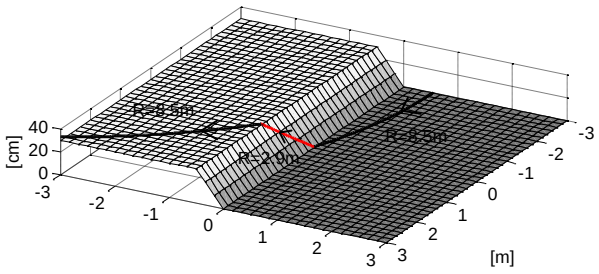


Figure 11: A track of right ski turn over a brim. The ski turn radius is reduced for almost a factor of 3.

C. *Ski pull in depression*

Replacing h in Eq. (2) with H_d from Eq. (5), we get the expression for r in a hollow depression, which indicates

that the depression significantly decreases the ski turn radius. The results are shown in Figure 12.

The ski turn radius is decreased to a value smaller than the depression radius. For example, in a depression with radius r_d = 6 m, the initial ski turn radius is reduced to 3.9 m at the skier inclination of 28.8°.

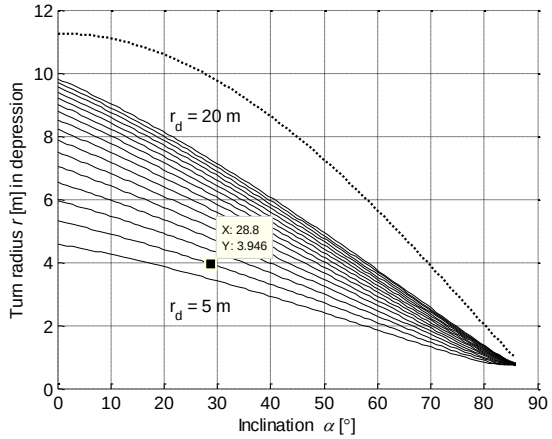


Figure 12: Reduction of ski turn radii in depressions with different radii r_d. Dotted curve shows the theoretical turn radius on a flat surface, all as a function of skier inclination and for a ski radius 11.5 m.

The speed of change in centrifugal force depends on the ski length and on skiing velocity. With shorter skis and higher velocities the change is more pronounced. For example at 40 km/h and ski length of 1,6 m the whole ski will approach the depression in 0.14 second and the centrifugal force will increase in the same period. It seems to be impossible to change the body inclination in such a short time. Skiers experience unstable positions, which can disable the proper musculature reaction and can trigger an accident or injury.

In Figure 13, two tracks of ski turns are shown: first, on a flat surface (blue) and second, through a depression (red) with radius r_d = 6 m and depth 1 m. The ski turn radius in the depression is reduced to 4 m, consequently, the centrifugal force increased for almost a factor of 3.

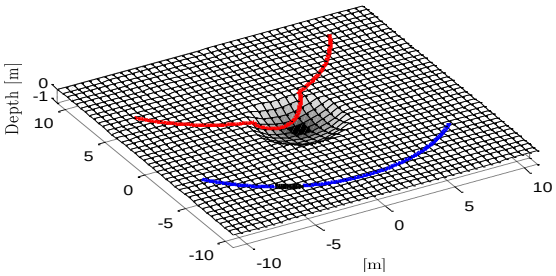


Figure 13: A track of ski turn on a flat surface with radius 12 m (blue). The track of same skies through a depression with radius 6 m (red).

IV. CONLUSION

The presented study explains a possible scenario for a tibial compression fracture, even in cases with no fall or collision. Our analysis suggests that a sudden inward turn of the inner ski with a significantly reduced radius and a subsequent swing of the inner leg, is a possible mechanism for compression fracture of the lateral tibial plateau that has not been described before. In the proposed mechanism the compression fracture of tibia is not caused by increased vertical forces arising from skier's dynamics or fall, but by axial forces resulting from the leg and ski swing.

We confirmed that the proposed mechanism can produce sufficient forces for the tibial compression fracture and other related injuries. We shown that the accident can be initiated by a sudden change of the ski turn radius due to a change of the shape of the ski surface, like a brim or depression, or due to hitting laterally to a fixed object on the skiing surface. Because the tibial compression fractures commonly occur with no fall, research of possible causes is vital to prevent further increase of injury rate. It is important that skiers and professionals engaged in ski sports are aware of the presented knee injury mechanisms in order to develop appropriate preventative measures.

REFERENCES

[1] Johnson, R.J., Ettlinger, C.F., Shealy, J.E., 2008. Update on injury trends in alpine skiing. J ASTM International 5, Paper ID JAI 102046.

[2] Schulz, D. 2010. http://www.ski-online.de/xfiles_a6/1291724839_7.pdf Unfälle und Verletzungen im alpinen Skisport - Zahlen und Trends der Saison 2009/2010. ARAG and SIS, Düsseldorf.

[3] Shealy, J.E., Johnson, R.J., Ettlinger, C.F., 2003. Femur and tibial plateau fractures in alpine skiing. Skiing trauma and safety: Fourteenth Volume, ASTM STP 1440, Jonhson RJ, Lamont M, Shealy JE, Eds., ASTM International, West Conshohocken, PA, 140-148.

[4] Burtscher, M., Gatterer, H., Flatz, M., Sommersacher, R., Woldrich, T., Ruedl, G., et al., 2008. Effects of modern ski equipment on the overall injury rate and the pattern of injury location in alpine skiing. Clinical Journal of Sport Medicine 18, 355-357.

[5] Hunter, R., 1999. Skiing injuries. Am J Sports Med 27, 381-389.

[6] Kennedy, J.C., Bailey, W.H., 1968. Experimental tibial-plateau fractures: studies of the mechanism and a classification. J Bone Joint Surg Am 50, 1522-1534.

[7] Senter, C., Hame, S.L., 2006. Biomechanical analysis of tibial torque and knee flexion angle: Implications for understanding knee injury. Sports Medicine 36, 635-641.

[8] McConkey, J.P., Meeuwisse, W., 1988. Tibial plateau fractures in alpine skiing. Am J Sports Med 16, 159-164.

[9] Bahr, R., Krosshaug, T., 2005. Understanding injury mechanisms: A key component of preventing injuries in sport. British Journal of Sports Medicine 39, 324-329.

[10] Howe, J., 1983. *Skiing mechanics*. Poudre Press: Colorado.

[11] Veselko, M., Kosec, G., Trobec, R., Compression fracture of the tibial plateau – a ski-injury mechanism, 2012, in review.

[12] Veselko, M., Polajnar, J., 2008. New skiing techniques – new injuries? Analysis of ski injuries in 2004/2005. Zdrav Vestn 77, 499-504.

[13] Ski prévu pour la pratique du ski alpin, EU-patent EP 1 484 091 B1, Salamon S.A., 2008.

[14] Smučka z asimetričnimi karakteristikami, SI-patent 23531, ELAN SKUPINA, holdinška družba, d.o.o., 2012.

A model of selective stimulation and ENG recording in the human left vagus nerve

Polona Pečlin and Janez Rozman

ITIS d. o. o. Ljubljana, Center for Implantable Technology and Sensors, Lepi pot 11
1000 Ljubljana, Republic of Slovenia

Abstract – In this study we have developed a model of using a thirty-nine-electrode spiral nerve cuff for selective stimulation of fibres in the left vagus nerve of a man to control the heart rate and tachycardia in particular. Furthermore, according to our recent work on experiments in human, we predicted the precisely defined stimulus shape and parameters. In a forthcoming study we are to conduct experiments to determine the effect of cervical selective vagus nerve stimulation (VNS) on the atrial fibrillation in a man using a developed model and designed implantable thirty-nine-electrode spiral cuff. Specifically, we intend to look at the effects on the atrial fibrillation when the superficial compartments of the left vagus nerve, including the cardiac branches containing nerve fibres A, B and C, are selectively stimulated with the precisely defined stimulation parameters.

I. INTRODUCTION

The use of VNS for treating and controlling a variety of medical, psychiatric and neurological disorders has seen significant growth over the past few decades [1]. In particular, stimulation of the left vagus nerve has been the subject of considerable research [2, 3]. One of its main functions is to monitor and control the activity of the internal organs and glands such as the heart, lung, stomach, bladder and pancreas. The gross anatomy of the vagus nerve of man has been described by numerous investigators [4, 5, 6, 7]. It carries both afferent and efferent nerve fibres, but conflicting conclusions have been drawn from the data available regarding these components and their distribution. According to observations made by Schnitzlein et al. [8], the total number of nerve fibres in the mid-cervical vagus nerve of a man varies from 45.110 to 153.123 fibres and considerable variation exists in different individuals and even in opposite sides of the same autopsy. The vagus nerve of a man is composed predominantly of nonmyelinated axons while among the myelinated fibres the small myelinated nerve fibres are predominance. In the cervical distribution, the myelinated nerve fibres were divided into large (10 μ or larger), medium (3 μ to 9 μ), and small (less than 3 μ) where approximately 80 % of the myelinated axons are less than 3 μ, 15 % are medium in size and a variable number 10 μ or larger. In adults, a number of the large fibres vary considerably.

The principal functions of the heart are regulated by vagal and sympathetic fibres. The right vagus nerve primarily innervates the SA node while the left vagus innervates the AV node; however, there can be significant overlap in the anatomical distribution.

The effect of vagal afferent signalling on cardio-inhibition has been well known for over 130 years. The carotid sinus nerve, a branch of the glossopharyngeal nerve, innervates the carotid sinus baroreceptors while aortic arch baroreceptors are innervated by the aortic nerve, which then combines with the vagus nerve.

VNS has been proposed as a method for treating various heart conditions, including supraventricular arrhythmias, angina pectoris, heart failure and atrial fibrillation f9, 10, 11]. It was demonstrated that favourable shifts towards increased cardiac vagal modulation can be achieved by vagal afferent nerve stimulation. This phasic effect appears to operate through central medullary pathways. Thus, by engaging vagal afferent fibres in humans, there is the possibility that one can exploit the benefits of central cardio-inhibition without adversely affecting heart rate, respiration or hemodynamics. Implantable prosthetic devices can be controlled using signals recorded in parts of the body where sensation has been retained [12].

Tripolar cuffs give global ENG, as nerve cuff signals enable considerable spatial and temporal averaging. Besides, when the outer electrodes of the tripolar cuffs are short-circuited and a bipolar amplifier is used, the tripolar cuffs assist in the additional reduction of interference from sources outside the cuff. Anyhow, nerve cuffs provide useful and reproducible whole-nerve recordings [13]. Unfortunately, these whole-nerve recordings provide only one channel of information from the aggregate activity of many nerve fibres.

Information recorded selectively with a multi-electrode cuff could be more effectively used for multiple channel closed loop control of implantable stimulators that selectively activate different neural pathways. Therefore, it is desirable to selectively record an autonomic nerve activity for the control of artificial heart systems. Yambe et al. showed [14] that the tonus of the vagal nerve was significantly altered before body motion with hemodynamic changes, suggesting the possibility of prediction. The results suggest that information from vagus nerves may help to control implantable devices that could assist in suppression of atrial fibrillation.

II. METHODS

Neurocardiac control by VNS requires the development of appropriate theoretical models and multi-electrode systems that are able to activate selectively a certain population of nerve fibres within different compartments in the nerve. Specifically, the acute chronotropic and inotropic cardiac responses of VNS generally requires mostly B fibre stimulation. Namely, by not stimulating large-diameter fibres, such stimulation generally avoids the adverse effects sometimes associated with the recruitment of such large fibres (dyspnea and hoarseness). Stimulation of small-diameter fibres is avoided because these transmit pain sensations and are important for the regulation of reflexes such as respiratory ones.

It was shown in the work of Schachter [15] that most of the large myelinated fibres A, and many of medium size B, are given off in the cervical branches of the vagus. As for other peripheral nerves, there are direct relationships between vagal fibre diameter and conduction velocity, and between fibre diameter and stimulation threshold.

The model adopted a conduction velocity of large, myelinated, efferent, preganglionic A fibres with diameters of $\geq$ 10 μm, of up to 50 m/s; medium, myelinated, efferent preganglionic B fibres, with diameters ranging from 3 to 9 μm, of between 2.5 and 15 m/s, and small, non-myelinated fibres C, with diameters of $\leq$3 μm, of between 2 and 2.5 m/s. In the model the following combination of cathodal and anodal current is presumed to result in the predominant stimulation of medium-diameter fibres (e.g. B-fibres). The stimulation pulses proposed to be used were current, biphasic, charge balanced with a quasitrapezoidal cathodal phase with a magnitude of approximately 1 mA, composed of a square leading edge and a plateau of 300 μs in width as well as an exponentially decaying phase of 500 μs, followed by a wide rectangular anodal phase of low magnitude.

It was calculated that action potentials of thick A fibres, arriving at the anodes within approximately 100 μs and before the action potentials of medium fibres B, arriving at the anodes within approximately 300 μs, should be blocked by hyperpolarization, while the action potentials of medium B fibres would pass. Precisely, hyperpolarization on the anodes should last at least 100μs, until the action potentials of large A fibres arrive at the anodes and more than 300 μs until the action potentials of B fibres arrive at the anodes. Subsequently, hyperpolarization should be reduced exponentially to prevent anodal excitation.

A magnitude of the anodal phase had to be lower than the threshold of activation of thick A fibres in order to prevent excitation due to reversed polarity on the connected electrodes during the anodal phase.

However, the width of the anodal phase was dependent upon the charge injected in the cathodal phase: the greater the charge, the wider the anodal phase. It is expected that VNS energies will be similar to those used in VNS for the treatment of epilepsy.

The relative position of the group of three electrodes (triplet) closest to the superficial region innervating the aortic arch baroreceptors is to be determined by delivering stimuli to all 13 triplets within the cuff. Considering the dimensions of the nerve and the cuff, certain triplets are to remain out of contact with the regions of the nerve. Therefore, the region in contact with the corresponding triplet is to be selectively stimulated using the above-defined stimuli with an intensity of between 1 and 1.5 mA, plateau of 300 μs, exponentially decaying phase of 500 μs, wide rectangular anodal phase of low magnitude and a frequency of 20 Hz.

To validate changes in heart rate, the high resolution ECG is to be recorded. Simultaneously, arterial blood pressures are to be measured with the disposable pressure transducer system (model: DPT-6000) for invasive blood pressure monitoring (Smiths Medical Deutschland GmbH, Kirchseeon, company section: pvb-Critical Care). The ECG and blood pressure signals are to be delivered to a custom designed amplifier and a Daq/216B™, high-performance data acquisition PC-Card for notebook PCs, designed and manufactured by IOtech, Inc., 25971 Cannon Road Cleveland, OH 44146-1833, USA, connected to a Lenovo T420 Portable PC.

In the model of recording the activity over the area innervating aortic arch baroreceptors we assumed that the aortic arch would respond to various performed interventions with a changed firing rate. We suggest also in the model that the afferent activity of baroreceptors, innervated by the aortic nerve, traveling via a certain superficial region of the left vagus nerve to the brainstem, could be selectively recorded with the cuff implanted on the left vagus nerve.

The indicated triplet and three randomly chosen triplets, are to be then simultaneously connected to the custom designed ENG amplifying system with a gain of A=100,000, powered by batteries. The raw ENGs will be fed to an aforementioned data acquisition system and sampled at 20 kHz. Heart rate and blood pressure will be modulated by various separate interventions.

All the experimental activities are to be performed during carotid endarterectomy. The experiment is to be carried out in a patient under general anesthesia. An initial incision is to be over the anterior border of the left sterno-cleidomastoid muscle, centred between the mastoid process and the clavicle. The vagus nerve is usually identified in the carotid sheath in a posterior groove between the carotid artery and the jugular vein. Vessel loops are to be used to mobilise the vagus nerve so that at least 3 cm are available for attachment of the spiral cuff. At least 4 cm of the nerve is to be completely freed from its surrounding tissue. Depending on the size of the exposed nerve, the 39-electrode cuff is to be installed around the nerve trunk. Then the cuff will be carefully attached to the exposed left vagus nerve for the time period between the carotid endarterectomy procedures and removed after they are finished.

III. RESULTS

Figure 1 presents the results of modelling a multi-electrode spiral cuff, to be implanted on the human left vagus nerve.

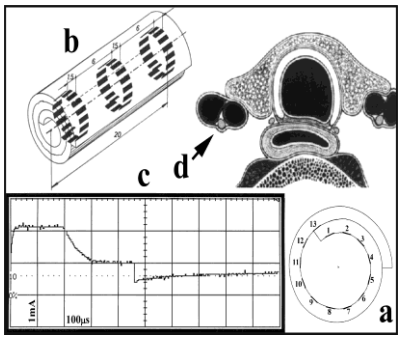


Figure 1: (a) geometric model of the 39-electrode spiral cuff, (b) 3D schematic diagram of the 39-electrode spiral cuff, c) stimulating current, biphasic, charge balanced pulse, and (d) site where the 39-electrode spiral cuff is to be implanted on the human left vagus nerve.

A cuff is made by bonding two 0.1 mm thick silicone sheets together [12, 13]. One sheet, stretched and fixed in that position, is covered by a layer of adhesive (MED-1511, NuSil, Carpinteria, California, USA). A second un-stretched sheet is placed on the adhesive and the composite is compressed to a thickness of 0.3 mm until the whole curing process is completed. When released, the composite curls into a spiral tube as the stretched sheet contracts to its natural length. The diameter of the cuff is related to the amount of stretch: the greater the stretch, the smaller the diameter.

Thirty-nine rectangular electrodes with a width of (0.6 mm x 1.5 mm) are made of 50 μm thick platinum ribbon (99.99 % purity) and connected to lead wires (AS 631, Cooner Wire, Chatsworth, California, USA) are mounted on a third silicone sheet with a thickness of 0.1 mm. They are arranged in three parallel groups each containing 13 electrodes at a distance of 0.5 mm. The distance between the spiral groups is 6 mm. Each electrode marked with the same number within each of the three parallel spiral groups has the same position. Accordingly, thirteen triplets in the same line in a longitudinal direction is formed. The central and two short-circuited outer electrodes of a particular triplet are then connected to the corresponding lead wires. The silicone sheet with the arranged electrodes is then bonded on the inner side of the mechanically opened cuff. The length of the cuff is optimized so that the surface of the nerve covered by the spiral cuff is as small as possible to prevent damage associated with a reduced blood supply and excessive mechanical trauma of the nerve. Therefore, the cuff with the inner diameter of 2.5 mm is trimmed to a length of 18 mm as shown in Figure 2b.

Finally, all the lead wires are connected to a male connector to be firmly fixed near the shoulder of a patient for the time between the experimental sessions. To connect the male connector to the outputs of the stimulator or inputs of the ENG amplifying system two cables have been developed (Figure 2a). For ENG recording the cable is shielded to minimize the pick-up of unwanted signals from the environment. At one end of the cables there is a female connector and at the other a connecting module permitting a particular GTE to be connected to the stimulator or ENG amplifying system individually or in combination with other GTEs. The short-circuited electrodes are to be connected to the stimulator as anodes while the corresponding central electrode is to be connected as a cathode. In the ENG recording mode the short-circuited electrodes are connected to one end of the ENG preamplifier, and the corresponding central electrode is connected to the other end.

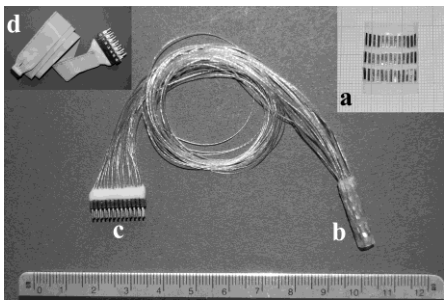


Figure 2: a) arrangement of electrodes on the silicone sheet, b) 39-electrode spiral cuff, c) subcutaneous connector, and d) connecting cable.

IV. CONCLUSION

In this modelling study we intended to elucidate whether a multi-electrode spiral cuff installed on the cervical vagus nerve can be used for controlling the heart rate, tachycardia and for suppression of atrial fibrillation in particular. According to our recent work, we have developed the model of using a multi-electrode spiral nerve cuff for selective stimulation of mainly fibres B within the left vagus nerve in a man to control the heart rate. Although most VNS research for controlling heart rate has been performed in animal studies, the consistency of the results obtained in preliminary human studies suggests that this technique may be used in humans. This potentially opens up the possibility to implant a device which, in contrast to a cardiac pacemaker, lowers the heart rate, and may be particularly beneficial for terminating specific paroxysmal arrhythmia or the non-pharmacological treatment of chronic heart failure. Furthermore, it seems logical to suggest that VNS delivered with different shapes and parameters, might produce different CNS effects that would in turn broaden the clinical indications. However, further research is needed to clarify the mechanisms of action of VNS and the potential clinical utility of this intriguing new somatic portal into the CNS.

ACKNOWLEDGMENT

This work was supported by the Ministry of Education, Science, Culture and Sport, Republic of Slovenia, research programme P3-0171. The authors thank Franci Vode, PhD, for providing the vagus nerve examines. The research was performed also as partial fulfilment of the requirements for the Doctor of Philosophy Degree (Polona Pečlin, MD).

REFERENCES

[1] Armour JA et al. eds. Neurocardiology. Oxford University Press 1994.

[2] Berthoud HR, Neuhuber WL. Functional and chemical anatomy of the afferent vagal system. Auton. Neurosci. 2000; 85: 1-17.

[3] Levy MN. Neural control of cardiac function. Baillieres Clin. Neurol. 1997; 6: 227-44.

[4] Mitchell GAG. Anatomy of the Autonomic Nervous System. E. and S. Livingstone Ltd., Edinburgh and London, 1953, chap, XII, pp. 160-189.

[5] Jones RL. Cell fiber ratios in the vagus nerve. The Journal of Comparative Neurology 1937; (67)3: 469-482.

[6] Chase MR. An experimental study of the vagus nerve. The Journal of Comparative Neurology 2004; (26)4: 421-428.

[7] Hoffman HH, Schnitzlein HN. The numbers of nerve fibers in the vagus nerve of man. The Anatomical Record 1961; 139(3): 429-435.

[8] Schnitzlein HN, Rowe LC, Hoffman HH. The myelinated component of the vagus nerves in man. The Anatomical Record 1958; 131(4): 649-667.

[9] Matheny RG, Shaar CJ. Vagus nerve stimulation as a method to temporarily slow or arrest the heart. Ann. Thorac. Surg. 1997; 63(6 Suppl): S28-9.

[10] Carlson MD, Geha AS, Hsu J, Martin PJ, Levy MN, Jacobs G, Waldo AL. Selective stimulation of parasympathetic nerve fibers to the human sinoatrial node. Circulation 1992; 85: 1311-7.

[11] Armour JA, Randall WC, Sinha S. Localized myocardial response to stimulation of small cardiac branches of the vagus. Am J Physiol 1975; 228: 141-8.

[12] Taylor J, Donaldson N, Winter J. Multiple-electrode nerve cuffs for low-velocity and velocity-selective neural recording. Med. Biol. Eng. Comput. 2004; 42: 634–643.

[13] Naples GG, Sweeney JD, Mortimer JT (inventors). Implantable cuff, method and manufacture, and method of installation. US Patent 4602 624, 29 July 1986.

[14] Yambe T, Nanka SS, Shiraishi Y, Tanaka A, Yoshizawa M, Abe K, Tabayashi K, Takeda H, Nitta S. Recording vagal nerve activity for the control of an artificial heart system. ASAIO J. 2003; 49(6): 698-700.

[15] Schachter SC. Vagus nerve stimulation: current status and clinical applications. Exp. Opin. Invest. Drugs 1997; 6(10): 1327-1335.

CTI

International Conference on

TELECOMMUNICATIONS & INFORMATION

Steering Committee

Chairs: Stjepan Golubić, Ericsson Nikola Tesla, Zagreb, Croatia
Branko Mikac, University of Zagreb, Croatia
Vlasta Hudek, Rezultanta, Zagreb, Croatia
Dragan Jevtić, University of Zagreb, Croatia
Davor Šarić, T-Croatian Telecom, Zagreb, Croatia

Members: Željka Car, University of Zagreb, Croatia
Tihana Galinac Grbac, Faculty of Engineering, Rijeka, Croatia
Miroslav Karasek, IREE, Czech Academy of Sciences Prague, Czech Republic
Marko Lacković, University of Zagreb, Croatia
Dina Šimunić, University of Zagreb, Croatia
Antonio Teixeira, Institute of Telecommunications, Aveiro, Portugal

INVITED PAPERS

New Challenges and Perspectives on Understanding and Managing the End User Experience: Presenting European Research Efforts on the Quality of Experience of Multimedia Systems

Lea Skorin-Kapov*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb
Croatia
lea.skorin-kapov@fer.hr

Sažetak – Dok se pojam kvalitete usluge (Quality of Service, QoS) primarno vezuje uz tehničke parameter koji opisuju performanse mreže i sustava, pojam iskustvene kvalitete (Quality of Experience, QoE) bavi se prvenstveno subjektivnom procjenom kvalitete od strane krajnjih korisnika, te predstavlja komplementarnu dimenziju i svojevrsnu nadopunu pojma QoS. Davatelji usluga i mrežni operatori prepoznali su važnost razumijevanja parametara koji utječu na iskustva krajnjih korisnika prilikom korištenja njihovih usluga. U ovom predavanju dati će se osvrt na istraživačke izazove vezane uz modeliranje i upravljanje iskustvenom kvalitetom korisnika, te će se predstaviti aktivnosti i ciljevi europske COST akcije QUALINET (European Network on Quality of Experience in Multimedia Systems and Services).

I. EXTENDED ABSTRACT

With the proliferation of multimedia services and advanced end user devices driving the transition from single medium communications based solely on voice to enriched and interactive multimedia, a key issue to address is meeting end-user perceived service quality expectations and demands. Studies dealing with the quality of media services date back to the early days of telephony speech and television system quality evaluation, to digital media services delivered via packet switched networks [1]. While today numerous ITU standards recommend various quality models and assessment methodologies [2], new and emerging service scenarios (e.g., 3D audiovisual services, Web and cloud-based services, collaborative services, gaming) are driving the need for new user-centered quality metrics and models.

The network-centric notion of Quality of Service (QoS) has been primarily associated with investigating and measuring technical network and system performance parameters (e.g., delay, jitter, loss). Various studies have further aimed at addressing the relationships between QoS parameters and end-user perceived quality [3][4][5]. Complementary to QoS, the notion of Quality of Experience (QoE) has recently gained great momentum in both academia and industry, aimed at providing a more holistic perspective on the factors impacting subjective quality perception [6][7]. Defined by the ITU as the "*overall acceptability of an application or service, as perceived subjectively by the end user*" [8], QoE extends the technology-oriented QoS concept by accounting for individual users' experiences derived from encounters with systems, impacted by expectations, prior experiences, context, etc. [9][10]. Consequently, besides depending on basic QoS networking parameters such as delay and bandwidth, a user's QoE may depend on parameters such as the usage context and intent of usage of the service, service content, and the user's socio-economic and psychological state.

In the context of increasing competition among stakeholders in the ICT domain, understanding the parameters that impact the quality of an application or service as subjectively perceived by the end-user has been recognized as a key issue providing a competitive advantage and reducing customer churn [11]. Looking to effectively manage and optimize QoE, knowledge regarding the root cause of unsatisfactory QoE levels or QoE degradations is necessary. This requires modeling the impact of different influence factors on QoE (including user-, application-, network-, and context- related factors). Resulting QoE models dictate the parameters to be monitored and measured, with the ultimate goal being effective QoE optimization strategies. The overall goal of QoE management may be related to optimizing end-user QoE (end-user perspective), while making efficient (current and future) use of network resources and maintaining a satisfied customer base (provider perspective) [12].

The widespread interest in the QoE field has brought together researchers under the umbrella of the European Cooperation in Science and Technology (COST) Action QUALINET (European Network on Quality of Experience in Multimedia Systems and Services). Following a general introduction to the domain of QoE, this talk will present some of the key research challenges

that are currently being addressed within QUALIENT, as related to the methodologies for quality evaluation and management of advanced multimedia services and communication systems. Croatian representatives in this community include the Faculty of Electrical Engineering and Computing and the Faculty of Graphic Arts, both from the University of Zagreb.

REFERENCES

[1] A. Raake et al. "IP-Based Mobile and Fixed Network Audiovisual Media Services: Current approaches to quality monitoring." *IEEE Signal Processing Magazine*, Vol. 28, no. 6, pp. 68-79, 2011.

[2] ITU-T Recommendation G.1011. Reference Guide to Quality of Experience Assessment Methodologies. Jun 2010.

[3] M. Fiedler, T. Hoßfeld, P. Tran-Gia, "A Generic Quantitative Relationship Between Quality of Experience and Quality of Service", *IEEE Network Magazine*, 24(2): pp. 36-41, Mar 2010.

[4] P. Reichl, B. Tuffin, R. Schatz, "Logarithmic laws in service quality perception: where microeconomics meets psychophysics and quality of experience", *Telecommunication Systems*, pp. 1-14, 2011.

[5] S. Junaid, M. Fiedler, and D. Collange. "Quality of Experience from user and network perspectives." *Annals of Telecommunications*, Vol. 65, no. 1-2 (2010): 47-57.

[6] R. Schatz, T. Hoßfeld, L. Janowski, S. Egger, "From Packets to People: Quality of Experience as a New Measurement Challenge", in *Data Traffic Monitoring and Analysis, From Measurement, Classification, and Anomaly Detection to Quality of Experience* LNCS, Volume 7754, pp. 219-263, 2013

[7] Z. Jingjing, and N. Ansari. "On assuring end-to-end QoE in next generation networks: Challenges and a possible solution", *IEEE Communications Magazine*, Vol. 49, No. 7, pp. 185-191, July 2011.

[8] ITU-T Recommendation P.10/G.100. Vocabulary for performance and quality of service. Amendment 2: New definitions for inclusion in Recommendation ITU-T P.10/G.100. Jul 2008.

[9] K. Kilkki, "Quality of experience in communications ecosystem." *Journal of Universal Computer Sciencem*, Vol. 14, no. 5, pp. 615-624, 2008.

[10] P. Le Callet, S. Möller, and A. Perkis, eds., "Qualinet White Paper on Definitions of Quality of Experience", European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, Version 1.1, June 3, 2012.

[11] R. Stankiewicz and A. Jajszczyk, "A survey of QoE assurance in converged networks," *Computer Networks*, vol. 55, no. 7, pp. 1459–1473, 2011.

[12] S. Barakovic, L. Skorin-Kapov, "Survey and Challenges of QoE Management Issues in Wireless Networks", *Journal of Computer Networks and Communications*, Vol. 2013, Article ID 165146, 28 pages

ICT i povećanje socijalne uključenosti osoba sa složenim komunikacijskim potrebama

Željka Car
University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb
Croatia
zeljka.car@fer.hr

Sažetak – Multidisciplinary research in the area of ICT-based systems for Alternative and Augmentative Communication (AAC) has been intensified recently, bringing together scientists and professionals from areas such as information and communication, computer science, education, rehabilitation, psychology and design. ICT-based systems for Alternative and Augmentative Communication provide support for communication, learning, and efficient execution of daily activities for persons with complex communication needs and may increase their social inclusion. In addition, these systems are aimed for family members, educators and therapist of persons with complex communication needs. The lecture presents state of the art in the area of development and implementation of ICT-based AAC systems, focusing on the results of Croatian researchers and challenges to be tackled within newly launched project financed by the European Union "IPA IIIc Science and Innovation Investment Fund".

Komunikacija je temeljni dio života. Svaka osoba ima potrebu razumjeti i komunicirati sa svojom okolinom, kao i da okolina prepozna potrebe, mogućnosti i osjećaje same osobe. Komunikacijske teškoće u ranoj dobi mogu imati ozbiljne posljedice na drugim razvojnim područjima, kao što su socijalne, emocionalne, i osobni razvoj.

Osobe sa složenim komunikacijskim potrebama iz različitih razloga nisu u mogućnosti komunicirati govorom te koriste potpomognutu komunikaciju (AAC - *Alternative and Augmentative Communication*), odnosno komunikacijski proces koji je ili potpomognut zamjenskim (*alternativnim*) sredstvima ili pojačan (*augmentacijom*) postojećih. Sustav potpomognute komunikacije pomaže osobama koje ne govore da se izraze, no on se može koristiti i kao sustav potpore koji osoba koristi za razumijevanje jezika ili komunikacije u cjelini. Uporaba potpomognute komunikacije može značajno unaprijediti kvalitetu života osoba s teškoćama, omogućiti im ravnopravnije sudjelovanje u društvu kao i mogućnost boljeg obrazovanja te veću razinu samostalnosti [1].

U grupu korisnika potpomognute komunikacije ubrajaju se osobe koje imaju poremećaj iz autističnog spektra, cerebralnu paralizu, intelektualne teškoće i razvojnu apraksiju govora ili stečene teškoće zbog različitih neuroloških bolesti ili oštećenja kao što su: progresivne neurološke bolesti, afazija, moždani udar, traumatska oštećenja mozga, demencija i sl [3].

Potpomognuta komunikacija (PK) uključuje sve metode komunikacije koje zamjenjuju ili potpomažu uobičajene metode komunikacije kada su one narušene ili kada ne odgovaraju potrebama pojedinca. Osnovu čine simboli koji predstavljaju izgovorene, grafičke ili pokretom izražene prezentacije ideja, osjećaja, objekata, radnji, ljudi, odnosa i događaja [1]. Simboli su organizirani u galerije, pri čemu odabir pojedine galerije treba biti prilagođen potrebama i mogućnostima i komunikacijskih partnera, pri čemu je nužno da oba komunikacijska partnera znaju značenje korištenih simbola .

Stupanj socijalne uključenosti osoba sa složenim komunikacijskim potrebama moguće je povećati razvojem ICT usluga koje će im omogućiti efikasniju komunikaciju, dostup informacijama i pružiti podršku pri edukaciji [2]. Razvoj novih tehnologije i pojava novih generacija pokretnih uređaja stvorile su široku lepezu mogućnosti za razvoj inovativnih aplikacija koje osobama sa složenim komunikacijskim potrebama pružaju podršku pri komunikaciji, učenju i učinkovitijem izvršavanju svakodnevnih aktivnosti, čime se nedvojbeno povećava stupanj njihove uključenosti u društveni život, a osobama s kojima žive, terapeutima i stručnjacima te edukatorima omogućuje se priprema informacijskih i edukacijskih sadržaja sukladno potrebama pojedinog korisnika.

Veliki nedostatak postojećih aplikacija i usluga zasnovanih na simbolima nedovoljno je korištenje mogućnosti Weba, čime im je ograničena komunikacijska komponenta. Postojeća rješenja zasnovana na simbolima mogu se svrstati u kategorije igara/vježbi (u kojima je potrebno odabirati simbole, određivati njihovo značenje i sl.) te pretvaranje poruka sastavljenih od simbola u glasovne ili tekstualne poruke. Spomenuta pretvorba simbola u tekstualni ili glasovni oblik ovisi o specifičnom jezičnom kontekstu jer u različitim zemljopisnim prostorima isti simboli označavaju različite stvari ili pojave. Također potrebno je za svaki jezik razvijati pretvorbu simbola što uvelike otežava brži razvoj usluga.

U Hrvatskoj su se zadnjih par godina intenzivirala istraživanja i razvoj u području potpomognute komunikacije. Tijekom projekta "ICT sustavi za osobe sa složenim komunikacijskim potrebama" uspostavljena je multidisciplinarna suradnje iz područja elektrotehnike, računarstva, logopedije, edukacije i rehabilitacije, psihologije i grafičke tehnologije na istraživanju i rješavanju složenih problema potpomognute komunikacije zasnovane na informacijskoj i komunikacijskoj tehnologiji. Tim projektom mobilizirali su se istraživačkih resursa i ojačali istraživački kapaciteti četiri sastavnice

Sveučilišta u Zagrebu: Fakulteta elektrotehnike i računarstva, Edukacijsko-rehabilitacijskog fakulteta, Grafičkog fakulteta i Filozofskog fakulteta - Odsjeka za psihologiju. Projekt je financirao Fond za razvoj Sveučilišta u Zagrebu u periodu od 1.1.2012. do 31.1.2013. godine, a u projekt su bile uključene i brojne roditeljske udruge kao predstavnici krajnjih korisnika, čiji je doprinos tijekom samih istraživanja, razvoja pilot aplikacija i njihove evaluacije bio neprocjenjiv Tijekom projekta stvorena je mreža kontakata i suradnje s predstavnicima različitih institucija (na primjer škola i centara inkluzivnog odgoja i obrazovanja, specijalnih bolnica, tvrtki za razvoj softvera), a medijska prezentacija projekta učinila ga je vidljivim široj javnosti, što je rezultiralo velikim brojem roditelja, stručnjaka, udruga i ustanove koje su osobno izrazile interes za provedena istraživanja i rezultate projekta. Tijekom projekta definiran je model tehnološke platforme [3] unutar koje su razvijene i implementirane prve pilot aplikacije namijenjene tablet uredjajima za komunikaciju, svladavanje osnovnih matematičkih vještina i poboljšanje jezičnih vještina [4].

U ožujku 2013. godine je pokrenut projekt „*ICT Competence Network for Innovative Services for Persons with Complex Communication Needs*" financiran u okviru fonda „IPA IIIc Science and Innovation Investment Fund"" koji okuplja četiri navedene sastavnice Sveučilišta u Zagrebu u svojstvu projektnih partnera, ali i značajni broj suradnika na projektu- centara inkluzivnog obrazovanja, roditeljskih i stručnih udruga i specijaliziranih poliklinika te tri mala i srednja poduzeća (SME), u okviru kojega će se dalje istraživati i rješavati složeni problemi potpomognute komunikacije podržane ICT uslugama

Zbog vrlo velike raznolikosti potreba i mogućnosti osoba sa složenim komunikacijskim potrebama, pri razvoju aplikacija za potpomognutu komunikaciju potrebno je implementirati mogućnosti dinamičke prilagodbe prikaza korisničkog sučelja i dinamičke prilagodbe samog sadržaja u okviru aplikacije. Sva rješenja za ovu nehomogenu korisničku skupinu trebaju imati mogućnost jednostavno se i učinkovito prilagoditi potrebama pojedinog korisnika. Sto su istraživački izazovi su potpuna implementacija poboljšane tehnološke platforme za razvoj i implementaciju ICT aplikacija zasnovanih na simbolima, korisnička sučelja s implementiranim mogućnostima prilagodbe potrebama korisnika i značajkama uređaja, prilagodba sadržaja potrebama korisnika, razvoj lokalizirane galerije simbola, mehanizmi za kategorizaciju simbola, mogućnosti kreiranja privatnih galerija simbola korinsika te implementacija elektroničkog javnog repozitorija simbolnih sadržaja za komunikaciju i edukaciju i pripade društvene mreže korisnika potpomognute komunikacije koje čine osobe sa složenim komunikacijskim potrebama, članovi njihovih obitelji, edukatori i terapeuti.

REFERENCES

[1] Katalog znanja o potpomognutoj komunikaciji. Projekt "ICT sustavi za osobe sa složenim komunikacijskim potrebama", 2012, dostupno na http://www.ict-aac.hr/index.php/katalog-znanja

[2] Ž. Car, I. Lovrek, "Predstavljanje projekta ICT sustavi za osobe sa složenim komunikacijskim potrebama". *Zbornik završne konferencije projekta*, Zagreb, 2013.

[3] I. Vučak, H. Belani, M. Vuković, "AAC Services Development: From Usability Requirements to the Reusable Components". *Proceedings of KES-AMSTA 2012: Agents and Multi-Agents Systems Technologies and Applications*, Springer, Dubrovnik, 2012, pp. 231-240.

[4] M Vuković, I. Blagajić, I. Vučak, "Aplikacije za potpomognutu komunikaciju razvijene u okviru projekta", *Zbornik završne konferencije projekta*, Zagreb, 2013.

[5] I. Šemanjski, T. Šaric, Ž. Janda-Hegediš, Ž. Car, M. Vuković, "e-Accessible service system: Calibrator and Communicator". *Proceedings of KES-AMSTA 2012: Agents and Multi-Agents Systems Technologies and Applications*, Springer, Dubrovnik, 2012, pp. 241-250.

[6] Ž. Car, M. Vuković, I. Vučak, J. Pibernik, J. Dolić, "A Platform Model for Symbol Based Communication Services". *Proceedings of the 11th International Conference on Telecommunications* ConTEL 2011, Graz 2011. 141-147.

[7] Z. Stančić, J. Frey Škrinjar, M. Ljubešić, Ž. Car, "Multidisciplinary Collaboration and ICT Sercices for People with Complex Communication Needs". *Zbornik radova 34. međunarodnog skupa* MIPRO 2011, Opatija, 2011, pp. 265-271.

PAPERS

SERVICE QUALITY

PSQM – Platform for service quality management

O. Jukić*, I. Heđi*
* Virovitica College, Republic of Croatia
{oliver.jukic|ivan.hedi}@vsmti.hr

Abstract - One of the most important criteria for evaluation of voice service quality is blocking probability – statistical probability that telephone connection cannot be established due to insufficient transmission resources in circuit switched network. Transmission path is established through number of network elements connected by transmission links with associated capacities. Hence, voice service is directly related to a number of functional and available speech and signalling channels on particular network parts and parts of transmission path. Further, in wireless networks, quality of voice service will depend on transmission capacities between access and core network and also on access network coverage. In this paper we have presented PSQM (platform for service quality management) aimed for on-line monitoring of relevant transmission KPIs (Key Performance Indicators) and its aggregation into the KQI (Key Quality Indicator). This monitoring should prevent operator from voice service degradation due to lack of transmission and signalling resources. Platform was tested in voice service domain. Data are collected from network elements using different techniques described briefly. Some other sources are also included in the KQI aggregation, giving integrated view on network and service status. Finally, implementation aspects of platform development as well as usage platform results are discussed.

I. INTRODUCTION

GSM architecture allows wireless access to voice service, including "classic" voice service as well as other services rely on "classic" voice service, such as IVR (Interactive Voice Recognition) services.

Although today's telecommunications are directed toward convergence of circuit switched ("classic" telephony) and packet switched (IP based) networks [1], monitoring of classic voice service based on circuit switching strategy is still of importance for telecom operators. In GSM network, service is implemented mostly in core network; however, it is necessary to ensure access to the service. Hence, it is also necessary to monitor access network allowing user access to the service.

During call setup in circuit-switched network communication channel is established "step-by-step". Transmission segments between nodes in network are reserved in first phase. In second phase, transmission segments will be grouped together, building "end-to-end" communication channel. Potential lack of transmission capacities will result with inability of building "end-to-end" communication channel. Those situations are not impossible and under specific circumstances are even normal. For instance, it is not unusual to be unable to make phone call in first trial during holiday such as Christmas or New Year's Day. Namely, networks are planned and optimized based on average information volume, not based on peak-values that are reached couple of times per year.

Hence one of the most important and relevant service quality indicators in circuit-switched networks is end-to-end blocking probability [5]. This is the probability that any call attempt will be unsuccessful due to a lack of network resources [9]. In optimal case, its value should be closed to zero.

In order to ensure voice service permanent availability it is necessary to ensure enough transmission capacities to be available during the call. These transmission capacities will participate in building a communication channel between calling and called subscriber. During call setup phase, it is also necessary to ensure enough signalling capacities (signalling links).

Finally, access to the core network should be realized through access network. Between access network and core network there are also transmission capacities participating in "end-to-end" communication channel.

In order to monitor and maintain "classic" voice service, it is necessary to collect real-time data from telecommunication network. In combination with network architecture and service implementation structure, these data will create real picture of availability of classic voice service. Data should be collected from different sources; the most important are [2]:

- fault management data
- performance management data
- end-to-end testing data
- customer complaints
- CDR (Call Details Records)
- other sources

PSQM platform primarily measures transmission and signalling capacities in circuit-switched network. Other performance indicators related to radio transmission or speech coding [6] are not in focus of PSQM.

In this paper we have described PSQM – platform for voice service quality management. Platform encapsulates several applications logically connected in one system.

Goal of the platform is systematical collection of real-time data from above mentioned sources, related to voice service quality as well as presentation of collected and aggregated data.

The platform is implemented in voice service monitoring process at one GSM operator in region and implementation results are satisfactory as it is described below.

Related works are briefly presented in second section, while in third section we have focused on content of collected data – KPIs (Key Performance Indicators).

Fourth section describes PSQM architecture with data sources mentioned as well as communication protocols and different data approach techniques. Parts of the platform are implemented in object-oriented manner; instantiation of new modules is possible without further development, only by using existing building blocks with specific configuration using XML language. Nodes in configuration XML file have specific names, and some of specific nodes are described in concrete examples.

In fifth section implementation aspects of platform components are described, as well as data storage and presentation key points.

Finally, usage results obtained from real telecommunication network are discussed together with hints for future work.

II. RELATED WORKS

According to an operator survey, with included results from 15 companies with a total of over 100 million customers, proactive activities in network and service management are desired, but these are not really integrated into existing OSS (Operational Support Systems) solutions. Also, one of the operators summarized his opinion as: "Services are not correctly managed well in any suite of applications and requires tremendous amount of work to maintain" [8].

There are a number of commercial platforms and tools aimed for service management. For instance, *HP Open View Service Quality Management Platform* provides service quality management solution, delivering full visibility on service levels across the entire network and IT infrastructure. *IBM Tivoli Quality Manager* combines service quality management and service level agreement.

However, according to our experience, many efforts are related to IT infrastructure and to IT service management. Management of classical voice service with systematical collection of relevant data is not in primary focus of current platforms. This gap can be avoided by implementation of proxy modules that will be able to collect relevant data and to "inject" it into service quality management platform.

III. PSQM MEASUREMENTS

In order to ensure adequate voice service quality level to end subscriber, it is necessary to ensure service access through access network coverage as well as service implementation in core network. Further, access network must be connected to the core network part using transmission capacities. All of those parts participate in temporary communication channel creation, during call setup phase.

PSQM platform is systematically collecting KPIs (Key Performance Indicators) from access and core network and it is aggregating collected data leading toward evaluation of KQIs (Key Quality Indicators).

Specific for access network is collection of data related to base station status which enables calculation of coverage percent for specific BSC (Base Station Controller) or RNC (Radio Network Controller) module.

Transmission capacity data are collected at transmission trunks between access and core network as well as within core network, between core network modules. Specific data are total number of trunks, number of currently occupied trunks, possible drilled-down by directions, number of idle trunks and number of non-operational (blocked) trunks. It allows evaluation of current load for any speech route, in absolute and relative (percentage) manner with possibility of generating alert to network operation personnel in the case of predefined threshold violation. Collected data are correlated with service structure data. For instance, different thresholds will be defined for speech routes having backup route and for speech routes without backup. Also, some specific routes can be related directly to subscribers with specific status or demanding SLA (Service Level Agreement).

In addition to transmission trunks, signalling links status is also measured. Signalling links are of great importance during communication channel establishment phase.

IV. PSQM PLATFORM ARCHITECTURE

A. PSQM basic architecture

Basic architecture of PSQM platform is shown on figure 1:

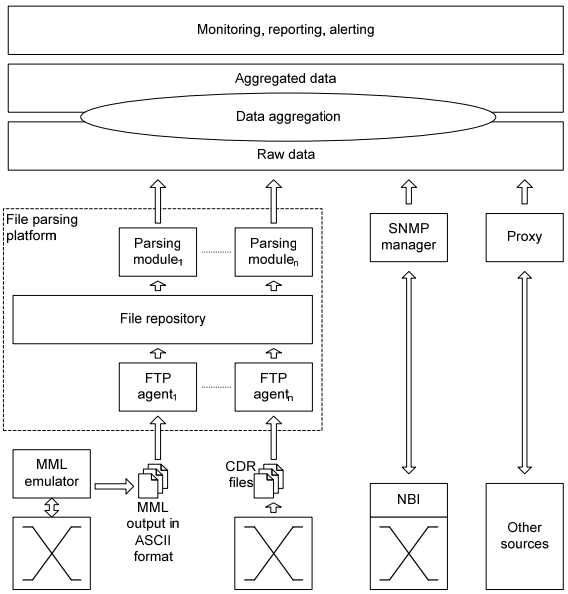


Figure 1 PSQM basic architecture

A number of FTP agents are aimed for systematical collection of network generated files into the file repository. Collected files are CDR files and files with MML (Man-Machine Language) output stored. Over file repository there is file parsing platform parsing collected files and extracting "raw" data from files. Then, raw data are stored in database.

After that, raw data are aggregated, together with data collected via northbound interfaces (NBI) and data from other sources. Service real-time monitoring, alerting in the case of threshold violation and report creation are all done based on aggregated data. A typical example is a number of idle speech trunks for some specific voice route – if that number achieves value less then N% (where N is configurable parameter for every voice route) it is necessary to generate alert to network operation personnel.

B. Data sources

For obtaining information about number of available, busy and non-operational speech trunks and signalling links MML-terminal emulation is used. Application **MML emulator** periodically executes MML commands on specific network elements. Response generated by network element is written into the ASCII file. Such files will be downloaded by FTP agents and "pushed" into the parsing platform. Files will be parsed respecting specific file format for every particular MML command type, which will produce raw data.

An example of MML command response for interrogating current state of voice trunks between MSC (Mobile Switching Center) and specific destination (in this case – BSC17, Base Station Controller 17) is shown below:

```
STATTRUNK:TGNO=BSC17,STATSUM=Y;     EXEC'D


INTERRUPTION TEXT JOB  1974

........

TRUNK STATUS SUM OVERVIEW

TGNO    STATUS          SUM
-------+---------------+-------
BSC17  IDLE             1825
BSC17  OUT                92
BSC17  INC                63
END JOB 1974
```

Parsing this response data stored in ASCII file allows collecting data about total number of trunks as well as number of busy trunks, grouped by direction (OUTgoing and INComing occupied).

Another data source are **CDR files**. CDR files are periodically generated by network elements and can be used as data sources for analysis within PSQM platform [7]. These files are not suitable for real-time monitoring, since typical generation interval is 15, 30 or 60 minutes. However, in this context, CDR files are of great importance for trend analysis and network capacity planning.

If **northbound interface** (NBI) exists for specific network element type, it is possible to obtain performance data via NBI. Content of information obtainable via NBI depends on NBI implementation and it is documented for every northbound interface type.

Data from any other source can be included in PSQM platform by **proxy module** implementation. In that way it is possible to include correlated fault management data ([3], [4]), end-to-end testing results, customer complaints and data from any other source.

Data collected from all above mentioned sources represent raw data. Actually, these data represent KPIs. After data aggregation, service quality indicators will be evaluated.

C. File parsing platform

In order to achieve easier and systematic approach to collecting CDR and MML-output files from the network, file parsing platform was developed. Base for data collection are N FTP agents acting as FTP clients collecting data from network elements. Number of FTP agent instances as well as configuration parameters for every particular FTP agent instance (e.g. time schedule, destination IP address, port number, source and destination folder, etc.) are defined in configuration file in XML format with specific node names predefined.

For every particular file type (binary or ASCII) there is parsing module processing incoming files and producing raw data. For every proprietary file format there is specific parsing module developed. However, there are common parsing modules that can be re-used for specific file types. It is only necessary to configure parsing module behavior in XML configuration file. Common parsing module example is parsing module for csv files parsing. For that module it is necessary to configure order number of every parameter within single line, parameter type and parameter name under which it will be stored in database (so-called variable). Another example is module for parsing files with fixed parameter's position. In that case, it is necessary to configure start and end position of every particular parameter, parameter type and parameter name.

The example of variables declaration within file parsing configuration file is shown in following XML segment:

```
<psqm:DECLARE_LPSTR name="@NE_NAME"/>
<psqm:DECLARE_LPSTR name="@TRUNK_NAME"/>
<psqm:DECLARE_LPSTR name="@DUMMY"/>
<psqm:DECLARE_I4 name="@TOTAL_TRUNKS"/>
<psqm:DECLARE_I4 name="@BUSY_TRUNKS"/>
<psqm:DECLARE_I4 name="@IDLE_TRUNKS"/>
<psqm:DECLARE_I4 name="@BLOCKED_TRUNKS"/>
```

Every XML node name related to PSQM platform starts with sequence: "psqm:". This ensures unambiguous parsing of XML configuration file and also allows any other configuration parameters, not related to PSQM platform, to be included in XML configuration.

There are 3 string variables declared as well as 4 integer variables, 4 octets length each. These variables can be used by all applications configured in the same XML configuration file.

After variables are declared, it is possible to configure one or more instances of FTP agent to download

appropriate files:

```
<psqm:FTP>
        <psqm:FTP_SERVER>
                ftpservername
        </psqm:FTP_SERVER>
        <psqm:FTP_USERNAME>
                someftpusername
        </psqm:FTP_USERNAME>
        <psqm:FTP_PASSWORD>
                someftppassword
        </psqm:FTP_PASSWORD>
        <psqm:REMOTE_FOLDER>
                D:\\
        </psqm:REMOTE_FOLDER>
        <psqm:LOCAL_FOLDER>
                C:\\TMN
        </psqm:LOCAL_FOLDER>
        <psqm:FILENAME_MASK>
                *.cdr
        </psqm:FILENAME_MASK>
        <psqm:TIME_INTERVAL>
                1
        </psqm:TIME_INTERVAL>
</psqm:FTP>
```

Node “psqm:FTP” will create one instance of FTP agent where all configuration parameters are defined in sub-nodes.

After files are stored in file repository, it is necessary to configure parsing module for each file type. The example is shown below:

```
<psqm:INPUT
        FileLocation="C:\\TMN"
        FileName="Test.cdr"
        ParsingModule="CSVParser.dll">
        <psqm:FIELDS>
                @DUMMY,
                @NE_NAME,
                @TRUNK_NAME,
                @DUMMY,
                @TOTAL_TRUNKS,
                @BUSY_TRUNKS,
                @IDLE_TRUNKS,
                @BLOCKED_TRUNKS
        </psqm:FIELDS>
        <psqm:SEPARATOR>,</psqm:SEPARATOR>
        <psqm:HEADER_LEN>2</psqm:HEADER_LEN>
</psqm:INPUT>
```

XML node “psqm:INPUT” defines location of file to be parsed, file name and executable file of parsing module (it defines parsing module type also). For this concrete example, parsing module is dedicated for parsing CSV files.

In XML node “psqm: FIELDS” there is a list of previously declared variables in which parsed data will be stored. Variables will be filled respecting parameter order in single CSV file line. Some of parsed values are stored in the same variable called @DUMMY – that is because those parsed values are irrelevant for PSQM, and variable @DUMMY is used just to respect order within single line.

After all CSV content is configured it is possible to specify CSV file separator as well as information about header length (number of lines from the beginning of the file to be ignored during processing).

After file is processed (parsed), all parsed values are stored within previously declared variables. At the end of file processing, it is necessary to store data from variables into database. Storing into database is configurable also. A segment of XML configuration file related to data storage (“data writing” in PSQM terminology, hence there is reference to „writer module“) is shown below:

```
<psqm:OUTPUT WriterModule="DBWritter.dll">
        <psqm:DB_CONNECTION_STRING>
                driver={SQL Server};
                server=127.0.0.1;
                database=Test;
                UID=someusername;
                PWD=somepassword
        </psqm:DB_CONNECTION_STRING>
        <psqm:TABLE_NAME>
                MSC_BSC_TRUNKS
        </psqm:TABLE_NAME>
        <psqm:COLUMN_NAMES>
                MSC_NAME,
                TRUNK_NAME,
                TOTAL_TRUNKS,
                BUSY_TRUNKS,
                IDLE_TRUNKS,
                BLOCKED_TRUNKS
        </psqm:COLUMN_NAMES>
        <psqm:COLUMN_VALUES>
                @NE_NAME,
                @TRUNK_NAME,
                @TOTAL_TRUNKS,
                @BUSY_TRUNKS,
                @IDLE_TRUNKS,
                @BLOCKED_TRUNKS
        </psqm:COLUMN_VALUES>
</psqm:OUTPUT>
```

XML node “psqm:OUTPUT” defines executable file for “writer“ module (it defines writer module type also). In this example database storage is shown. However, data can be written in some other ways (e.g. “written” to any other system by sending SNMP traps).

Node “psqm:DB_CONNECTION_STRING” defines connection string needed for ODBC connection to database. Table name is defined within node “psqm:TABLE_NAME” while column names are defined within node “psqm:COLUMN_NAMES”. Finally, node “psqm:COLUMN_VALUES” defines variable names to be written in specific table column, respecting type, number and order of variables/columns.

Using file parsing platform allows fast and easy introduction of new file types in PSQM platform. In the case of CSV files and “fixed position” files there is no need for developing – it is enough just to configure new instance of parsing module.

In the case of proprietary format which is not suitable for existing common parsing modules, existing development framework can be used in order to improve development of new parsing modules.

D. Data aggregation

After KPIs from different file types are collected, data aggregation is performed. Aggregation will generate new indicators, which will be closer to the customer point of view of service quality. Hence, we refer to these indicators as Key Quality Indicators or Service Quality Indicators.

For instance, for coverage area of single BSC or RNC module it is possible to collect data about status of every particular base station instance. Aggregation of collected raw data enables evaluation of coverage percentage for every BSC or RNC.

E. Presentation layer

Concrete outputs of PSMQ platform should be obtained key performance and quality indicators, presented through real-time monitoring tools, reports and visualized status of monitored service.

Aggregated data are base for report creation, but also raw data can be included in reports if needed – especially in drill-down reports. Not only reports are of importance – there is also possibility for real-time monitoring of service status, based on aggregated data. Figure 2 shows current access network coverage grouped by RNC modules – percentage of operational base stations.

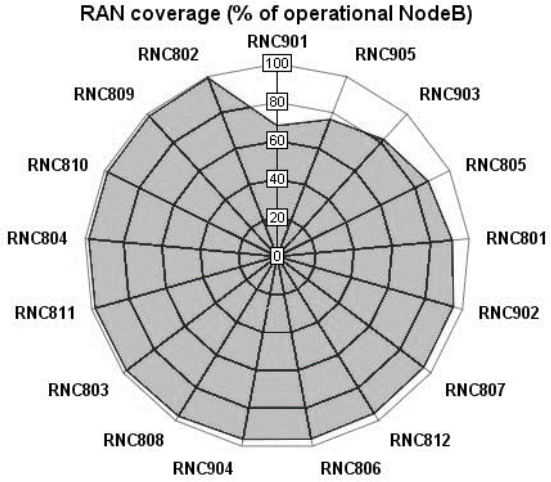


Figure 2 Access network coverage presentation

V. IMPLEMENTATION ASPECTS

PSQM platform was developed using C/C++ programming language and Microsoft Visual Studio compiler. Database engine was Microsoft SQL Server.

Presentation layer required development of number of smaller applications (utilities) for current data presentation. Service structure and its implementation on network is shown on main GUI (Graphical User Interface). In current phase, application dedicated main GUI is under development, but commercial tool HP Open View Network Node Manager is used for service presentation. This tool is not primarily aimed for service management, but it allows usage of very flexible and powerful API (Application Programming Interface). Usage of this API has enhanced performance of GUI for usage in service management domain.

VI. CONCLUSION

In this paper we have presented PSQM – platform for service quality management used in voice service domain. The main goal of this platform is real-time monitoring of basic KPIs related to resource availability in "classic" voice service for GSM/UMTS network. Measured KPIs are focused on blocking probability evaluation as one of the main parameters for service quality evaluation in circuit-switched networks.

Proposed data sources were CDR files, ASCII files with MML command responses and other sources, presented here and presented in previous works [2].

By platform implementation proof of concept was achieved – real-time monitoring of voice service quality was established, based primarily on transmission and signalling resources availability.

Future work should include development of PSQM dashboard, main GUI dedicated for service management. Data from any other source should be included, if their existence will increase confidence in data presented. Finally, platform should be verified in monitoring of other service types.

REFERENCES

[1] U. Koivukoski and V. Raisanen, "Managing mobile services – technologies and business practices", Wiley, 2005.

[2] O. Jukić and M. Kunštić, "Integrated view on telecommunication network status", Proceedings of the 34rd International Convention MIPRO 2011, pp. 141-145, MIPRO, Opatija, 2011.

[3] O. Jukić and I. Heđi, "Service monitoring and alarm correlations", 4th International congress on ultra moder telecommunications and control systems, pp. 330-334, Saint Petersburg, 2012.

[4] D.K.Udupa, "TMN – Telecommunications Management Network", McGraw-Hill Telecommunications, New York, 1999.

[5] A. Bažant et al., "Osnovne arhitekture mreža", Element, Zagreb, 2007.

[6] T. Halonen, J. Romero and J. Melero: "GSM, GPRS and Edge Performance – Evolution Towards 3G/UMTS", Wiley, 2003

[7] O. Jukić and I. Heđi, "The use of call detail records and data mart dimensioning for telecommunication companies", 20th Telecommunications Forum – Proceedings of papers, pp. 292-295, Beograd, 2012.

[8] S. Wallin and V. Leijon, "Telecom Network and Service Management: An operator Survery", MMNS 2009, pp. 15-26, 2009.

[9] ITU-T: "Recommendation E.721: Telephone network and ISDN quality of service, network management and traffic engineering – network grade of service parameters and target values for circuit-switched services in the evolving ISDN", ITU-T, Geneva, 1991.

A Possibility of Estimation of IPTV QoS Parameters by Using DSL Transceivers Physical Layer Parameters

N. Škaljo*, A. Begović* and N. Behlilović**
*BH Telecom, J.S.C. Sarajevo, Department Zenica, Masarykova 46, Zenica, Bosnia and Herzegovina
**Faculty of Electrical Engineering Sarajevo, Zmaja od Bosne bb, Kampus, Sarajevo, Bosnia and Herzegovina
namir.skaljo@bhtelecom.ba, alen.begovic@bhtelecom.ba, narcis.behlilovic@etf.unsa.ba

Abstract – this paper analyses a dependence of standard definition IPTV service continuity error on code violation block codes. A very reliable method of evaluating of several IPTV QoS parameters by observing the parameters of physical layer of DSL transceivers was introduced. Also, an ability to use the procedures, which have already implemented at common ATU-R and ATU-C ADSL2+ modems for estimation of IPTV QoS parameters, without appropriate performance monitoring system implemented was discussed.

Keywords - access network, code violation, continuity error, digital subscriber line, error correction block codes, error correction seconds, errored seconds, physical layer parameter, local loop, severely errored seconds, twisted pair

I. INTRODUCTION

The provisioning of triple play services over DSL (Digital Subscriber Line) technology has shown that the multimedia services, especially the IPTV (Internet Protocol TV) require strong conditions on the physical layer. They are more susceptible to problems than the traditional POTS (Plain Old Telephone Service) and flat-rate high speed data services. Problems of delivering of IPTV based services are various. These problems may be caused in every part of telecommunication network. It is often said that access networks with symmetrical copper pairs as their basic elements, cause most of the problems that occur on IPTV services.

In order to reliably determine if there are any problems with providing of IPTV services, and in which part of telecommunication network, appropriate mechanisms for monitoring of IPTV service performances need to be set. One method is evaluation of performances using QoE (Quality of Experience) parameters. QoE is subjective concept of service with components that are very difficult to measure in practice. In another words, it is the overall user "experience" of service that is provided to him. For instance, if user notices occasional short of prolonged "freezing" of picture, pixelisation of picture, delays in changing of TV stations etc.

For this subjective concept of service there is quite effective and objective way of evaluation of the entire telecommunication network that provide services, and which directly affects user's service "experience". In another words, there are QoS (Quality of Service) measurable network parameters at packet level.

In order to monitor performances of the entire telecommunication network, and all of its individual domains by using IPTV QoS parameters, monitoring of IPTV QoS parameters has to be implemented in the entire network [1], [2]. However, in practice it is possible that implemented IPTV monitoring of the entire telecommunication network does not exist due to several reasons [1]. One of those reasons is real possibility of existence of more providers, such as: VSP (Video Service Provider), RNP (Regional Network Provider), NAP (Network Access Provider), LP (Loop Provider) and NSP (Network Service Provider). Some possibilities of monitoring of IPTV service performances can be limited only for one of those domains and unavailable for others or for the entire telecommunication network. Also, it is possible to implement IPTV services even before implementing IPTV service performances monitoring.

However, regardless of existence or absence of IPTV service performances monitoring, it is possible, even with some already existing mechanisms implemented in ATU-C (ADSL Transceiver Unit-Central Office) and ATU-R (ADSL Transceiver Unit-Remote) modems indirectly, to monitor IPTV service performances from aspect of DSLAM (Digital Subscriber Line Access Multiplexer) access nodes, access networks and CPE (Customer Premises Equipment). This is related to physical layer parameters monitoring. These possibilities were also specified in appropriate recommendations for production of ATU-R and ATU-C ADSL (Asymmetric Digital Subscriber Line) modems [3], [4] when only data transfer service were provided through them, but not IPTV services.

The goal of this paper is to show how some IPTV QoS parameters from aspect of access networks can be estimated by using physical layer parameters for DSL transceivers, even when implemented IPTV service performance monitoring does not exist.

TABLE I. RECOMMENDED MINIMUM TRANSPORT LAYER PARAMETERS FOR SATISFACTORY QOE FOR MPEG-2 ENCODED SDTV SERVICES [5]

Transport stream bit rate (Mb/s)	Delay	Jitter	Maximum duration of a single error	Corresponding loss period in IP packets	Loss distance	Corresponding average IP video stream packet loss rate
3,0	<200 ms	<50 ms	≤16 ms	6 IP packets	1 error event per hour	$\leq 5{,}85 \times 10^{-6}$
3,75	<200 ms	<50 ms	≤16 ms	7 IP packets	1 error event per hour	$\leq 5{,}46 \times 10^{-6}$
5,0	<200 ms	<50 ms	≤16 ms	9 IP packets	1 error event per hour	$\leq 5{,}26 \times 10^{-6}$

II. IPTV QOS AND PHYSICAL LAYER PARAMETERS

Generally, the most important IPTV QoS parameters are [5]:

- transport stream bit rate,
- packet loss,
- delay,
- jitter.

Except these four QoS parameters, there are some other QoS parameters, such as: Error indicator, continuity error (it shows number of lost media packets), PCR (Program Clock Reference) jitter, PSI (Program Specific Information) data error, IGMP latency. There are two kinds of jitter [6]: IP packet jitter and PCR jitter. IP packet jitter can influence PCR jitter occurrence, while PCR jitter can cause some other occurrences that are not related to transport network.

In terms of what we want to show in this paper regarding usability of data on number of errors on lines in certain period of time, it is important to focus on parameter packet loss. There are two kinds of packet loss [6]: IP packet loss and video packet loss (MPEG-2 or MPEG-4 packet loss).

Regarding these four basic QoS parameters, copper pair as a part of access network, can certainly impact on attainable bit rate and packet losses to the extent that it affects IPTV service [7]. However, if the packet loss due to copper pair was extremely high, then measurements of delay and jitter could ascertain that the copper pair affected both jitter and delay. But, that would be mainly caused by packet losses. "Real" jitter and delay can be present even if there are no packet losses. Accordingly, hereafter, the influence of copper pair on the IPTV service QoS parameters will be observed only through packet losses.

For this study, it is important to define what appropriate recommendations say about allowed media packet losses. Regarding video packet loss, IP packet loss and other QoS parameters for SDTV (Standard Definition TV) defines [5] (Table I).

Regarding recommended IPTV QoS parameters and relationship between IP packets and video packets, it is important to emphasize that one IP packet contains seven uniform MPEG-2 video packets [6].

Also, it is important to adduce some physical layer parameters defined by [8]. This recommendation defines considerable number of physical layer parameters for DSL transceivers. In order to accomplish the goal of this paper, it is important to point out these parameters:

- FEC (Forward Error Correction) codes,
- CV (Code Violation) codes,
- ES (Errored seconds),
- ECS (Error correction seconds),
- SES (Severely errored seconds).

According to definitions of these parameters [8], FEC block code counters show count of all block codes that have detected and corrected all errors, in certain period. Counters for ECS show count of one-second intervals in which these FEC codes occurred. Counters for CV show count of all block codes that have detected but not corrected those detected errors in certain period. Counters for ES show number of one-second intervals in which CV codes occurred. Counters for SES show number of seconds in which 18 or more CRC anomalies occurred.

III. RESULTS AND DISCUSSION

In accordance with previous descriptions, and in order to show how can we estimate number of lost video packets based on the number of errors, FEC block code counters are not important indeed, but just CV block code counters. Thus, estimation of lost video packets will be based only on CV block code counters data. As can be seen from the description, only these codes can cause packet losses. However, in order to obtain an overall picture of observed line, it is well to know ES, ECS and SES parameters too.

Table II shows testing results made on a real system in practice.

If we analyse the testing results from Table II, we can see the dependence of number of video packet loss from CV block codes number. We have been testing on an ADSL2+ line several times in few months period. We have been doing it under very different circumstances regarding copper pair primary parameters, and conditions on the line regarding bit rate required for transfer of requested services.

TABLE II. VIDEO PACKETS LOSS DEPENDENCE OF CV CODES

Numeral of testing		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Line rate (kb/s)	DOWN:	12.283	12.159	12.283	12.287	7.049	7.049	7.176	5.656	5.788	4.988
	UP:	893	869	890	894	916	937	850	855	855	850
Maximum attainable rate (kb/s)	DOWN:	13.016	12.248	12.772	12.656	13.088	12.624	7.188	5.668	5.804	5.000
	UP:	1.170	1.170	1.170	1.170	1.170	1.170	1.170	1.170	1.170	1.170
Output power (dBm)	DOWN:	22,0	22,0	22,0	22,0	21,0	21,0	21,0	21,0	21,0	21,0
	UP:	12,6	12,6	12,6	12,6	12,6	12,6	12,6	12,6	12,6	12,6
Signal attenuation (dB)	DOWN:	24,0	24,0	23,0	23,0	20,0	20,0	21,0	21,0	21,0	21,0
	UP:	9,0	16,0	16,0	16,0	16,0	16,0	17,0	17,0	17,0	17,0
SNR margin (dB)	DOWN:	7,5	6,0	6,0	6,5	18,5	18,0	6,5	6,5	10,5	5,5
	UP:	8,0	7,5	7,5	7,0	7,5	7,5	7,0	7,0	7,5	7,0
FEC	DOWN:	11	97	582	10	0	0	70.948	68.011	69.074	71.531
CV	DOWN:	**3**	**2**	**3**	**1**	**0**	**0**	**26**	**28**	**12.963**	**4**
Count of initialisation attempts		0	0	0	0	0	0	0	0	0	0
UAS (s)	DOWN:	0	0	0	0	0	0	0	0	0	0
SES (s)	DOWN:	0	0	0	0	0	0	0	0	125	3
ES (s)	DOWN:	3	2	3	1	0	0	23	23	128	0
LOS	DOWN:	0	0	0	0	0	0	0	0	0	0
ECS (s)	DOWN:	6	79	212	7	0	0	590	587	589	602
Testing duration (minutes)		10	20	20	10	10	40	20	20.	20	20
Received IP packets		176.618	332.873	323.986	167.797	170.295	694.017	298.671	296.854	267.400	300.182
Count of continuity error		**0**	**0**	**2**	**1**	**0**	**6**	**48**	**35**	**22.606**	**5**

First four columns represent testing results when the observed line was carrying triple-play service whose video component consists of two simultaneous SDTV IPTV streams. In the last six columns we presented testing results when the line had been carrying triple-play service with one SDTV IPTV stream.

Regarding quality condition of the line primary parameters, it is also important to emphasize that the first six columns refer to testing results when there was not any malfunction on the line. In these conditions, there was a wideband noise on the copper pair whose maximum amount on the side of main distributor was -34.4 dBm.

In the last four columns we showed testing results when very significant malfunction on the customer end of the line existed.

It is important to emphasize that CV block codes refer to all packets on the line regardless of the type of service (IPTV, VoIP) that they contained in their informational part. In another words, considering that this line transfers three different services simultaneously, if CV codes occur, that does not necessarily mean that only video packet loss will happen. Of course, VoIP or data loss can occur too.

From presented testing results, we can notice that cases in which there are several CV codes can occur, but with no video packet losses. In those cases, such CV codes were probably carrying VoIP or data. Then, there are cases when there are not any CV codes, but video packet losses occur. Of course, this brings us to the fact we cannot forget- packet loss can occur in other parts of transport network, and not just in access network.

However, for cases when there is large number of CV codes, and especially when there is very significant number of CV block codes, we can notice that one CV code on average causes loss of two packets maximum. By observing and analysing for this kind of triple-play line, from the viewpoint of number of CV block codes in some period, we can estimate if a line satisfies necessary conditions regarding loss packet number in accordance with Table I., of course depending on the referent level regarding packet losses.

It is important to emphasize that facts that we pointed out, we showed for this kind of triple-play line does not match to other triple-play lines that have different distribution of band for certain service. However, similar

experiment can be done for other broadband lines that contain IPTV service.

IV. CONCLUSION

As it can be seen from provided results of testing on the real system, there is quite feasible possibility of indirectly estimation of IPTV QoS parameters in terms of access network even when IPTV performance monitoring of the entire telecommunication network is not implemented. For this purpose, existing mechanisms implemented in DSLAM devices and ATU-R modems, which existed before IPTV service implementation, are used. Also, which is very important, this possibility of QoS parameters assessment is performed on probably the most sensitive segment of the entire telecommunication network. Also, due to often need to solve malfunctions and disturbances, daily monitoring is performed on the user lines. The results of this monitoring are stored for a long period of time. These results can be used anytime. Thus, this method becomes even more significant.

However, we can see that this indirect way of QoS parameters evaluation cannon tell us what is QoE of delivered service. Of course, it is possible that in certain circumstances this indirect measurement of QoS on segment of access network shows quite satisfactory results regarding attainable bit rate and packet losses on segment of access network, but the customer still has unsatisfactory results regarding QoE. This is because other segments of telecommunication network can sometimes affect the overall QoS of provided IPTV service.

Also, it is important to know what the relations between continuity errors are and CV codes for lines with different total bit rate and different distribution of that bit rate for certain services. In this regard, it is important to do necessary testing of those lines too.

There are also several interested topic for further work, e.g. a comparison of measured IPTV QoS with DSL parameters on different broadband lines; measurement of IPTV QoE in parallel with DSL parameters and its correlation, etc.

REFERENCES

[1] Broadband Forum Technical Report 160: IPTV Performance Monitoring, November 2010,

[2] J. Zhang, Y. Wang, B. Rong, "QoS/QoE techniques for IPTV transmissions", in Proceedings of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting BMSB, Bilbao, Spain, June 2009, pp. 1-6,

[3] ITU-T Recommendation G.992.5: Asymmetric digital subscriber line transceivers - Extended bandwidth ADSL (ADSL2+), May 2003,

[4] ITU-T Recommendation G.992.3: Asymmetric digital subscriber line transceivers 2 (ADSL2), July 2002.

[5] DSL Forum Technical Report 126: Triple-play service Quality of Experience (QoE) requirements, December 2006.

[6] "Triple-Play service deployment - A comprehensive guide to test, measurement and service assurance", JDS Uniphase Corporation, Milpitas, CA, USA, October 2007.

[7] "How network conditions impact IPTV QoE", Case Study, Agilent Technologies, Englewood, CO, USA, October 2008,

[8] ITU-T Recommendation G.997.1: Physical layer management for digital subscriber line (DSL) Transceivers, June 2012,

An Impact of Access Network Physical Impairments on IPTV QoE in DSL Enviroment

N. Goran*, M. Hadžialić ** and A. Begović*
* BH Telecom JSC, Sarajevo, Bosnia and Herzegovina
** University of Sarajevo, Faculty of Electrical Engineering, Bosnia and Herzegovina
nermin.goran@bhtelecom.ba; mesud.hadzialic@etf.unsa.ba; alen.begovic@bhtelecom.ba

***Abstract* - this paper analyses the impact of access network of physical impairments in DSL environment caused by primary parameters imperfections. It also analyses the correlation of BER (Bit Error Rate) and error control of DSL system with values of MDI (Media Delivery Index) which gives an indication of expected IPTV (Internet Protocol Television) QoE (Quality of Experience). Also this paper gives a review of different network impairments, proper values of QoS (Quality of Service) parameters, like MLR (Media Loss Rate) factor, and at the same time, their impact on quality of IPTV. At the end, practical examples are given of how different types of network physical impairments cause problems in delivering IPTV service.**

***Keywords* — access network; bit error rate; digital subscriber line; media delivery index; media loss rate; quality of experience; quality of service;**

I. INTRODUCTION

Today, in the world, there is the biggest growth of DSL (Digital Subscriber Line) lines in comparison to other broadband technologies [1]. With using DSL vectoring, operators can offer huge bit rates [2], while at the same time penetration of pure FTTH (Fibre to the Home) lines is far from anticipated. Operators of fixed network try to find optimum between offering of bit rates and OPEX (Operating Expenditure) and CAPEX (Capital Expenditure) for future FTTH network. In many countries, without intervention of the government, deployment of FTTH network will be very slow. Problems which appear in twisted pair environment are identified as problems of cable primary parameters [3]. Also after shortening lines, as in some FTTX (Fibre to the X) solutions, the same problems appear. At the same time, impact of network physical impairments of DSL lines, caused by primary parameters imperfections, is in correlation with BER and error control of system [4].For triple play services, which are delivered by DSL lines, consideration of twisted pair problems is very important. But these network impairments cause problems in higher layers [5]. Error control of DSL systems in this case is very important and we will give example of how problems appear that error control cannot solve [6]. At the same time QoS network parameters is defined by [7] influence QoE of video stream. In [8] are defined demands for proper QoE in DSL environment. One example of relation between QoE and QoS is defined in [9]. This paper defines relation between quality of DSL lines and error control system versus quality of experience at delivering IPTV service. Also the main aim of this work is demonstration of influence of bit errors on QoS or QoE that occur as a result of different cause in access network. This is explained and analysed in chapter III and IV. For examples we used delivering of SDTV (Standard Definition Television) and HDTV (High Definition Television) MPEG-2 (Moving Picture Experts Group) video over common protocols like TS/RTP/UDP/IP/ATM over ADSL2+.

II. QOS PARAMETERS OF MPEG-2 TS VIDEO IN CORRELATION WITH QOE

QoS of transport network defines QoE i.e. IPTV QoE is a function of different QoS parameters. Also, choice of compression technique and choice of transport methods influence QoE. But if we ignore those choices we can appoint that QoE depends on QoS [9], [10]. In this case different network parameters have different weighting factors that affect in some manner a level of QoE. It can be shown as a function of following parameters:

$$\mathrm{QoS} = f_1\left(MDI,\ \text{loss, bandwidth, IGMP latency}\right) \quad (1)$$

Also we can write dependence of QoE as:

$$\mathrm{QoE} = f_2\left(\mathrm{QoS}\right) \quad (2)$$

MDI in (1) indicates media delivery index that consists of two factors. First of them is delay or variation delay known as jitter. For appropriate video content delivery, constant bit rate has to be assured for STB (Set Top Box) i.e. constant coding rate. Every modification of rate affects user perception end to end route between headend and user's STB. Packets with constant rate have null value of variation of delay or jitter. Jitter or variation delay can be shown as equation:

$$VD = \left|1 - \frac{R_{\mathrm{rcv}}}{R}\right| \quad (3)$$

where R denotes projected rate of packets and R_{rcv} is current rate. Choice of buffer proper size can provide constant rate to decoder settled in STB. Higher level STB can solve and tolerate 50 ms variation of packet rate. Also that indicates the situation where R_{rcv} is 66 % value of R. Possibility of solving jitter problem depends on quality of STB i.e. solving correction of variation rate can be solved with proper projecting buffer in STB. With the above mentioned, jitter causes less of total values that influence

complete QoE at customer side. Besides the jitter, MDI is defined over media loss rate or MLR. MLR indicates losing of continuity of packets or out of order packets. MDI with 4:0.001 has 4 ms jitter and MLR presented by 0.001 loss packets in second [11]. Maximum loss rate is 5 IP packets in a second in a half hour observation for SDTV and VoD, while HDTV is 5 IP packets in second in four hour observation. Maximum allowed MLR is 0.019 for SDTV, but for zapping time that must be 0. Packet loss in network layer occurs in case of constant error that affects out of sequence packets in video stream, lack of packets and duplicate of packets. Appearance of packet loss can happen in all network layers. Particularly if there are no physical disturbances on DSL network, problem with packet loss can be produced by errors in distribution network like congestion or coding problem at video headend side. Packet loss is not allowed over 0.1 % [12]. Congestion on network level can cause variation and PCR jitter that produces pixelization and blurring of video stream. There is no variation of MPEG-2 packets if this parameter is less than or equal 5 ms. Proper reproduction of video stream requires enough broadband and rates have to satisfy values between 2.5 and 3.5 Mbps. Improper broadband causes significant jitter in packets. IGMP is used for multicast. IGMP message, sent by network, is received at customer side and proper program is selected by STB. Time from sending IGMP message to receiving first video stream is called IGMP delay. If there is one headend, every program has the same IGMP delay. If it is less than or equal 200 ms, there is no influence on QoE. Every network element processes and adds packet queue. Also every element generates jitter or cause degradation in packet queue. In a real system, such as the one SMP (Significant Market Power) operators have, there are many points that can cause problems in packet delivery. QoS for these networks can be shown as summation QoS of every element that is included in distribution of video steam. Problems may be introduced by headend at the reception of signal. Also every point in IP/MPLS network can cause degradation of QoS parameter. From the above mentioned, total QoS can be written as:

$$\mathrm{QoS} = \sum_i \alpha_i \mathrm{QoS}_i \tag{4}$$

where or of QoS_i is QoS and α_i is weighting factor of i^{th} element, respectively. Weighting factor indicates the relative importance of impact of i^{th} element as compared to other observing elements. But very important part of distribution of video stream is access network as one of problematic part of all network segments. In this case, there is customer and QoE noticed by him at the end of network. Every problem in transportation of video stream and some degradation of QoS parameters can be probably noticed here. Also there are many disturbances in DSL environment that cause degradation in this part while physical level of QoS influences higher level of protocols.

III. Impact of Network Impairments on Content Quality of IPTV as Correlation of QoPh and QoE

Presence of different source of noise on DSL lines can cause appearance of errors in bit transfer. Some of those errors can be controlled but some cannot. Stationary noise that appears from other active DSL lines in cable binder and wideband noise can produce little errors. At the same time, noise that is generated due to the on-off switching of a DSL by neighbouring users produces stepwise changing crosstalk. But if there is proper SNR margin we can neglect that noise because every network converges after some time. With using trellis or RS (Reed-Solomon) codes we have additional improvement. Impulse noise that appears in noise bursts is big source of bit errors especially if it is REIN (Repetitive Electrical Impulse Noise). Also impulse noise in bad balanced pairs in DSL environment can introduce enormous bit errors. When we have bad balancing and long twisted pair it is difficult to provide 6 dB SNR (Signal-to-Noise Ratio) margin. High resistance i.e. improper primary parameters of cable impact the quality of physical links. Bad quality of home wire installation couples impulse noise very well. Although some errors cannot be avoided with using of correction codes, trellis or RS coding help with solving of errors caused by impulse noise. While correction coding introduces delay, depth of interleaving increases delay. Maximum delay with interleaving is 16 ms. Fast method causes delay of 2 ms but this method cannot fight with impulse noise. For video streaming, using of interleaving is inevitable. If using interleaving, delay in DSL access network can be shown as:

$$D = D_{\mathrm{int}} + D_{\mathrm{proc}} \tag{5}$$

where D_{int} and D_{proc} are delay caused by interleaving and processing packets in ATU-C (ADSL Termination Unit – Central) and ATU-R (ADSL Termination Unit – Remote) respectively. As we mentioned in DSL environment, there are stationary and non-stationary sources of noise. Disturbances appear because those noises can be noticed as network physical impairments. If we use error control of DSL system we can correlate QoPh (Quality of Physical link) of DSL lines with QoS and IPTV QoE. Quality of DSL lines can be defined by number of disturbances that happen on pairs. This can be written as:

$$\mathrm{QoPh} = f_3\begin{pmatrix}\text{FEC, CV, ES, SES, SNR margin,}\\ \text{HEC, IGMP latency, attenuation}\end{pmatrix} \tag{6}$$

where FEC, CV, ES, SES and HEC are Forward Error Correction, Code Violation, Error Seconds, Several Error Seconds and Header Error Correction respectively. Also, other valid equation is:

$$\mathrm{QoS} = f_4\left(\mathrm{QoPh}\right) \tag{7}$$

We can write, by combining equations (2) and (7)

$$\mathrm{QoE} = f_2(\mathrm{QoS}) = (f_2 \circ \quad \text{'h}) = f_5(\mathrm{QoPh}) \quad (8)$$

Indication of FEC or FEC ES bits is consequence of correction errors. Presences of enormous number of errors show presence stationary noise or noise from systems that converge in stationery noise. Also impulse noise can cause appearance of errors. A huge number of FEC indications do not mean that QoS has low weighted factors or low QoE at customer side, because FEC has good result in correction stationary noise. Indication of code violation means appearance of some errors in CRC word caused by degradation of SNR margin. Also disruption of bit continuity can generate QoS parameter degradation and low IPTV QoE. Bigger number of CV errors (10 or more) generates one error second (ES). Depending on the place in frame where the packet were lost, deviation of video quality will be less or more noticeable. The worst case is when I and B frame are lost. Every ES indication will cause a problem in continuity of video stream. Therefore, there will be blurring of IPTV video stream after ES indication. This will be especially noticeable when ES is repeated. Several error seconds happen after several consecutive ES indications. In every case weighted factor of QoS parameters decrease and IPTV QoE becomes unsatisfactory. Stability of SNR margin has the big role in DSL environment. SNR margin defines size of received noise which can be added on observed DSL line but still with existing service working. Size of margin is only parameter but stability of SNR margin is a key factor that defines QoE. Although there is a big size of SNR margin that does not mean that video streaming will be reliable. Every change of SNR margin indicates one or more error CV bits. As [10] defines only one ES can occur in 60 min of video at SDTV. For HDTV we have more rigorous criteria. One ES can occur in 4 hours of HDTV video. But from customer view it is not presented strictly. He does not lose interest but he will be more unsatisfied. This means he will assess it with less MOS. In particular DSL line, CV indication can be noticed in five or less minutes. Attenuation in comparison with length of cable is physical problem regardless of the quality of primary parameters.

TABLE I. SNR MARGIN VS MAX. RATE

Length CO - STB	SNR margin DS	Attenuation DS	Max att. Rate DS	Dedicated rate DS
300 m	35	1	23476	5886
1000 m	20	12	17654	9015
1500 m	17	18	15640	9015
2000 m	10	25	12758	9015

Every signal attenuates with propagation along the cable. Physical restrictions and primary parameters of cable sections can cause huge fall of SNR margin. In Table I, there are values of SNR margin and attenuation of cable type TK 59-45 GM 100×2×0.4 [13]. Maximum attainable rate is defined by SNR margin and attenuation. As mentioned above, although there is enough size of margin, that does not provide non error state. Maximum attainable rate is a rate when SNR margin would have value about 6 dB and BER = 10^{-7}. Especially in non-stationary noise environment, where there is no stability, it is negligible. HEC indication shows error in ATM cells and depends on ES and SES.

IV. EXAMPLES

Above mentioned observation is checked by some examples. Quality of physical network impacts content quality of IPTV. Error control system defines quality of DSL lines and impairments in network. Imperfections in primary parameters can be noticed over error control system. Examples show correlation between impairments of physical DSL link and QoS parameters and QoE of IPTV. We observed factors from (1) and (6) and showed how they influence QoE. We used one pair in DSL environment for all measurements. First example is measurement of error on ADSL2+ line that was disturbed by non-stationary noise, depicted in table II. Line used fast method and transported SDTV MPEG-2 video. We used 10 period of time observing and measuring. It can be noticed that weighted factors of QoS parameters are satisfactory and do not show some deviations. But error control system of DSL shows bit indications like CV, ES and SES. Every indication of ES caused short time freezing of video stream. There were 12 indications in 10 minutes. Also appearance of SES indication caused the same problem as ES on IPTV QoE. At the same time we could not notice any deviations in QoS parameters. Probably errors that appeared in packets caused little loss but losses were in I and B frame. Continuity of packet video stream in MPEG-2 TS, as well as PMT and PAT errors were used for indication of QoE.

TABLE II. MPEG-2 SDTV SERVICE ADSL2+ FAST METHOD

QoPh	QoS	QoE
Local (Dn) CV: 207	Packet Loss = 0.1 %	PCR jitter = 2 ms
Local (Dn) ES: 12	Packet Jitter = 1 – 2 ms	PAT = 2
Local (Dn) SES: 6	Packet Delay = 28 ms	PMT = 2
Local (Dn) UAV: 0	Bandwidth = 7156 Mbps	PID timeouts = 1
Local (Dn) HEC: 818	IGMP latency = 28 ms	Continuity errors = 26

Second case was line with interleaving method that transported SDTV MPEG-2 TS (Table III). FEC indication showed corrected errors but still every ES caused freezing of video stream. In this case, continuity errors, PMT and PAT errors followed similar pattern like in previous case. That strongly depended on ES and SES errors. Worse situation than previous was due to appearance of impulse noise with unpredictable form although we used interleaved method.

Next case scenario relates to less bandwidth in access network, depicted in table IV. This implicates little current rate and maximum attainable rate. This situation represents huge higher layer problems although there is very big size of SNR margin. As we see in table IV, there are no errors in physical layer, produced by less bandwidth. At the same time, errors of higher layers implicate users' disappointment because weighted factors of QoS parameters have low values. Therefore IPTV QoE, in this case, is very low.

TABLE III. MPEG-2 SDTV ADSL2+ INTERLEAVED METHOD

QoPh	QoS	QoE
Local (Dn) FEC: 380	Packet Loss = 0.86 %	PCR jitter = 2 ms
Local (Dn) FEC ES: 11	Packet Jitter = 1 – 48 ms	PAT = 6
Local (Dn) CV= 319	Packet Delay = 105 ms	PMT = 6
Local (Dn) ES= 10	Bandwidth = 7166 Mbps	PID timeouts = 0
Local (Dn) SES: 7	IGMP latency = 3 ms	Continuity errors = 800
SNR Margin: 30 dB		
Local(Dn) HEC: 9523		

TABLE IV. MPEG-2 HDTV ADSL2+ BANWIDTH LIMITED

QoPH	QoS	QoE
Local (Dn) FEC: 0	Packet Loss = 9.75 %	PCR jitter = 5 ms
Local (Dn) FEC ES: 0	Packet Jitter = 1 – 5 ms	PAT = 321
Local (Dn) CV= 0	Packet Delay = 158 ms	PMT = 330
Local (Dn) ES= 0	Bandwidth = 2056 Mbps	PID timeouts = 0
Local (Dn) SES: 0	IGMP latency = 128 ms	Continuity errors = 48261
SNR Margin = 45 dB		
Local(Dn) HEC: 0		

The last scenario used interleaved DSL line that transports HDTV MPEG-2 TS. In this scenario, there were very little ES and SES indications. At the same time, FEC and FEC ES indications are numerous. Error control system corrected wrong bits and only a few bits that reached to the STB were bad. In network layer, there are no indication QoS parameters. Also continuity errors did not significantly affect IPTV QoE at customer side.

TABLE V. MPEG-2 HDTV SERVICE ADSL2+

QoPH	QoS	QoE
Local (Dn) FEC: 837248	Packet Loss = 0.00 %	PCR jitter = 2 ms
Local (Dn) FEC ES: 7084	Packet Jitter = 1 ms	PAT = 0
Local (Dn) CV= 165	Packet Delay = 25 ms	PMT = 0
Local (Dn) ES= 2	Bandwidth = 11568 Mbps	PID timeouts = 0
Local (Dn) SES: 2	IGMP latency = 72 ms	Continuity errors = 37
SNR Margin = 21.5 dB		
Local(Dn) HEC: 3968		

V. CONCLUSION

From above mentioned examples it can be noticed that there is strong dependence between network impairments and quality of DSL lines on one side and QoS/QoE on the other side. Precisely defined model that shows dependence between QoPh and QoE directly can help in finding mistakes on all OSI model layers. Since all network parameter errors can be noticed at the end of network i.e. at the customer side, model can be helpful in the defining quantity of their influence on QoE. After defining the model OPEX can be decreased due to possibility of revealing the part of the network where the problem appeared without unnecessary additional expenditure of technician work. It is obvious from the mentioned examples that QoS/QoE problems cannot be treated separately without modelling of QoPh i.e. physical layer influences in DSL environment. QoE strongly depends on error control of DSL and monitoring. It is obvious there is relation between equations (1) and (6). For further work, function should be found that maps errors from physical layer to higher layer. Also in DSL, there is relation between deviations of primary parameters and network impairments. Problem impulse noise or non-stationary noise will be expressed after implementation of vectored system. Relation between non-stationary noise and QoE should be examined.

REFERENCES

[1] World Broadband Statistics, „A Short Report from Global Broadband Statistics Q1/2012", Point Topic, Ltd., London, UK, June 2012,

[2] V. Oksman et al., "The ITU-T's new G.vector standard proliferates 100 Mb/s DSL," IEEE Communication Magazine, vol. 48, no. 10, October 2010, pp. 140–148,

[3] A. Begović, N. Goran, N, Škaljo, "An example of estimation of twisted pair channel capacity in DSL enviroment", accepted for publication,

[4] A. C. Begen., "Error Control for IPTV over xDSL Networks" in Proceeding of 5th IEEE Consumer Communications and Networking Conference CCNC, January 2008, pp. 632–637,

[5] J. A. Andersson, M. Kihl and S. Höst, "Impact of DSL link impairments on higher layer QoS parameters", in Proceedings of 8th Swedish National Computer Networking Workshop, Stochholm, Sweden, June 2012, pp. 95-98,

[6] Broadband forum TR-105, "ADSL2/ADSL2plus Functionality Test Plan", issue 2, November 2011,

[7] Broadband forum TR-059, "DSL Evolution - Architecture Requirements for the Support of QoS-Enabled IP Services", September 2003,

[8] Broadband forum TR-126, "Triple-Play Services Quality of Experience (QoE) Requirements", December 2006,

[9] J. Bae, H. S. Park, Y. M. Chin, H. C. Park, „An identification of correlation between network QoS and video quality index for IPTV service", in Proceedings of 11th International Conference on Advanced Communication Technology ICACT, Februar 2009, vol. 2, pp. 1309-1402,

[10] H. J. Kim, S. G. Choi, „A study on a QoS/QoE correlation model for QoE evaluation on IPTV service",in Proceedings of 12th International Conference on Advanced Communication Technology, February 2010, vol. 2, pp. 1377-1382,

[11] Agilent Technologies, „IPTV QoE: Understanding and interpreting MDI values", white paper, available online: http://cp.literature.agilent.com/litweb/pdf/5989-5088EN.pdf, date of access: 25.01.2013.

[12] Reference guide, „HST-3000 IP video testing user's guide", 211093453-003, Revision 000, JDS Uniphase Corporation, Mitipias, CA, USA, August 2008,

[13] Telecommunications cables, Catalogue B 02 05 08 07, Elka kabeli, d.o.o, Zagreb, Croatia, pp. 38-45.

METODOLOGIJA VREDNOVANJA KVALITETE USLUGE E-POZIVA U SKLOPU PILOT PROJEKTA HeERO

Pavao Britvić[1], Marijan Rimac[2], Krešimir Vidović[3]

[1]Državna uprava za zaštitu i spašavanje, Nehajska 5, 10 000 Zagreb
[2]Hrvatski autoklub,Avenija Dubrovnik 44, 10 000 Zagreb
[3]Ericsson Nikola Tesla d. d., Krapinska 45, 10 000 Zagreb

Sažetak: ***Usluga e-Poziva nova je telekomunikacijska usluga automatske dojave o prometnoj nesreći prema jedinstvenom operativno-komunikacijskom centru (Centar 112) u pilot primjeni u dvije pokretne elektroničke komunikacijske mreže. Usluga je operativno provedena u pokretnim mrežama Tele2 i VIPnet. Nacionalni konzorcij za provedbu pilot projekta e-Poziva predvode Državna uprava za zaštitu i spašavanje (DUZS), Ericsson Nikola Tesla d. d. i Hrvatski autoklub. Hrvatski pilot projekt djeluje u okviru paneuropskog projekta HeERO (Harmonised eCall EuRopean pilOts), koji za cilj ima evaluaciju i validaciju postavljenih normi i specifikacija usluge e-Poziva u okviru nacionalnih pilot-projekata, a financiran je od strane Europske komisije.***
U ovom radu predstavljena je metodologija testiranja e-Poziva u hrvatskom pilot projektu. Opisani su scenariji vrednovanja e-Poziva u laboratorijskim uvjetima te scenariji vrednovanja u operativnom okruženju (javna pokretna mreža). Predstavljeni su problemi i nedostaci uočeni u prvoj fazi testiranja e-Poziva te su prikazani prijedlozi aktivnosti i mjera za njihovo uklanjanje u narednim fazama vrednovanja.

Ključne riječi:

1. e-Poziv, HeERO

I. UVOD

e-Poziv je hitni poziv iniciran automatski od strane informacija iz osjetila u automobilu ili ručno od strane putnika u vozilu. Po aktivaciji, pruža obavijesti i relevantne informacije o položaju najprikladnijoj Centru 112 (pristupnoj točki javne sigurnosti) putem javnih pokretnih mreža te prenosi definiran i normiran minimalni set podataka, ukazujući na prometnu nesreću koja zahtijeva pažnju i intervenciju hitnih službi te uspostavlja izravnu glasovnu vezu između Centra 112 i putnika u vozilu. Usluga e-Poziva predstavlja važan doprinos povećanju sigurnosti u cestovnom prometu, ali i javnoj sigurnosti uopće. Republika Hrvatska potpisala je s Europskom komisijom Sporazum o razumijevanju u području implementacije usluge e-Poziva i aktivan je član paneuropskog projekta HeERO (Harmonised ecall EuRopean pilOts), koji za cilj ima evaluaciju i validaciju postavljenih normi i specifikacija usluge e-Poziva u okviru nacionalnih pilot-projekata.

e-Poziv je utemeljena na arhitekturi hitnog poziva 112. e-Poziv spada u skupinu telekomunikacijskih usluga vezanih za lokaciju, primijenjenih u području inteligentnih transportnih sustava. U ovom radu prikazani su scenariji testiranja u laboratorijskom okruženju i u stvarnom okružju, predstavljeni su problemi uočeni u prvoj fazi testiranja te su prikazani prijedlozi aktivnosti i mjera provedenih za uklanjanje uočenih nedostataka prije početka druge faze testiranja.

II. FUNKCIONALNI I OPERATIVNI ZAHTIJEVI

U uvodu europske norme ETSITS 102 689[1], e-Poziv je opisan kao hitni poziv iniciran automatski od strane informacija iz osjetila u automobilu ili ručno od strane putnika u vozilu. Po aktivaciji, pruža obavijesti i relevantne informacije o položaju najprikladnijem Centru 112 putem javnih pokretnih mreža te prenosi definiran i normiran minimalni set podataka, ukazujući na prometnu nesreću koja zahtijeva pažnju i intervenciju hitnih službi te uspostavlja izravnu glasovnu vezu između najprikladnije pristupne točke javne sigurnosti te putnika u vozilu. Nakon uspostave žurnog glasovnog poziva (112/E112) između vozila i Centra 112, uređaj za glasovnu komunikaciju koji sadrži mikrofon i zvučnik u vozilu odspaja se s linije za vrijeme transmisije minimalnog seta podataka kroz glasovni kanal prema Centru 112. Putnicima u vozilu na odgovarajući se način signalizira da je proces e-Poziva u tijeku. Po završetku prijenosa minimalnog seta podataka uređaj za glasovnu komunikaciju ponovno se spaja na glasovni kanal i glasovna komunikacija se ostvaruje između operatera Centra 112 i unesrećenih u vozilu. U trenutku uspostave glasovne komunikacije, a po prijenosu minimalnog seta podataka, ti podaci dostupni su na uvid operateru koji je zaprimio poziv. Nakon početnog zaprimanja poziva i rješavanja hitne situacije, operater Centra 112 može prekinuti poziv, ali, uređaj ugrađen u vozilu ostaje registriran na mobilnoj mreži, sukladno normi EN 16072 kako bi se omogućilo eventualno dodatno uspostavljanje komunikacije od strane Centra 112 ili žurnih službi prema unesrećenim osobama.
Tehnički zahtjevi e-Poziva, koji su vezani za uspostavu TS12 žurnog poziva i prijenos podataka iz modema ugrađenog u vozilo, su definirani u ETSI TS 122 101 i ETSI TS 124 008 normama. Ove norme također definiraju korištenje diskriminatora (tzv. eCall zastavice) e-Poziva unutar javne pokretna mreže koji je nužan za ispravno prepoznavanje, filtriranje i usmjeravanje e-Poziva prema

najprikladnijem Centru 112. Uređaj za ugradnju u vozilo, s ugrađenim modemom (eng. eCall in-band modem), koji se koristi za prijenos minimalnog seta podataka, definiran je normama ETSI TS 126 267 i ETSI TS 126 268.
U uvjetima normalnog rada, hodogram uspostave i tijeka e-Poziva sastoji se od devet koraka:

Korak 1 – Postupci nakon aktiviranja i inicijalizacije sustava u vozilu.
Korak 2 – Aktivacija sustava: a) generator e-Poziva inicira e-Poziv koristeći informacije iz osjetila u vozili ili ručnu aktivaciju. e-Poziv se sastoji od dva elementa, minimalni set podataka i glasovni poziv temelj na TS12 pozivu.
Korak 3 – Uspostava poziva, uključujući i identifikaciju vrste poziva, odabir mreže i registraciju na mrežu, bazno pozicioniranje, uspostava audio veze s modemom Centru 112. Mreža operatora pokretne telefonije omogućava uspostavu hitnog poziva (TS12) i usmjeriti glasovni poziv, uključujući i minimalni set podataka na najprikladnijem Centru 112, sukladno nacionalnoj regulativi. Mreža pokretne telefonije, s podrškom za e-Poziv, prepoznat će diskriminator e-Poziva, kao što je definirano u ETSI TS 124 008 normi, zaprimljen prilikom uspostave poziva, kako bi omogućio razlikovanje e-Poziva od „običnog“ hitnog TS12 poziva. Diskriminator e-Poziva može se koristiti za preusmjeravanje poziva sukladno lokalnim postavkama.
Korak 4 – Transfer minimalnog seta podataka, uključujući i odspajanje mikrofona i zvučnika s linije, slanje tona e-Poziva, sinkronizacija, zahtjev za slanje minimalnog seta podatak (engl. Minimum Set of Data, MSD), slanje MSD-a i provjeru pogreške, zatim potvrdu mrežnog sloja uključujući i zaustavljanje prijenosa MSD-a.
Korak 5 – Potvrda aplikacijskog sloja, Centar 112 prenosi potvrdu modemu u vozilu o uspješno zaprimljenom minimalnom setu podataka.
Korak 6 – Uspostava audio veze (uključujući i provjeru govornog kanala prema unesrećenima u vozilu, dekodiranje i prikaz minimalnog seta podataka te preusmjeravanje poziva i minimalnog seta podataka na drugi Centar 112.
Korak 7 – Stvaranje opće slike o hitnoj situaciji i određivanje položaja hitne situacije.
Korak 8 – Pokrenuti standardni operativni postupak za nesreću te glasovnim putem obavijestiti unesrećene da je obrada informacije o nesreći u tijeku i proslijeđena odgovarajućim hitnim službama.
Korak 9 – Prekid poziva
Procedure definirane normama moraju biti prihvaćene od strane lokalnih vlasti u pojedinim zemljama zbog potencijalnih nedostataka ili poteškoća u komunikaciji poput:

- uspostavljen samo glasovni poziv bez prijenosa podataka,
- neispravan transfer minimalnog seta podataka,
- neispravno generiranje/obrada diskriminatora e-Poziva,
- pogreška prilikom prijave na mobilnu mrežu,
- opća pogreška prilikom uspostave poziva,
- mreža mobilnog operatera ne podržava e-Poziv,
- e-Poziv preusmjeren na pristupnu točku javne sigurnosti koja nije opremljena za prihvat e-Poziva,
- kvar modema na strani pristupne točke javne sigurnosti,
- kvar mreže ili opreme na strani pristupne točke javne sigurnosti,
- problem s aplikacijom (programskom podrškom) na strani pristupne točke javne sigurnosti,
- operator na strani pristupne točke javne sigurnosti se ne javlja na poziv,
- Minimalni set podataka nije poslan,
- Minimalni set podataka nije zaprimljen,
- Nije uspostavljen glasovni kanal,
- Glasovni kanal uspostavljen ali dolazi do prekida,
- Ponovni pokušaj uspostave poziva i ponavljanja procedure slanja u slučaju neplaniranog prekida komunikacije,
- Automatski broj ponavljanja u slučaju neuspješnog poziva,
- Modem u vozilu ne zaprima obavijest o prekidu poziva.

III. SCENARIJI TESTIRANJA E-POZIVA

Usluga e-Poziva ima nekoliko karakteristična scenarija primjene. Identifikacija karakterističnih scenarija primjene i odgovarajućih indikatora uspješnosti koji opisuju ponašanje usluge e-Poziva (engl. Key Performance Indicators - KPIs) od ključnog je značenja za objektivnu provedbu postupka ispitivanja i validacije karakteristika usluge e-Poziva. Hrvatski e-Poziv Pilot, kao dio pan-europskog HeERO projekta, identificirao je četitri karakteristična seta scenarija korištenja usluge e-Poziva od kojih ćemo za potrebe ovog rada obraditi dva. Prvi set scenarija uključuje testiranje e-Poziva u laboratorijskom okruženju, dok drugi set scenarija uključuje testiranje u stvarnom okruženju. Opis karakterističnih scenarija korištenja usluge e-Poziva definiran je sljedećim temeljnim opisnim atributima: okruženje, broj uključenih vozila, broj uključenih vozila sa stranim registarskom oznakama, te način aktivacije usluge e-Poziva.
U laboratorijskom okruženju predviđeno je sedam različitih scenarija testiranja. Scenariji se razlikuju ovisno o broju IVS uređaja, o broju uređaja u roamingu (korištene su SIM kartice stranih mobilnih mreža), način aktivacije uređaja (automatski ili ručno) te broj ponovljenih aktivacija. U prvoj fazi testiranja u laboratorijskim uvjetima izvršeno je cca 800 različitih e-Poziva. Laboratorijski uvjeti sadržavali su sve komponente koje su potrebne za uspostavu usluge e-Poziva, što uključuje uređaj za ugradnju u vozilu, zatim emulirano mobilnu mrežu s radiopristupnim dijelom te pristupnu točku javne sigurnosti. Mobilna mreža i pristupna točka javne sigurnosti bile su zauzete samo testiranjem e-Poziva te preko njih nije vršen ostali promet. Također, mobilna mreža i pristupna točka javne sigurnosti bili su

nadograđeni kako bi omogućili realizaciju e-Poziva po propisanim normama, s korištenjem zastavice e-Poziva.

Tablica 1. Scenariji laboratorijskog testiranja

Oznaka	Broj IVS uređaja	Uređaji u roamingu	Način aktivacije	Broj ponavljanja aktivacije	Broj ponavljanja
L1	1	0	A	0	> 1000
L2	1	1	A	0	> 1000
L3	1	0	M	0	> 1000
L4	1	1	M	0	> 1000
L5	1	0	M	3	> 100
L6	1	1	M	3	> 100
L7	1	0	A	3	> 100

Tablica 2. Scenariji testiranja u realnom okruženju

Oznaka	Lokacija	Broj vozila	Broj uređaja u roamingu	Način aktivacije	Broj ponavljanja
R1	Zagreb, Centar, naseljeno	1	0	M	> 100
R2	Zagreb, obilaznica, nenaseljeno	1	1	M	> 100
R3	Zagreb, obilaznica,nenaseljeno	1	0	M	> 1000
R4	Lokalne ceste (okolica Zagreba)	1	0	M	> 1000

U realnom okruženju testiranje je vršeno kroz četiri različita karakteristična scenarija korištenja usluge e-Poziva. Scenariji testiranja razlikuju se u odnosu na okružje, u smislu nenaseljenog ili naseljenog okruženja, te u slučaju naseljenog okruženja testovi su vršeni na različitim lokacija ovisno o kategorizaciji prometnica, na lokalnim cestama ili na autocesti. Unutar scenarija razlikuju se načini aktivacije e-Poziva, kao i činjenica da li se koristio uređaj s lokalnom SIM karticom ili SIM karticom u roamingu. Cijeli set testiranja vršen je na području pokrivenosti signalom mobilne mreže partnera koji sudjeluje u projektu. Mobilna mreža nadograđena je funkcionalnošću za ispravno prepoznavanja i usmjeravanje e-Poziva, te je za potrebe testiranja korišten, sukladno normama, 112 broj s diskriminatorom e-Poziva. U realnom okruženju izvršeno je cca 4.000 e-Poziva.

III. OPAŽENI NEDOSTACI I PRIJEDLOZI ZA NJIHOVO UKLANJANJE

Cilj pilot projekta jest evaluirati da li zahtijevane performanse e-Poziva mogu biti realizirane u realnom okruženju sa postojećim javnim pokretnim mrežama i postojećim operativno-komunikacijskim centrom. To znači da su testovi fokusirani na provjeru ispunjavanja normi koje opisuju e-Poziv kao i na mjerenje ključnih identifikatora performansi. Zaključak prve faze testiranja jest da je paneuropsko rješenje e-Poziva robusna i pouzdana usluga koja se odvija sukladno propisanim normama uodgovarajućem okružju (visoka razina signala javne pokretne mreže, dostupnost GPS signala) odnosno da dolazi do smanjenja performansi usluge u uvjetima u kojima ti parametri nisu na odgovarajućoj razini. Većina testova izvršena u prvoj operativno fazi e-Poziva imala je snažan operativni fokus s glavnim ciljem u smislu verifikacije normi koje su razvijene i odobrene od strane europskih standardizacijskih tijela, kao i operativnih procedura razvijenih u sklopu pilot projekta, kako bi se utvrdile performanse usluge e-Poziva kao i potvrdilo da nova usluga nema utjecaja na postojeću uslugu žurnog poziva 112. Tijekom prve faze testiranja utvrđeno je da se scenariji testiranja i operativne procedure moraju prilagoditi prije početka naredne faze testiranja. Izvršeni su nužni laboratorijski testovi, kao i testovi u realnom okruženju kako bi se utvrdilo da su mjereni i kasnije analizirani svi ključni pokazatelji performansi koji ukazuju na pouzdanost i robusnost usluge. Kako bi se potvrdilo da je riječ o harmoniziranoj i kompatibilnoj usluzi, predviđen je i izvršen set testova interoperabilnosti između različitih zemalja koje sudjeluju u pilot projektu. Prva faza testiranja ukazala je da su u pilot projektu e-Poziva u Republici Hrvatskoj rezultati zadovoljavajući i u skladu s propisanim normama, ali su uočeni nedostaci koji će biti uklonjeni do početka druge faze testiranja.

Uočen je problem vremenske sinkronizacije između operativno komunikacijskog centra i uređaja za ugradnju u vozilo. Obzirom na zatvorenost sustava operativno komunikacijskog centra koja iz sigurnosnih razloga ograničava sučelja za razmjenu podataka s javnim mrežama, potrebno je definirati novi način za precizno usklađivanje vremena, kako bi se smanjile pogreške u mjerenjima uzrokovane neusklađenosti vremena između uređaja za ugradnju u vozilo i operativnog komunikacijskog centra. Rješenje je pronađeno u poslužitelju vremenskih podatka koji vrijeme ažurira iz istog izvora kao i IVS (engl. In-Vehicle System, IVS), te je implementacijom tog sustava minimizirana potencijalna pogreška u usklađivanju vremena. Primijećeno je da prve, inicijalne inačice uređaja za ugradnju u vozilo (IVS) uslijed vibracija mogu ostati bez napajanja, što uzrokuje ponovno pokretanje uređaja odnosno gubitak pohranjenih postavki. Ovaj problem adresiran je proizvođaču uređaja te se očekuje da će u narednim revizijama biti ispravljen. Uređaj za ugradnju u vozilo funkcionira sukladno zahtjevima okružju gdje su signali kako pokretne mreže, tako i GPS sustava na zadovoljavajućoj razini. Ukoliko signal javne pokretne mreže nije na zadovoljavajućoj razini, uređaj pokušava generirati e-Poziv koristeći signal druge mreže, koji je u tom trenutku jačeg intenziteta. Zbog nepostojanja zastavice e-Poziva u drugoj mreži pomoću koje IVS pokušava ostvariti poziv, uređaj ponavlja pokušaje u petlji. Ovaj problem uobičajen je u svim državama koje sudjeluju u projektu a uzrokovan je

činjenicom da sve javne pokretne mreže nisu nadograđene za funkcionalnost e-Poziva. Ovaj problem će biti riješen kad svi mobilni operateri nadograde svoje javne pokretne mreže nadogradnjom za ispravno prepoznavanje i usmjeravanje e-Poziva.

IV. ZAKLJUČAK

Usluga e-Poziva nova je telekomunikacijska usluga automatske dojave o prometnoj nesreći prema jedinstvenom operativno-komunikacijskom centru (Centar 112) u pilot primjeni u dvije pokretne elektroničke komunikacijske mreže. Po završetku implementacije e-Poziva u laboratorijskim uvjetima, kao i u izdvojenom sustavu u realnom okruženju izvršena je prva faza testiranja usluge e-Poziva sukladno definiranim scenarijima. Većina testova izvršena u prvoj operativno fazi e-Poziva imala je snažan operativni fokus s glavnim ciljem u smislu verifikacije normi koje su razvijene i odobrene od strane europskih standardizacijskih tijela, kao i operativnih procedura razvijenih u sklopu pilot projekta, kako bi se utvrdile performanse usluge e-Poziva kao i potvrdilo da nova usluga nema utjecaja na postojeću uslugu hitnog poziva 112. Prva faza testiranja pokazala je da je usluga pouzdana i robusna, te sukladna normama koje ju opisuju kako u laboratorijskim uvjetima tako i u realnom okruženju. Uočeni su nedostaci u scenarijima i u realizaciji testiranja koji su modificirani za drugu fazu testiranja. Nedostaci su ispravljeni uvođenjem novih elemenata u implementaciji e-Poziva kao i modifikacijom scenarija.

V. LITERATURA

[1] ETSI TS 102 689 v1.1.1 (2010-08) Machine to Machine communications (M2M) ; M2M service requirements

[2] Filjar, R., G. Jezic, M. Matijasevic. (2008). Location-Based Services: A Road Towards Situation Awareness. *Journal of Navigation*, **61**, 573-589.

[3] Filjar R., Simunic D., Vidovic K., A. Sarjanovic, and P. Britvic, "A methodology for eCall performance laboratory examination," in 19th ITS World Congress, 2012.

[4] Filjar R., K. Vidovic, P. Britvic, and M. Rimac, "eCall: Automatic notification of a road traffic accident," in 2011 Proceedings of the 34th International Convention MIPRO, 2011, vol. I, no. May 2010, pp. 600–605.

[5]Brown, M C. (2006).Hacking Google Maps and Google Earth.John Wiley & Sons. Indianapolis, IN

Utjecaj kvalitete usluga i zadovoljstva korisnika na lojalnost korisnika usluga

Harun Kuč, dipl. ing. el. *, mr. sci. Adnan Kuč, dipl. ing. el.**
* BH Telecom d.d. Sarajevo/Služba za nadzor i analizu TKS, Zenica, Bosna i Hercegovina
** BH Telecom d.d. Sarajevo/Služba za podršku BH Mobile, Zenica, Bosna i Hercegovina
harun.kuc@bhtelecom.ba
adnan.kuc@bhtelecom.ba

Sažetak - U uvjetima liberalizacije tržišta telekomunikacija, mobilni operateri sve više prepoznaju i prihvaćaju činjenicu da je zadržavanje postojećih korisnika znatno profitabilnije od privlačenja novih. Zato se zadržavanje postojećih korisnika i lojalnost korisnika sve češće nalazi u središtu njihovog zanimanja i predstavlja prioritet svakog perspektivnog mobilnog operatera. U radu prezentiramo rezultate istraživanja utjecaja kvaliteta usluga i zadovoljstva korisnika na lojalnost korisnika usluga. Testirane su hipoteze o tome kako kvaliteta, imidž, cijena, zadovoljstvo i povjerenje korisnika utječu na lojalnost korisnika.

I. UVOD

Početkom 2006. godine liberalizirane su sve usluge na tržištu telekomunikacija u BiH i time su stvoreni dodatni uvjeti za razvoj konkurencije te se kao osnovni cilj nalaže daljnje jačanje konkurentnosti u svim segmentima tržišta. Kako bi bili konkurentni na tržištu, mobilni operateri trebaju pružati uslugu u skladu s potrebama i stavovima korisnika usluga.

U situaciji kada svi konkurentni mobilni operateri imaju kvalitetnu uslugu, korisniku nije bitno koga će odabrati osim ako nema osjećaj da mobilnom operateru može više vjerovati i da mobilni operater želi da se korisnik osjeća zadovoljno prilikom korištenja njihovih usluga. Zadovoljni korisnici postaju s vremenom lojalni i tako nam omogućavaju sigurnije i bolje prodajne rezultate. Lojalni korisnici našim dodatnim zalaganjima postaju oduševljeni kupci tako što nas preporučuju drugima i na taj način nam ponovo osiguravaju bolji prodajni rezultat i skraćuju naše vrijeme pronalazaka novih korisnika [18]. Dakle, zadovoljan korisnik je jedan od preduvjeta lojalnosti. Zadovoljan i lojalan korisnik nije slučajnost, već koncept koji se gradi, stoga ponašanje kupaca može i treba biti mjereno kako bi rezultate tog mjerenja koristili u strategiji razvoja zadovoljnih kupaca. Ovo ukazuje da bi mobilni operateri trebali redovno mjeriti zadovoljstvo korisnika. Nastojanje mobilnog operatera na zadržavanju korisnika ne znači potpuno odustajanje od privlačenja novih potencijalnih korisnika. Upravo oni predstavljaju stabilan izvor za osiguranje profitabilnosti kao i izbor korisnika koje vrijedi zadržati i čiju lojalnost vrijedi pridobiti. Mobilni operateri također moraju znati kad treba i odustati od nastojanja da se gradi lojalnost među korisnicima koji ne predstavljaju vrijednost za mobilnog operatera [20] [27] [28] [29].

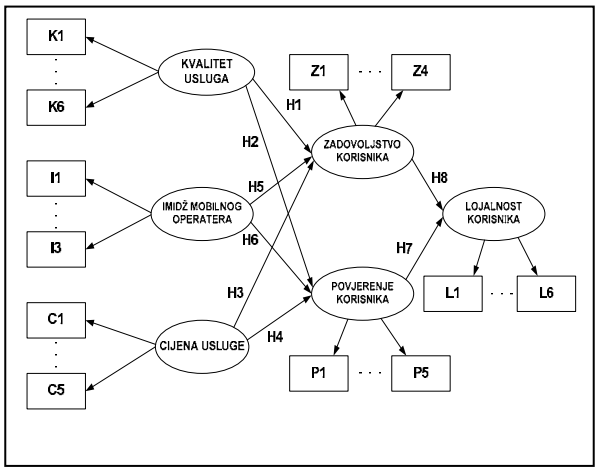
U skladu s prethodno navedenim provedeno je istraživanje na modelu čiji je cilj bio da se utvrdi koji su to faktori koji će zadržati postojeće i privući nove korisnike, a samim tim i osigurati veći prihod postojećem mobilnom operateru.

II. PRETHODNA ISTRAŽIVANJA

Kvaliteta usluga predstavlja ključni faktor za zadovoljstvo korisnika, prema istraživanju [30], iako to nije jedini faktor [9] u prethodnim istraživanjima pokazali su da zadovoljstvo korisnika ima pozitivan i veliki utjecaj na lojalnost korisnika. Prethodna istraživanja [26] su pokazala da zadovoljstvo korisnika ima značajan utjecaj prilikom izgradnje dugoročnih i profitabilnijih odnosa sa svojim korisnicima [14] [23]. Zadovoljstvo korisnika se može definirati kao "razlika između očekivanja i stvarne percepcije usluga" i smatra se najvažnijim faktorom koji vodi prema konkurentskoj prednosti i uspjehu, a to je ono što čini korisnike lojalnim prema mobilnom operateru. Povjerenje korisnika predstavlja preduvjet za izgradnju i održavanje dugoročnih odnosa i ima direktan utjecaj na lojalnost korisnika [4] [5].

Kvaliteta usluga ima utjecaj na zadovoljstvo i lojalnost korisnika [4] [6]. Kvaliteta usluga predstavlja preduvjet zadovoljstva korisnika i direktno utječe na povjerenje korisnika [7] [9]. Mnogi istraživači su dokazali da cijena ima utjecaj na zadovoljstvo i povjerenje korisnika. [19] [21] .

Također, u prethodnim istraživanjima [25] uočeno je da imidž i cijena značajnije utiču na stupanj zadovoljstva korisnika usluga, od bilo kojeg drugog faktora. Cijena igra važnu ulogu u telekomunikacijskom tržištu, posebno za mobilne operatere [23] [24]. Cijena pozitivno utiče na zadovoljstvo i lojalnost korisnika [15], [16]. Imidž ima „marginalno značajan“ direktan utjecaj na lojalnost korisnika, ali značajan uticaj putem zadovoljstva korisnika. Dakle, imidž može utjecati na lojalnost direktno i indirektno [2] [14].



Slika 1. Predloženi model mjerenja utjecaja kvalitete usluga i zadovoljstva korisnika na lojalnost korisnika usluga

Na osnovu prethodnih istraživanja [1] [22] [17], kvaliteta usluga, imidž i cijena predstavljaju faktore koje korisnici koriste prilikom odabira mobilnog operatera. Imidž predstavlja sveobuhvatnu sliku o mobilnom operateru koja se temelji na percepciji korisnika. Percipirana cijena od strane korisnika predstavlja uvjerenje da je kvaliteta usluga ostala na istom nivou. Na osnovu percipirane cijene i kvalitete usluga može se održati imidž mobilnog operatera. U skladu s navedenim možemo zaključiti da postoji korelacija između faktora kvalitete usluga, imidža i cijene [22].

III. PREDLOŽENI MODEL MJERENJA

Kako je lojalan korisnik izvor stabilnih prihoda i promotor usluga, vrlo je važno testirati hipoteze o tome kako kvaliteta, imidž, cijena, zadovoljstvo i povjerenje korisnika utječe na lojalnost korisnika [8] [11].

Polazeći od pretpostavke da kvaliteta, imidž i cijena direktno utječu na zadovoljstvo i povjerenje korisnika usluga, a zadovoljstvo i povjerenje na lojalnost korisnika [1] [22] [31], predložen je model mjerenja zadovoljstva i lojalnosti korisnika usluga mobilne mreže koji je prikazan na slici 1., pri čemu se oznake K1-K6, I1-I3, C1-C5, Z1-Z4, P1-P5 i L1-L6 odnose na niz pitanja koja odgovaraju pojedinim faktorima u anketnom upitniku dat u tablici I.

U skladu s istim možemo definirati sljedeće hipoteze o odnosima između kvalitete, cijene, imidža, zadovoljstva i lojalnosti korisnika usluga:

H1: Visoka kvaliteta usluga ima pozitivan i direktan utjecaj na zadovoljstvo korisnika.

H2: Visoka kvaliteta usluga ima pozitivan i direktan utjecaj na povjerenje korisnika.

H3: Cijena usluge ima pozitivan i direktan utjecaj na zadovoljstvo korisnika.

H4: Cijena usluge ima pozitivan i direktan utjecaj na povjerenje korisnika.

H5: Imidž mobilnog operatera ima pozitivan i direktan utjecaj na zadovoljstvo korisnika.

H6: Imidž mobilnog operatera ima pozitivan i direktan utjecaj na povjerenje korisnika.

H7: Povjerenje korisnika ima pozitivan i direktan utjecaj na lojalnost korisnika.

H8: Zadovoljstvo korisnika ima pozitivan i direktan utjecaj na lojalnost korisnika.

IV. UZORAK I NAČIN ISTRAŽIVANJA

Istraživanje je provedeno na uzorku od 739 slučajno odabranih korisnika mobilnih usluga. Podaci su prikupljeni putem anketnog upitnika (radeći klasičnu anketu presretanja) na području općine Zenica.

Ispitanici su popunjavali anketni upitnik u kojem se od njih tražilo da putem Likertove skale (od 1-u potpunosti se ne slažem do 5-u potpunosti se slažem) ocijeni stupanj slaganja s 29 pitanja u kojima su iznesene pretpostavke o elementima ponude koju mobilni operateri moraju osigurati da bi stvorili pretpostavke za zadovoljstvo korisnika. Prikaz osnovnih karakeristika uzorka istraživanja dan je u tablici II.

TABLICA II. OSNOVNE KARAKTERISTIKE UZORKA

Varijabla	Kategorija	Frekvencija	%
Spol	Muški Ženski	472 267	63,9 36.1
Starosna struktura	Manje od 20 godina 20-24 godina 25-34 godina Više od 34 godina	110 294 259 76	14,9 39,8 35,0 10,3
Vremenski period korištenja usluga mobilnog operatera	Manje od 6 mjeseci 7 – 12 mjeseci 13 – 24 mjeseci 25 – 36 mjeseci Više od 36 mjeseci	10 77 224 289 139	1,4 10,4 30,3 39,1 18,8
Mobilni operater	Eronet BH Telecom M:tel	122 357 260	16,5 48,3 35,2

V. REZULTATI ISTRAŽIVANJA

Prikupljeni podaci analizirani su putem statističkih programa SPSS (eng. Statistical Package for the Social Sciences) i AMOS (eng. Analysis of Moment Structures). Kao preduvjet za ocjenu prikladnosti predloženog modela neophodno je ispunjenje visoko zahtjevnog uvjeta koji se odnosi na validnost i pouzdanost mjernog instrumenta i sastoji se u provjeri jednodimenzionalnosti mjerne skale, odnosno provjeri sposobnosti mjernog instrumenta da njegovim korištenjem dobijemo konzistentne rezultate.

A. Ispitivanje pouzdanosti mjerne skale

Mjerna skala za potrebe istraživanja treba biti pouzdana. Pouzdanost se može posmatrati s više aspekata. Jedno od glavnih pitanja odnosi se na unutrašnju saglasnost skale, to jest stepen srodnosti iskaza od kojih se skala sastoji. Da li svi oni mjere isti faktor ili konstrukt? Među najčešće upotrebljavanim pokazateljima unutrašnje saglasnosti je Cronbach-α koeficijent. U idealnom slučaju bi ovaj koeficijent trebalo da bude veći od 0,7. Međutim, njegove vrijednosti su veoma osjetljive na broj iskaza na skali. Na opaženim podacima iz našeg uzorka, ispitivanje unutrašnje saglasnosti svake podskale je izvršeno u skladu sa iskazima koji se odnose na odgovarajuću faktor izmjerenu posredstvom mjernog instrumenta. Dakle, na odgovorima korisnika usluga analiziramo srodnost određenih iskaza da li oni mjere odgovarajući faktor [12].

U tablici III. dat je prikaz vrijednosti Cronbach-α koeficijenata odgovarajućih varijabli koje se odnose na pojedine faktore imidž, cijenu, kvalitet, povjerenje, zadovoljstvo i lojalnost korisnika.

TABLICA III. VRIJEDNOST CRONBACH-α KOEFICIJENTA

Faktor	Varijable	Cronbach-α
Imidž mobilnog operatera	p27 - p29	0,845
Cijena	p22 - p26	0,919
Kvalitet usluga	p16 - p21	0,915
Povjerenje korisnika	p11 - p15	0,892
Zadovoljstvo korisnika	p7 - p10	0,853
Lojalnost korisnika	p1- p6	0,909

Na temelju Cronbach-α koeficijenata dobivenih putem SPSS programa moguće je zaključiti kako primijenjene mjerne podskale posjeduju idealne nivoe pouzdanosti.

B. SEM analiza

SEM (eng. Structural Equation Modeling) predstavlja multivarijantnu statističku tehniku koja predstavlja kombinaciju CFA (eng. Confirmatory Factor Analysis), PA (eng. Path Analysis) i regresijske analize i primjenjuje se u analizi hipotetičkih odnosa izmedu endogenih (varijable koje se ne mogu direktno izmjeriti) i egzogenih (varijable koje se mogu direktno mjeriti) varijabli. Koristeći statistički program AMOS koji se koristi za SEM modeliranje i simulacije, SEM statističku analizu i obradu podataka kao i za grafički prikaz rezultata SEM analize, izvršena je provjera postavljenih hipoteza [3].

Na slici 2. je prikazan dijagram modela utjecaja kvalitete usluga i zadovoljstva korisnika na lojalnost korisnika usluga sa izračunatim parametrima posredstvom AMOS programa, pri čemu oznake e1- e29 predstavljaju rezidualnu varijansu (nastaje pod uticajem nekontrolisanih faktora) egzogenih varijabli p1- p29, dok oznake e30 - e32 predstavljaju rezidualnu varijansu endogenih varijabli [9] [10].

Na temelju provedene SEM analize prvo se pristupilo utvrđivanju stupnja prikladnosti predloženog modela s empirijskim modelom. U tablici IV. dat je prikaz indeksa kojima se mjeri stupanj prikladnosti:

TABLICA IV. INDEKSI PRIKLADNOSTI PREDLOŽENOG MODELA S EMPIRIJSKIM MODELOM

Indeksi prilagođenja	Vrijednost	Granične vrijednosti
$\chi 2/df$	2,493	< 3
RMR	0,019	< 0.08
GFI	0,926	> 0.9
AGFI	0,912	> 0.9
NFI	0,941	> 0.9
RFI	0,934	> 0.9
IFI	0,964	> 0.9
TLI	0,959	> 0.9
CFI	0,963	> 0.9
RMSEA	0,045	< 0.05

Indeksi $\chi 2/df$ (eng. Chi square to degrees of freedom), NFI (eng. Normed Fit Indeks), TLI (Tucker Lewis Indeks), CFI (eng. Comparative Fit Indeks), IFI (eng. Incremental Fit Indeks) i RFI (eng. Relative Fit Indeks) predstavljaju relativne pokazatelje prikladnosti postavljenog modela s dva referentna modela: mjernog i idealnog modela, dok indeksi GFI (eng. Goodness of Fit Indeks), AGFI (eng. Adjusted Goodness of Fit Indeks), RMR (eng. Root Mean square Residuals) i RMSEA (eng. Root Mean Square Error of Approximation) predstavljaju apsolutne pokazatelje prikladnosti postavljenog modela i opaženih podataka [13]. Vrijednosti dobivenih indeksa upućuju na zaključak da je nivo prikladnosti definiranog modela podacima zadovoljavajuća, tj. postoji preklapanje empirijske i ciljne matrice, te je definirani model prihvatljiv za daljnju analizu.

Tablica V. prikazuje korelacije između faktora i pokazuje da su sve varijable pozitivno korelirane jedna s drugom. Faktor lojalnost i zadovoljstvo korisnika pokazuju najveći stupanj povezanosti 0,739.

Rezultirajući korelacijski koeficijenti upućuju na zaključak da se hipoteze H1 do H8 mogu smatrati potvrđenima. Svi korelacijski koeficijenti su pozitivni.

Tablica VI. prikazuje standardizirane strukturne koeficijente kojima se procjenjuju direktne uzročne veze između latentnih varijabli, specificirane u ranije definiranom strukturnom modelu. Iz tablice V. se vidi da je p-vrijednost za svaku korelaciju manja od 0.001 (*** $p < 0.001$) što znači da postoji statistički značajna razlika između ovih faktora. Na faktor zadovoljstva najviše utječe faktor kvalitete, jer ima najveći procijenjeni koeficijent 0,324, a najmanje faktor imidža (jer ima najmanji koeficijent 0,227). To znači da zadovoljstvo najviše ovisi o kvaliteti servisa, najmanje o imidžu mobilnog operatera. Također, vidimo da na faktor povjerenja najviše utječe faktor imidža, jer ima najveći procijenjeni koeficijent 0,342, a najmanje faktor kvalitete, jer ima najmanji procijenjeni koeficijent 0,161. Na faktor lojalnosti najveći utjecaj ima faktor zadovoljstva s koeficijentom 0,528.

Dakle, možemo zaključiti da se lojalnost direktno može postići iz zadovoljstva korisnika, a indirektno iz kvalitete servisa.

Rezultirajući standardizirani strukturni koeficijenti upućuju na zaključak da se hipoteze H1 do H8 mogu smatrati potvrđenima. Svi strukturni koeficijenti su statistički značajni te su predviđenog smjera.

VI. ZAKLJUČAK

Ovo istraživanje produbljuje spoznaje o utjecaju zadovoljstva na lojalnost korisnika usluga mobilne mreže. Primjenom SEM analize testirane su postavljene hipoteze. Potvrđeno je svih osam hipoteza. Strukturni koeficijenti su statistički značajni i predviđenog su smjera. Na osnovu rezultata pokazali smo da kvaliteta (indirektno) i zadovoljstvo (direktno) imaju veliki utjecaj na lojalnost korisnika usluga mobilne mreže.

U ovom istraživanju potvrdili smo da kvaliteta, imidž i cijena pozitivno i indirektno utječu na lojalnost korisnika prema mobilnom operateru. Prema istraživanju [33], kvaliteta usluga, imidž i cijena pozitivno i direktno utječu na lojalnost korisnika prema mobilnom operateru.

Također, potvrdili smo da kvaliteta, imidž i cijena pozitivno i direktno utječu na zadovoljstvo korisnika, kao i da postoji korelacija između faktora kvalitete, imidža i cijene. Prema istraživanju [22], kvaliteta usluga, imidž i cijena pozitivno i direktno utječu na zadovoljstvo korisnika prema mobilnom operateru. Prema istraživanju [22] postoji korelacija između faktora kvalitete, imidža i cijene.

Prema istraživanju [32], kvaliteta (indirektno) i zadovoljstvo (direktno) imaju pozitivan i velik utjecaj na lojalnost korisnika usluga mobilnog operatera. To znači da povećanjem kvalitete usluga dolazi do povećanja zadovoljstva korisnika, a samim tim i do povećanja lojalnosti korisnika usluga mobilnog operatera. Ovo smo potvrdili našim istraživanjem.

Iako ova metodologija istraživanja ima i svojih prednosti i mana, u svrhu detaljnijeg uvida u uzročno-posljedične odnose između varijabli potrebno je provesti niz dodatnih istraživanja kojima bi se ispitali odnosi utjecaja kvalitete, imidža, cijene na lojalnosti korisnika, kao i utjecaj dimenzija kvaliteta usluga na zadovoljstvo korisnika.

Za buduća istraživanja bilo bi zanimljivo istražiti utjecaj usluge prenosivosti mobilnog broja (prelazak korisnika kod drugih mobilnih operatera zadržavajući postojeći mobilni broj) na lojalnost korisnika, kao i utjecaj programa nagrađivanja lojalnosti korisnika od strane mobilnih operatera na samu lojalnost korisnika.

LITERATURA

[1] A. Raza, and Z. Rehman, "Impact of relations marketing tactics on relationship quality and coustomer loyalty: A case study of telecom sector of Pakistan,", African Jurnal of Business Managment Vol.6(14), 2012, pp. 5085-5092.

[2] A. Hart, and P. Rosenberger, "Effect of corporate image in the formation of customer loyalty: an Australian replication", Australian Marketing Journal, 2004, pp. 88-96.

[3] B. Byrne, "Structural Equation Modeling With AMOS, Basic Concepts, Applications, and Programming, Second edition, 2010.

[4] B. Lewis, "Service quality: recent developments in financial services", International Journal of Bank Marketing, 1993, pp. 19-26.

[5] B. Corbitt, T. Thanasankit, and H. Yi, "Trust and e-commerce: a study of consumer perceptions", Electronic Commerce Research & Applications, 2003, pp. 203-215.

[6] D. Rousseau, S. Sitkin, R. Burt, and C. Camerer, "Not so different after all: a cross-discipline view of trust", Academy of Management Review, 1998, pp. 393-404.

[7] E. Anderson, C. Fornell, and D. Lehmann, "Customer Satisfaction, Market Share, and Profitability: Findings from Sweden", Journal of Marketing, 1994, pp. 53-66.

[8] E. Garbarino, and M. Johnson, "The different roles of satisfaction, trust and commitment in customer relationships", Journal of Marketing, 1999, pp. 70-87.

[9] H. Kuč, "Statističke funkcije u Excel-u kroz primjere", Zenica, 1999.

[10] H. Kuč, "Statistika u Excelu", Zenica, 2001.

[11] J. Cronin, and S. Taylor, "Measuring service quality: a reexamination and Extension. Journal of Marketing", 1992, pp. 55-68.

[12] A. Kuč, "Faktorska struktura funkcionalnog kvaliteta telekomunikacijskih usluga", Magistarski rad, Sarajevo, 2011.

[13] R. Schumacker, and R. Lomax, "A beginner's guide to structural equation modeling", 2-nd ed., 2004.

[14] S. Aydin, and G. Ozer, „The analysis of antecedents of customer loyalty in the Turkish mobile telecommunication market", European Journal of Marketing, 2005, pp. 910-925.

[15] S. Varki, and M. Colgate, "The role of price perceptions in an integrated model of behavioral intentions", Journal of Services Research, 2001, pp.232–240.

[16] R. Bolton, and J Drew, "A multistage model of customers' assessments of service quality and value", Journal of Consumer Research, 1991, pp. 375-384.

[17] A. Eshghi, and D. Haughton, and H Topi, "Determinants of customer loyalty in the wireless telecommunication industry", Telecommunication policy, 2007, pp. 93-106.

[18] H. Herzog, "Behavioral Science Concepts for Analyzing the Consumer", Marketing and the Behavioural Sciences, 1963, pp.76-86p.

[19] R. Oliver, „When Consumer Loyalty", The Journal of Marketing, Fundamental Issues and Directions for Marketing 63, 1999, pp.33-44.

[20] D. Tse, and P. Wilton, „Models of Consumer Satisfaction Formation: An Extension", Journal of Marketing Research 25, 1988, pp.204-12.

[21] T. Cheng, L. Lai, and A. Yeung, "The Driving Forces of Customer Loyalty: A Study of Internet Service Providersin Hong Kong", International Journal of E-Business Research , 2008, pp.26-42.

[22] M. Muhammad, G. Muhammad and I. Hafiz, "Impact of Brand Image, Service Quality and price on customer satisfaction in Pakistan Telecommunication sector", Internacional Journal of Business and Social Science, 2012, pp. 123-129.

[23] A. Herrmann, L. Xia, K. Monroe, and Huber, "The Influence of Price Fairness on Consumer Satisfaction: An Empirical Test in the Context of Automobile Purchases", Journal of Product and Brand Management, 2007, pp.49-58.

[24] T. Kollmann, "The Price /Acceptance Function: Perspectives of a Pricing Policy in European Telecommunication Markets", European Journal of Innovation Management, 2000, pp. 7-14.

[25] W. Olatokun, and S. Nwonne, "Determinants of Users Choise of Mobile Services Providers in the Nigerian Telecommunications Market", African Journal of Computing and ICT, 2012, pp.19-32.

[26] B. Afsar, Z. Rehman, A. Quereshi and A. Shahjehan, "Determinants of coustomer loyalty in the banking sector: The case of Pacistan", African Journal Business and Management, 2010, pp.1040-1047.

[27] I. Ahmed, M. Nawaz, A. Usman, M. Shaukat, N. Ahmed, and U. Rehman, "A mediation of coustomer satisfaction relationship between service quality and repurchase intentions for the telecom sector in Pakistan: A case study of university students", African Journal Business and Management, 2010, pp.3457-3462.

[28] I. Ali, K. Rehman, A. Yilmaz, S. Nazir, and J. Ali, "Effects of corporate social responsibility on coustomer retention in cellular industry of Pakistan", African Journal Business and Management, 2010, pp.457-485.

[29] M. Butt, and E. Run, "Modeling coustomer satisfaction in cellular phone services", J. Kemanusiaan Bil., 2009, pp. 79-93.

[30] A. Maiyaki, N. Noor, and S. Mokhtar, "The influence of service of service quality of mobile phone on coustomer satisfaction in Malaysia: A students feedback survey", Journal Business and Management, 2011, pp.79-97.

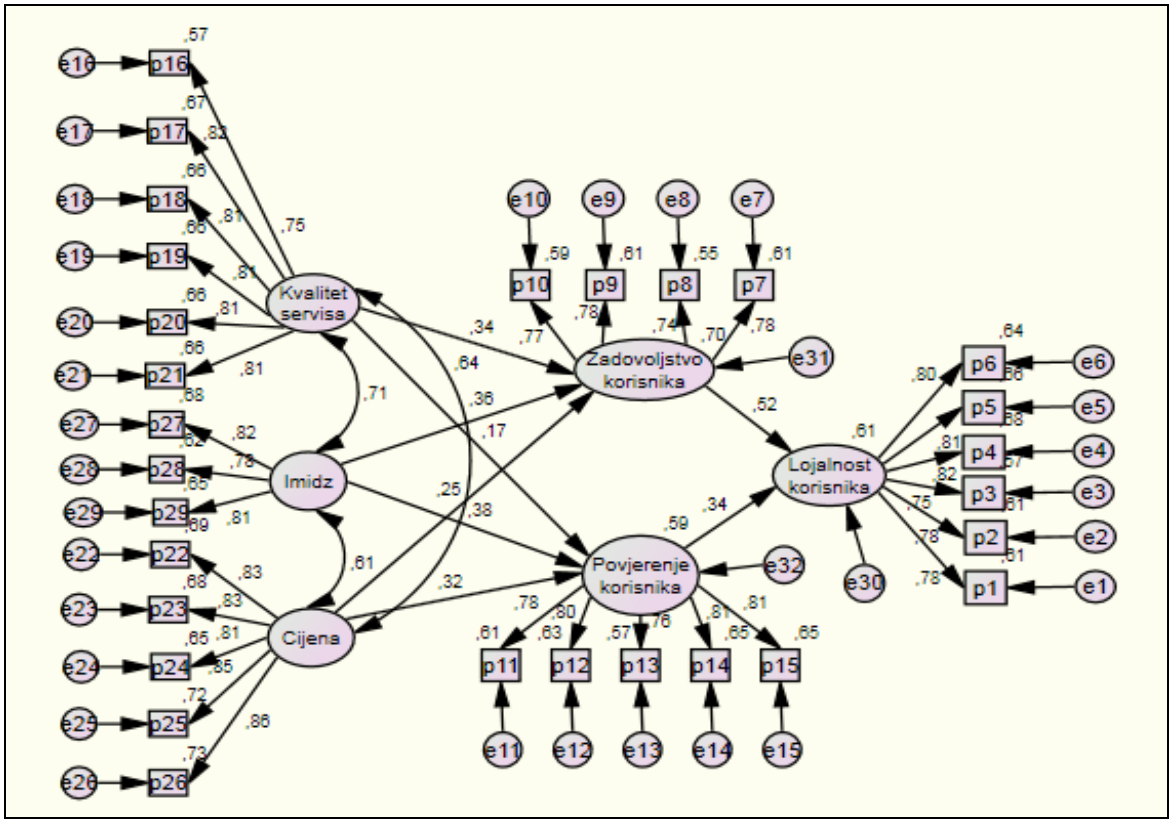
[31] S. Zaim, A. Turkyilmaz, M. Tarim, B. Ucar, and O. Akkas, "Measuring coustomer satisfaction in Turk Telekom Company using structural equation modeling technique", Journal of Global Strategic and Management, 2010, pp.89-99.

[32] M. Sattar, and B. Sattar, "Coustomer satisfaction affects the coustomer loyalty: Evidence from telecommunication sector in Pakistan", Asian Journal of Business Management, 2012, pp. 252-259.

[33] K. Young, and L. Jung, "Relation between corporate image and coustomer loyalty in mobile communications service markets", African Journal Business Management, 2010, pp.4035-4041.

TABLICA I. PITANJA KOJA SE ODNOSE NA POJEDINE FAKTORE U ANKETNOM UPITNIKU

Oznaka	Pitanja
L1- p1	1. Namjeravam nastaviti koristiti mobilne usluge ovog mobilnog operatora dugo vremena.
L2- p2	2. Ako želim koristiti dodatne mobilne usluge, ja sam spreman nastaviti ponovo odabrati ovog mobilnog operatora.
L3- p3	3. Čak i ako je cijena je niža kod drugog mobilnog operatera, ja ću nastaviti koristiti usluge ovog mobilnog operatera.
L4- p4	4. Ja sam spreman reći pozitivne stvari o ovom mobilnom operateru drugim korisnicima.
L5- p5	5. Ja ću potaknuti prijatelje i rodbinu da koristite usluge koje nudi ovaj mobilni operater.
L6- p6	6. Za mene, ovaj mobilni operater je u mogućnosti pružiti najbolju mobilnu uslugu.
Z1- p7	7. Zadovoljan sam s ukupnom kvalitetom usluga koje nudi ovaj mobilni operater.
Z2- p8	8. Zadovoljan sam s profesionalnošću ovog mobilnog operatora.
Z3- p9	9. Zadovoljan sam nastupom zaposlenika ovog mobilnog operatora.
Z4- p10	10. Zadovoljan sam prijatnim okruženjem za poslovanje s ovim mobilnim operatorom.
P1- p11	11. Ovaj mobilni operater je pouzdan jer se uglavnom bavi interesom korisnika.
P2- p12	12. Sistem naplate koji ima ovaj mobilni operater je pouzdan.
P3- p13	13. Ugled koji ima ovoj operator je pouzdan.
P4- p14	14. Politika i praksa koju vodi ovaj mobilni operator je pouzdana.
P5- p15	15. Osjećam sigurnost prilikom korištenja usluga ovog mobilnog operatera.
K1- p16	16. Ovaj mobilni operater pravovremeno rješava zahtjeve korisnika.
K2- p17	17. Zaposlenici ovog mobilnog operatora su uvijek spremni pomoći.
K3- p18	18. Ovaj mobilni operater brzo rješava pritužbe korisnika.
K4- p19	19. Ovaj mobilni operator kada uvodi nove usluge osigurava pravovremene informacije.
K5- p20	20. Ovaj mobilni operater je dosljedan u pružanju kvalitetnih usluga.
K6- p21	21. Ovaj mobilni operater pruža uslugu govornih poziva, slanja SMS i MMS poruka i pristupa internet uslugama bez greške.
C1- p22	22. Lako razumljiva politika cijena usluga ovog mobilnog operatora.
C2- p23	23. Politika cijene usluga ovog mobilnog operatora je atraktivna.
C3- p24	24. Cijena uspostave govornog poziva koje nudi ovaj mobilni operator je razumna.
C4- p25	25. Ovaj mobilni operater nudi fleksibilne cijene za razne usluge koje zadovoljavaju moje potrebe.
C5- p26	26. Nastavit ću koristiti usluge ovog mobilnog operatora, osim ako je cijena znatno viša za istu uslugu.
I1- p27	27. Imidž ovog mobilnog operatora je popularan i dobro cijenjen od strane mnogih.
I2- p28	28. Imam dobar osjećaj o socijalnoj i društvenoj odgovornosti koje nudi ovaj mobilni operater.
I3- p29	29. Korisnici vole biti povezani s imidžem ovog mobilnog operatora.



Slika 2. AMOS dijagram modela utjecaja kvalitete usluga i zadovoljstva korisnika na lojalnost korisnika usluga s izračunatim parametrima

TABLICA V. KORELACIJE IZMEĐU FAKTORA MODELA

	Kvalitet	Cijena	Imidž	Zadovoljstvo	Povjerenje	Lojalnost
Kvaliteta	1					
Cijena	0,637	1				
Imidž	0,707	0,612	1			
Zadovoljstvo	0,755	0,688	0,755	1		
Povjerenje	0,648	0,664	0,700	0,640	1	
Lojalnost	0,614	0,584	0,631	0,739	0,673	1

TABLICA VI. STANDARDIZIRANI KOEFICIJENTI

Relacija			Procjena parametara	Standardna greška	t -vrijednost	p -vrijednost
Zadovoljstvo	<---	Kvaliteta	0,324	0,044	7,156	***
Povjerenje	<---	Kvaliteta	0,161	0,046	3,509	***
Zadovoljstvo	<---	Imidž	0,227	0,036	6,266	***
Povjerenje	<---	Imidž	0,342	0,046	7,374	***
Zadovoljstvo	<---	Cijena	0,314	0,044	7,349	***
Povjerenje	<---	Cijena	0,290	0,039	7,420	***
Lojalnost	<---	Zadovoljstvo	0,528	0,048	11,076	***
Lojalnost	<---	Povjerenje	0,343	0,044	7,738	***

Istraživanje zadovoljstva korisnika multidimenzionalnih usluga na tržištu telekomunikacija Bosne i Hercegovine

Mr.sc. Nermin Karahmetović
BH Telecom d.d. Sarajevo - RD Tuzla, Tuzla, Bosna i Hercegovina
nermin.karahmetovic@bih.net.ba

Sažetak - Istraživanje percepcije zadovoljstva korisnika multidimenzionalnih usluga na tržištu telekomunikacija, između ostalog, ima za cilj izdvajanje dimenzija koje imaju najveći uticaj na zadovoljstvo korisnika. Istraživanje je usmjereno na asembliranu uslugu korisnika mobilnih mreža koja obuhvata: govorne usluge, mobilne podatkovne usluge te pripadajuću terminalnu opremu neophodnu za optimalno korištenje date usluge. Faktorskom analizom rezultata istraživanja utvrdit će se ključne determinante zadovoljstva, kao i ključne karakteristike koje imaju utjecaj na kupovinu i korištenje asemblirane usluge. Dobiveni rezultati istraživanja zadovoljstva korisnika asemblirane usluge poslužit će za kreiranje mape važnosti/performanse.

Telekom operatori kao davaoci usluga moraju stalno provjeravati i uspoređivati stvarno stanje zadovoljstva korisnika sa onim kako ga doživljavaju uposlenici kompanije. U tom kontekstu istraživanje će, dodatno, biti usmjereno i na utvrđivanje postojanja jaza u percepciji zadovoljstva korisnika, sa jedne strane, i percepcije tog zadovoljstva od strane menadžmenta i zaposlenih u prodaji. Dobiveni rezultati biti će testirani primjenom ANOVA testa, a sve kako bi se utvrdilo postojanje navedenih jazova i značajnijih odstupanja.

Rezultat istraživanja, analize i izdvajanja ključnih determinanti i karakteristika zadovoljstva korisnika asemblirane usluge, jeste kreiranje i poduzimanje konkretnih aktivnosti i mjera za stalno povećanje zadovoljstva a samim time i percipiranog kvaliteta na strani korisnika asemblirane usluge.

I. UVOD

Poznato je da je zadovoljstvo korisnika najbolji preduslov buduće prodaje, i to preko pokazatelja lojalnosti i spremnosti širenja pozitivne usmene predaje. Međutim, cjelokupno zadovoljstvo korisnika uslovljeno je time u kojoj mjeri su ispunjena njihova očekivanja sa aspekta primljene vrijednosti. Zadovoljavanje potreba i očekivanja korisnika podrazumjeva obezbjeđenje onoga šta je potrebno i kada je potrebno (prava kombinacija usluge u pravo vrijeme). Ovdje je važno napomenuti da se ni u kom slučaju ne smije pretpostavljati potreba korisnika. Ozbiljne kompanije ulažu ogromne napore i sredstva da bi spoznale i istražile potrebe i želje korisnika.

U uslovima globalne konkurencije, a telekomunikacijske usluge možemo posmatrati u takvim uslovima, telekom operatori ne mogu postići konkurentsku prednost samo na osnovu jeftinije usluge odnosno, da svoju konkurentsku prednost grade samo na nižim cijenama usluga [1].

Za stvaranje konkurentske prednosti, telekom operatori moraju određene (tražene) usluge pružiti prije i kvalitetnje u odnosu na konkurenciju kako bi poboljšali i zadržali svoju tržišnu poziciju, koju treba posmatrati kao prolaznu i promjenjivu kategoriju [2]. Uvažavajući prethodno rečeno, primarni cilj istraživanja a potom i upravljanja zadovoljstvom ne bi trebao biti samo ispunjenje očekivanja korisnika u pogledu konkretne multidimenzionalne usluge, već da se nadmaše očekivanja korisnika kako bi se stvorio dojam da konkurencija nije u stanju njegove potrebe ispuniti na bolji način [3].

Upravljanje kvalitetom primjenom ISO 9001 standarda, također, ima za cilj zadovoljnog korisnika. U prilog tome ide i činjenica da je već u ISO 9001:2000 standardu utemeljen pojam *Zadovoljstvo korisnika*, korisnikove potrebe, usmjerenost na korisnika, komunikacija sa korisnikom i sl. U novije vrijeme, pored već dobro poznatih pojmova poput *Quality of Services(QoS[1]),* sve češće u uporabi je pojam *Quality of Experience (QoE[2]).* To znači da se kvaliteti iskustva korisnika pridaje sve veća pažnja, te da se vrlo često poistovjećuje sa pojmom *Zadovoljstvo korisnika.*

Značajan broj radova koji analiziraju zadovoljstvo korisnika telekomunikacijskim uslugama baziran je prema jednoj, specifičnoj, usluzi. U tom kontekstu analiziraju se uticaji tih, pojedinačnih, usluga na zadovoljstvo korisnika ili na njihovo iskustvo korištenjem usluga (QoE) [4]. Međutim, u savremenom telekomunikacijskom okruženju tzv. *bundliranje/ asembliiranje* usluga postalo je standardno, bez obzira o kojem segmentu telekomunikacijskih usluga je riječ (bez obzira preko koje tehnologije se usluga pruža, jedna ili kombinacija različitih tehnologija).

U takvom okruženju davaoce telekomunikacijskih usluga itekako interesuje koje dimenzije i karakteristike usluga, trebaju prilagoditi ili bolje rečeno uskladiti sa željama, očekivanjima i značajnostima koje iste imaju za korisnike. Normalno je da svaka dimenzija ili pojedina karakteristika nema istovjetan uticaj na zadovoljstvo korisnika, te je cilj profiliranje i poboljšanje onih dimenzija i karakteristika koje imaju najveću važnost za korisnike [5].

U ovome radu polazi se upravo od prethodno pomenutih postavki. U cilju maksimiziranja zadovoljstva korisnika,

1. Kvaliteta određenog servisa odnosno usluge,
2. Kvaliteta sveukupnog iskustva kupca/korisnika sa određenim proizvodom ili uslugom.

potrebno je konstantno povećavati nivo percipiranog kvaliteta asemblirane usluge. Da bi se postigao taj cilj, aktivnosti su usmjerene u pravcu: upravljanja dimenzijama koje imaju najveći značaj za korisnika, poduzimanju konkretnih aktivnosti za poboljšanje stvarnog kvaliteta asemblirane usluge, ciljanih promotivnih i akcija oglašavanja, kako bi se djelovalo na uspostavljanje realnog nivoa očekivanja kod korisnika, te na taj način indirektno uticalo na povećanje zadovoljstva i nivoa percipiranog kvaliteta [6].

Kako bi se navedeni cilj adekvatno i sveobuhvatno realizirao, kroz različite faze internog i eksternog istraživanja prikupljeni su neophodni podaci, te su njihovom detaljnom analizom i obradom izvučeni određeni zaključci, te će biti predložene mjere za unapređenje zadovoljstva korisnika asemblirane usluge.

Nakon provedenog preliminarnog istraživanja, te analizom reklamacija, primjedbi i sugestija korisnika, detektovane su najučestalije dimenzije i determinante zadovoljstva odnosno nezadovoljstva, te su iste i korištene u provedenom istraživanju. U istraživanju je posmatrano zadovoljstvo korisnika sljedećim dimenzijama asemblirane usluge: govorna usluga, mobilna podatkovna usluga te terminalna oprema. Za potrebe analize zadovoljstva korisnika različitim dimenzijama i karakteristikama, asemblirana usluga je raščlanjena na 10 determinanti zadovoljstva, i to: kvalitet govorne usluge, cijena govorne usluge, prodajna i postprodajna podrška govornim uslugama, kvalitet usluge za prenos podataka, cijena mobilne podatkovne usluge, prodajna i postprodajna podrška mobilnim podatkovnim uslugama za prenos podataka, kvalitet terminalne opreme, cijene terminalne opreme, prodajna i postprodajna podrška prodatoj terminalnoj opremi, te ljubaznost, stručnost i profesionalnost zaposlenih u prodaji i podršci asemblirane usluge.
Utvrđivanje i analiza ključnih determinanti i karakteristika zadovoljstva korisnika asemblirane usluge omogućavaju kreiranje i poduzimanje konkretnih aktivnosti i mjera za stalno povećanje zadovoljstva, a samim time i percipiranog kvaliteta korisnika navedene usluge .

U cilju dobijanja validnih spoznaja o zadovoljstvu korisnika asemblirane usluge, davaoci usluga moraju kontinuirano uporeĎivati stvarno stanje zadovoljstva korisnika sa onim kako ga doživljavaju uposlenici kompanije. U vezi s tim, istraživanje će biti usmjereno i na utvrđivanje postojanja jaza u percepciji zadovoljstva korisnika usluge i percepciji tog zadovoljstva od strane menadžmenta i zaposlenih u prodaji asemblirane usluge. Prikupljeni podaci obradit će se ANOVA testom, a sve kako bi se utvrdilo postojanje navedenih jazova, odnosno statistički značajnih odstupanja.

II. METODOLOGIJA ISTRAŽIVANJA

U cilju boljeg razumjevanja i lakše interpretacije podataka koji će biti prikazani u nastavku rada, predstavljen je kratak pregled metodologije istraživanja.
Na tržištu telekomunikacija Bosne i Hercegovine trenutno se pruža veći broj multidimenzionalnih usluga, što je trend na telekomunikacionim tržištima širom svijeta, a sve to kako bi se korisnici vezali za telekom operatere, ujedno ostvarujući i za sebe određene pogodnosti. Jedna od multidimenzionalnih usluga koja je prisutna na telekomunikacijskom tržištu Bosne i Hercegovine je i asemblirana usluga kojom se korisnicima nudi govorna/podatkovna usluga, zajedno sa adekvatnom terminalnom opremom, bilo u vidu mobilnog telefonskog aparata ili druge terminalne opreme potrebne za optimalno korištenje date usluge.

Opšti cilj istraživanja je upravljanje zadovoljstvom korisnika asemblirane usluge, i stvaranje konkurentske prednosti telekom operatora u Bosni i Hercegovini.
Osnovna istraživačka hipoteza glasi: Upravljanje određenim, manjim, brojem ključnih dimenzija asemblirane usluge moguće je povećati zadovoljstvo korisnika.

Prema tome, istraživanje je fokusirano na: stepen zadovoljstva korisnika asemblirane usluge isporučenim kvalitetom, percepcija menadžmenta i zaposlenih po pitanju zadovoljstva korisnika asemblirane usluge. Također, jedan od ciljeva istraživanja je da se provođenjem ankete među korisnicima asemblirane usluge na području Tuzlanskog kantona istraže stavovi korisnika po pitanju njihovog zadovoljstva različitim dimenzijama i karakteristikama usluge, utjecaju pojednih dimenzija na sveukupno zadovoljstvo.

Za potrebe realizacije prikupljanja primarnih podataka provedeno je istraživanje na jednom dijelu telekomunikacijskog tržišta Bosne i Hercegovine (koje predstavlja dio homogenog skupa korisnika multidimenzionalnih usluga).
U okviru eksternog istraživanja (*eng. field research*) podaci su prikupljani primjenom direktnih kontakata i razgovora sa korisnicima multidimenzionalnih usluga metodom ispitivanja i strukturiranog posmatranja, te podataka prikupljenih anketiranjem korisnika.

Sastavni dio istraživanja predstavljat će anketa i intervju sa prodajnim osobljem i rukovodiocima prodaje na području RD Tuzla kao i menadžmentom BH Telecoma, a sve to kako bi se dobio uvid u njihove stavove o stepenu zadovoljstva korisnika multidimenzionalnih usluga, najčešćim negodovanjima korisnika i sl.
Ovo je potrebno kako bi se utvrdila eventualna odstupanja (jaz) između stvarnog i očekivanog zadovoljstva korisnika asemblirane usluge, na strani korisnika i na strani zaposlenih i menadžmenta. Obzirom da sveukupan doživljaj zadovoljstva korisnika ima više dimenzija, te da svaka dimenzija nema istovjetan uticaj i značaj kod korisnika, potrebno je istražiti vrijednosti i značaj svih dimenzija. Cilj ovakvog pristupa je efikasno upravljanje zadovoljstvom korisnika asemblirane usluge, odnosno upravljanje zadovoljstvom kod onih dimenzija za koje se ispostavi da imaju najveći uticaj na zadovoljstvo korisnika.

Za sevobuhvatan pristup istraživanju teme rada, u prvom dijelu Anketnog upitnika, asemblirana usluga je raščlanjena na četiri (4) dimenzije koje je čine, i to: govorna usluga, usluga prenosa podataka (SMS, MMS, GPRS) i terminalna oprema koja se prodaje u sklopu usluge (mobilni telefon). Svaka od ovih dimenzija na svoj način doprinosi sveukupnoj vrijednosti samog paketa.
Svaka dimenzija, odnosno zadovoljstvo korisnika njenom vrijednošću, istražiti će se kroz: stvarni kvalitet usluge, cijenu, te kvalitet prodajne i postprodajne podrške. Za svaku navedenu dimenziju usluge istražiti će se usklaĎenost njene značajnosti kod korisnika sa stepenom njegovog zadovoljstva.
U istraživačkom projektu obuhvaćeni su svi korisnici asemblirane usluge na području koje pokriva Regionalna direkcija BH Telecom-a Tuzla, u periodu od 01.01.2007. do 31.12.2010 godine.

Ispitanik je fizičko ili pravno lice – korisnik asemblirane usluge sa područja koje pokriva Regionalna direkcija Tuzla. Odabir ispitanika za uzorak je slučajan i izvršen je korištenjem programa, za ovu potrebu napravljenog, u softverskom programu Excell. Samo anketiranje provedeno je primjenom kombinovane metode anketiranja, unutar Regionalne direkcije Tuzla. Svim korisnicima izabranim u uzorak poštom je poslat

Anketni upitnik, a tome je prethodio telefonski kontakt i priprema korisnika za prijem i popunjavanje anketnog upitnika. Korisnici su telefonom kontaktirani i nakon slanja anketnih upitnika, a sve u cilju povećanja broja odgovora na anketu.
Na ovakav način anketiranja i odabira ispitanika u uzorak ostvareno je osnovno načelo da svaki korisnik asemblirane usluge ima jednaku mogućnost da bude izabran.

Korisnicima je u Anketnom upitniku ponuđeno da ocijene svoje zadovoljstvo/nezadovoljstvo na rangirajućoj skali izreženoj ocjenama od 1 do 5 (standardna MOS skala) tj. da svoje zadovoljstvo izraze na skali od „Izrazito nezadovoljan“ do „Izrazito zadovoljan“. Isto tako, na pitanja o ocjeni važnosti/značajnosti koju svaka od dimenzija i karakteristika ima na njihovo sveukupno zadovoljstvo, korisnici su se izjašnjavali na odgovore od 1 do 5, odnosno od „Izrazito nevažno“ do „Izrazito važno“.

Uzorak obuhvata 200 korisnika. Od tog broja, 159 korisnika je dostavilo ispravno popunjene Anketne upitnike, 32 korisnika nije nikako dostavilo Upitnike, a njih 9 je dostavilo neispravno popunjene Upitnike. Ukupan odziv na Anketu iznosi 79,5%, što je ocijenjeno kao izrazito prihvatljiv nivo.

III. REZULTATI ISTRAŽIVANJA

Prvo je testirana značajnost razlike stavova zadovoljstva korisnika usluge prezentiranim determinantama. Osnovni zadatak ove analize je da se na osnovu određenog broja originalnih promjenjivih, koje međusobno koreliraju, izdvoji manji broj međusobno nezavisnih latentnih promjenjivih kojima se može objasniti korelacija između originalnih promjenjivih.
Između većeg broja pristupa koji se mogu naći u literaturi, ovdje smo se opredjeliti za postupak faktorske analize koji se realizuje kroz tri koraka [7]:

1. ocjenjivanje faktorskog modela,
2. određivanje broja faktora,
3. interpretacija faktora.

Faktorskom analizom tretirat će se sledeće determinante zadovoljstva:

VAR01 = Kvalitet govorne usluge,
VAR02 = Cijene govorne usluge,
VAR03 = Prodajna i postprodajna podrška govornim uslugama,
VAR04 = Kvalitet mobilnih podatkovnih usluga,
VAR05 = Cijene usluga za prenos podataka,
VAR06 = Prodajna i postprodajna podrška mobilnim podatkovnim uslugama,
VAR07 = Kvalitet ponude mobilnih uređaja (terminalne op.),
VAR08 = Cijene terminalne opreme,
VAR09 = Prodajna i postprodajna podrška prodatoj terminalnoj opremi,
VAR10 = Ljubaznost, stručnost i profesionalnost zaposlenika u prodaji i podršci asemblirane usluge.

Prema rezultatima izdvajanja faktora (TABLICA I.) utvrđeno je da prva dva faktora imaju kritične vrijednosti veće od 1, te da se sa ova dva faktora objašnjava 58,799 varijanse. Kao metod izdvajanja faktora primjenila se procedura glavnih komponenti, i kao kriterij izbora broja faktora izabrane su kritične vrijednosti pridružene faktorima koje su veće od 1.

Objašnjenje varijabiliteta: faktor kvaliteta (determinante koje određuju pojedine dimenzije kvaliteta usluge) 48,526% ukupnog varijabliteta posmatranih varijabli (1-10), a faktor cijene (determinante koje određuju poojedine aspekte cijene usluge) dodatnih 10,274% varijabliteta.

TABLICA I. FAKTORI ZADOVOLJSTVA KORISNIKA

	Početne vrijednosti faktora		
	Ukupno	Varijansa (%)	Kumulativno (%)
1	4,853	48,526	48,526
2	1,027	10,274	58,799
3	0,851	8,513	67,312
4	0,747	7,473	74,785
5	0,647	6,467	81,252
6	0,557	5,575	86,827
7	0,462	4,621	91,448
8	0,363	3,626	95,073
9	0,276	2,763	97,837
10	0,216	2,163	100,000

Rezultati ortogonalne rotacije faktora (TABLICA II.) pokazuju da prvi faktor – kojeg smo nazvali faktor kvaliteta, između ostalog, čini zadovoljstvo: podrškom uslugama za prenos podataka, kvalitetom usluga za prenos podataka, kvalitetom govorne usluge. Faktor cijena čini zadovoljstvo cijenama mobilnih aparata i cijenama usluga za prenos podataka.

Rezultati iz tablice 2. pokazuju da f*aktor* kvaliteta čine kriteriji: varijabla 6 (k*valitet podrške uslugama za prenos podataka*) sa težinskim koeficijentom 0,770, varijabla 4 (*kvalitet usluga za prenos podataka*) sa težinskim koeficijentom 0,753, varijabla 1 (*kvalitet govorne usluge*) sa težinskim koeficijentom 0,703, itd.

Faktor cijene čine kriteriji: varijabla 8 (c*ijene mobilnih aparata*) sa težinskim koeficijentom 0,837, varijabla 5 (c*ijene usluga za prenos podataka*) sa težinskim koeficijentom 0,752, itd.

Dalje, potrebno je navedene težinske koeficijente kvadrirati kako bi se utvrdile proporcije varijanse određene promjenjive (određenog kriterija) i utvrdilo dejstvo datog faktora. U ovom slučaju, se može zaključiti da faktor kvaliteta objašnjava 59,29% varijanse promjenjive broj 6 (kvalitet prodajne i postprodajne podrške uslugama za prenos podataka), $0,77^2$ = 0,5929 = 59,29%, itd.

TABLICA II. MATRICA KOMPONENTI ZADOVOLJSTVA

	Faktori	
	1	2
Varijabla 06	0,770	0,304
Varijabla 04	0,753	0,370
Varijabla 01	0,703	
Varijabla 09	0,696	0,333
Varijabla 10	0,642	
Varijabla 03	0,461	0,444
Varijabla 08		0,837
Varijabla 05	0,356	0,752
Varijabla 07	0,406	0,689
Varijabla 02		0,665

Faktorskom analizom analiziran je uticaj i značaj pojedinih dimenzija i karakteristika kod donošenja odluke o kupovini i korištenju asemblirane usluge.
Prema rezultatima analize (TABLICA III.) prva dva faktora imaju kritične vrijednosti veće od 1, te se sa ova dva faktora objašnjava 59,903% varijanse.

TABLICA III. FAKTORI ZNAČAJNOSTI KOD ODLUČIVANJA

Faktori	Početne vrijednosti faktora		
	Ukupno	Varijansa (%)	Kumulativno (%)
1	4,890	48,900	48,900
2	1,100	11,002	59,903
3	0,981	9,811	69,713
4	0,612	6,123	75,837
5	0,563	5,632	81,469
6	0,510	5,099	86,568
7	0,475	4,749	91,317
8	0,372	3,723	95,040
9	0,278	2,783	97,823
10	0,218	2,177	100,000

Prema podacima u ovoj tablici vidi se da prva dva faktora imaju kritične vrijednosti veće od 1, te se sa ova dva faktora objašnjava 59,903% varijanse.

Objašnjenje varijabliteta: Prvi faktor objašnjava 48,900% ukupnog varijabliteta posmatranih varijabli (1-10), a drugi faktor dodatnih 11,002% varijabliteta.

Rezultati ortogonalne rotacije faktora pokazuju da faktor kvaliteta, između ostalog, čini važnost kojeg za korisnika ima: kvalitet govorne usluge, prodajna i postprodajna podrška govornim uslugama, i kvalitet mobilnih podatkovnih usluga. Faktor cijene i podrške, čini važnost kojeg za korisnika ima: cijena terminalne opreme, prodajna i postprodajna podrška vezana za terminalnu opremu.
Rezultati ortogonalne rotacije faktora su predstavljeni u sledećoj tablici.

TABLICA IV. MATRICA KOMPONENTI ZNAČAJNOSTI

	Faktori	
	1	2
Varijabla 02	0,797	
Varijabla 04	0,759	
Varijabla 05	0,725	0,377
Varijabla 03	0,696	0,333
Varijabla 07	0,557	0,477
Varijabla 09		0,797
Varijabla 10		0,726
Varijabla 11		0,708
Varijabla 06	0,414	0,668
Varijabla 08	0,529	0,539

Rezultati pokazuju da *faktor kvaliteta* čine: značaj/važnost koju za korisnika ima kvalitet govorne usluge, sa težinskim koeficijentom 0,797, značaj/važnost koju za korisnika ima prodajna i postprodajna podrška govornim uslugama, sa težinskim koeficijentom 0,759, značaj/važnost koju za korisnika ima kvalitet mobilnih podatkovnih usluga, sa težinskim koeficijentom 0,725, itd.
Faktor cijene i podrške čine: značaj/važnost koju za korisnika ima cijena terminalne opreme, sa težinskim koeficijentom 0,797, značaj/važnost koju za korisnika ima prodajna i postprodajna podrška vezana za terminalnu opremu, sa težinskim koeficijentom 0,726, itd.

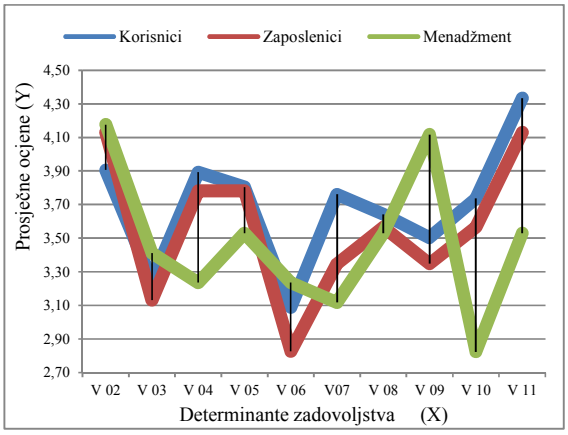
Primjenom ANOVA testa, testirana je i značajnost razlike u percepciji prodajnog osoblja i menadžmenta o stavovima i zadovoljstvu korisnika (Slika 1.) .

Analiza je imala za cilj da ukaže na odstupanja u stavovima korisnika po pitanju njihovog zadovoljstva, te percepcije tih stavova od strane prodajnog osoblja i menadžementa kompanije. Rezultati pokazuju da po pitanju percepcije zadovoljstva korisnika pojedinim dimenzijama asemblirane usluge između posmatrane tri grupe postoje značajne razlike.

Te razlike su posebno izražene kod percepcije podrške govornim uslugama, podrške uslugama za prenos podataka, cijene terminalne opreme, podrške prodatoj terminalnoj opremi, te ljubaznosti, stručnosti i profesionalnosti zaposlenika u prodaji i podrški. Utvrđeno je da zaposlenici u prodaji imaju bolju percepciju o stavovima korisnika po dimenzijama zadovoljstva iskazane kroz navedene varijable u odnosu na menadžment kompanije.

U cilju upotpunjavanja slike o percepciji zadovoljstva korisnika pojedinim dimenzijama i karakteristikama asemblirane usluge na slici 1. grafički su prikazane prosječne ocjene ispitanika u kategorijama: korisnici, zaposleni (prodajno osoblje) i menadžment. Iz grafikona jasno su vidljiva odstupanja prosječnih vrijednosti po pojedinim dimenzijama i karakteristikama usluge. Na slici 1. jasno je vidljivo da zaposlenici u prodaji imaju bolju percepciju o stavovima korisnika po dimenzijama zadovoljstva iskazane kroz varijable 4, 7, 9, 10 i 11 u odnosu na menadžment kompanije.

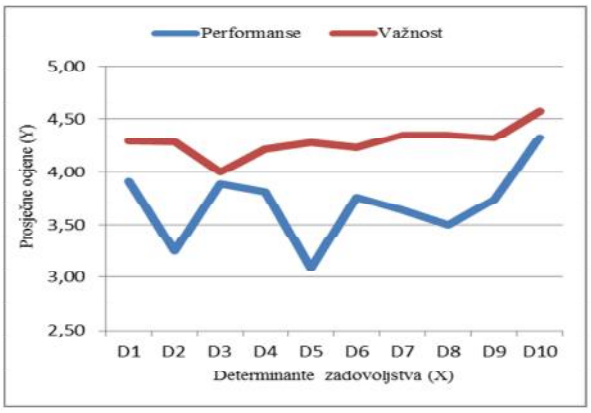
Također, na slici je moguće na vrlo jasan način vidjeti odstupanja u percepciji zadovoljstva korisnika između menadžmenta i prodajnog osoblja i po ostalim posmatranim dimenzijama zadovoljstva.



Slika 1. Zadovoljstvo i percepcije zadovoljstva između različitih kategorija ispitanika

Dodatno, kako je i planirano, rezultati istraživanja omogućavaju da se korištenjem mape važnost / performanse ukaže na pravce djelovanja u smislu poboljšanja onih

determinanti zadovoljstva kod kojih postoje najveća odstupanja u odnosu na važnost/značajnost koju iste imaju kod korisnika asemblirane usluge. Slika 2. predstavlja grafički prikaz rezultata istraživanja zadovoljstva korisnika pojedinim determinantama kao i važnosti koje svaka od tih determinanti ima za korisnika i njegovo korištenje usluge.



Slika 2. Mapa Važnost / Performanse

Kako je i vidljivo performanse (iskazani stavovi) zadovoljstva korisnika pojedinim determinantama u značajnoj mjeri odstupaju od važnosti koju svaka od tih determinanti ima za korisnike asemblirane usluge.

Navedena odstupanja posebno su izražena kod sledećih determinanti:

- Cijene govorne usluge,
- Cijene mobilnih podatkovnih usluga,
- Cijene treminalne opreme,
- Ponuda terminalne opreme,
- Prodajna i postprodajna podrška prodatim uređajima.

Vidljivo je da korisnici (prema dodjeljenim važnostima) posebno važnim doživljavaju ljubaznost, stručnost i profesionalnost zaposlenih te kvalitet ponude terminalne opreme (asortiman ponude mobilnih uređaja). Također, vidljivo je, da je po pojedinim determinantama zadovoljstva telekom operator dostigao performanse koje su u skladu ili su približne važnostima koje one imaju za korisnike. Primjer tih determinanti su:

- Prodajna i postprodajna podrška govornim uslugama,
- Ljubaznost, stručnost i profesionalnost zaposlenika u prodaji i podršci,
- Kvalitet govorne usluge, i
- Kvalitet mobilnih podatkovnih usluga.

U skladu sa prethodno rečenim, poželjno je aktivnosti telekom operatora usmjeriti prije svega ka poboljšanju onih determinanti zadovoljstva kod kojih je percepcija zadovoljstva na znatno nižem nivou u odnosu na važnost koje iste imaju za korisnike. Istovremeno, potrebno je osigurati zadržavanje performansi determinanti koje su jednake ili približne njihovim važnostima kod korisnika.

IV. ZAKLJUČAK

Rezultati analize ključnih dimenzija i karakteristika asemblirane usluge koje utiču na zadovoljstvo korisnika (primjenom faktorske analize) pokazuju da je aspekt kvaliteta usluge dominantniji u odnosu na aspekt cijene usluge.

Upravljanje kvalitetom asemblirane usluge treba da obezbjedi brojne koristi, a njihov uticaj se, prije svega, treba posmatrati kroz uticaj kvaliteta na zadovoljstvo korisnika. Različite dimenzije i karakteristike asemblirane usluge imaju različit značaj i uticaj na zadovoljstvo korisnika. Upravljanje kvalitetom, za korisnika, ključnih dimenzija može povećati nivo zadovoljstva korisnika.

U okviru aspekta kvaliteta usluge najznačajniji uticaj imaju sljedeće determinante: prodajna i postprodajna podrška uslugama za prenos podataka, kvalitet usluge za prenos podataka, kvalitet govorne usluge, prodajna i postprodajna podrška prodatoj terminalnoj opremi.

Kako bi se povećao nivo zadovoljstva korisnika, poboljšanju kvaliteta treba prije svega težiti kod ovih determinanti. To se može ostvariti kreiranjem dodatnih vrijednosti unutar pojedinih dimenzija asemblirane usluge.

U okviru aspekta cijene usluge najznačajniji uticaj imaju slijedeće determinante: cijena terminalne opreme, cijena mobilne podatkovne usluge. Obzirom da cijene ove dvije dimenzije asemblirane usluge imaju značajan uticaj na zadovoljstvo korisnika, potrebno je konstantno voditi računa da se cjenovnom politikom potiče obim korištenja, prije svega mobilne podatkovne usluge.

Analiza značajnosti uticaja pojedinih dimenzija i karakteristika na odluku o kupovini i korištenju asemblirane usluge pokazuje da je moguće izdvojiti dva faktora koja imaju najveći uticaj kod korisnika.

Prvi je faktor kvaliteta u okviru kojeg najveći uticaj na odluku o kupovini i korištenju usluge imaju slijedeće karakteristike kvaliteta: kvalitet govorne usluge, podrška govornim uslugama, kvalitet mobilnih podatkovnih usluga.

U okviru drugog faktora ključne karakteristike koje utiču na odluku o kupovini i korištenju usluge su: cijena terminalne opreme, prodajna i postprodajna podrška terminalnoj opremi, ljubaznost, stručnost i profesionalnost zaposlenih.

Komparacija dobivenih rezultata o zadovoljstvu korisnika pojedinim dimenzijama asemblirane usluge i percepcije tog istog zadovoljstva od strane zaposlenih i menadžmenta pokazuje da realniju percepciju o zadovoljstvu korisnika imaju zaposleni u prodaji i podršci.

Rezultati ANOVA testa pokazuju da je percepcija zaposlenih u prodaji i podršci o zadovoljstvu korisnika dimenzijama: prodajne i postprodajne podrške govornim uslugama, prodajne i postprodajne podrške mobilnim podatkovnim uslugama, cijene terminalne opreme, prodajne i postprodajne podrške prodatoj terminalnoj opremi, ljubaznost, stručnost i profesionalnost zaposlenika u prodaji i podršci asemblirane usluge. Na osnovu ovoga podatka potrebno je u internim procesima obezbjediti nesmetani protok informacija o zadovoljstvu korisnika od zaposlenih u prodaji i podršci ka menadžmentu, a sve kako bi donosioci odluka imali što realniju sliku o zadovoljstvu korisnika pojedinim dimenzijama asemblirane usluge.

Svakako da menadžment mora pri planiranju budućih aktivnosti upravljanja kvalitetom asembliranih usluga, kao i upravljanja odnosima s korisnicima usluga, te kod planiranja prodajnih aktivnosti imati u vidu navedene determinante zadovoljstva korisnika i nihovog uticaja na odluku o kupovini i korištenju telekomunikacionih usluga.

LITERATURA

[1] Naylor G., Frank K.E., The effect of price bundling on consumer perceptions of value, Journal of Services Marketing, vol. 15., no. 4., MCB UP Ltd., 2001.

[2] Sajko H., Utjecaj mobilnih komunikacijskih usluga na stjecanje konkurentske prednosti, Doktorska disertacija, Ekonomski fakultet, Zagreb, 2008.

[3] Ferguson R., Brohaugh B., Telecom's search for the ultimate customer loyalty platform, Journal of Consumer Marketing, vol. 25., no. 5., Emerald Group Publishing Ltd., 2008.

[4] Marušić S., Marketing podatkovnih usluga mobilne komunikacije, Tržište, Vol. XVIII, br.1/2, UDK: 658.8:654.165, pp. 41-52, Zagreb, 2006.

[5] Cassab H., MacLachlan L. Douglas, A consumer-based view of multi-channel service, Journal of service Management, vol. 20., no. 1., Emerald Group Publishing, 2009.

[6] Bergman B., Klefsjö B., Quality from Customer Needs to Customer Satisfaction, 2nd ed., Studentlitteratur, Lund 2003.

[7] Zahirović S., Multivarijaciona analiza, Osnove teorije i primjene u marketinškim istraživanjima, Infograf d.o.o. Tuzla, pp. 168-174, Tuzla 2005.

INNOVATIONS AND BUSINESS IMPROVEMENTS

A blend of Semi-Open Innovation & Open Sourceness: accelerating innovation within a large Telecommunications company

Marko Bervanakis, Saša Dešić
Ericsson Nikola Tesla d.d.
Research and Development center, Zagreb Croatia
marko.bervanakis@ericsson.com

Abstract – Due to intense competition and time-to-market pressures, Telecommunications companies across the world are fostering and stimulating Research and Development (R&D) organizations to 'open-up' to various degrees. This is a response to both an immediate 'need' and perhaps due to the fact that the well proven open innovation paradigm has shown to be an effective strategy. Naturally, the degree of this 'openness' may vary from one firm to the next. Additionally, this open paradigm may not be necessarily confined only to the R&D (traditional innovation drivers) parts of the business. In this paper, for example we witness that its influence can have far more reaching and valuable impacts. Subsequently, the paper outlines how a 'Semi-Open paradigm' approach can be an essential ingredient in *accelerating* innovation capability growth - company wide. We present a firms recipe for success, which includes several novel innovation process methods and tools to capture and measure the valuable key innovation aspects the firm has identified. We find that one of the key culture aspects of this approach involves both a potpourri blend of top-down strategic alignment (fostering semi-open innovation) and combining pure (bottom-up) creative enthusiasm (using open-source software) via its people.

I. INTRODUCTION

Today, open innovation has become a very popular and widely used term, to the point that in many cases it has become mis-used or even *flippant*. For example due to the binary approach (or oversimplification) around this term, by a significant amount of research in this area [1], we feel there is an important *gap* that needs to be filled and requires more support. Namely, the inclusion and existence of *sub-levels* for a *third state*, when a firm is neither completely open or closed, but more say, in a *semi-open innovation* [2] (dynamic) state.

Open innovation *per se*, has been characterized as an enabler to bring down boundaries between firms and other value adding external stakeholders. Hence, allowing the flow of ideas, innovations and problem solving into a wider and more diverse resource pool. This, in terms of innovation literature - maps to an essential innovation characteristic (or DNA) for innovative people and organizations alike, namely the ability to *network* with a more diverse cohort or the creation of the *Medici effect* [3].

The development of this paper in turn, is bourne from two perspectives and ambitions. On the one hand, to help explore how the *semi-open-approach* (controlled open innovation & open source tools) has allowed a large Telecommunications firm to add that much needed ('need for') innovative premium to their portfolio. With the backdrop that the firm had in place a significant and formal long-term (over thirteen years) commitment to innovations *per se*. And on the other hand, to help identify the additional key 'open-people' that have combined with the available 'open-technologies' to help facilitate this new found success within this Telecommunications [2] based firm.

As we will see in this work, the success of the innovation management system within the firm in question often is the result of a combination of many factors, like structural changes (top-down leadership) and creative people from the lowest levels (bottom-up). These two opposing (management-workers) forces and paths seem to intersect at some mid-point (equilibrium) to *burst out* with innovative results that are much needed for the firm to be both successful and competitive. Additionally, we will reflect on some common open innovation *pitfalls* and how this company has managed to keep these common negative impacts at bay [1].

In turn, this paper details the firms novel approach to address these concerns (avoiding firm boundary syndrome), which could be applied to various degree within other Telecommunications firms or other industries. Additionally, this paper subsequently aims to advance the branch of research that examines what strategies and groups of people need to co-exist in a large firm to help make the *open-paradigm* a success.

Firstly, we begin by outlining the company's historical innovation management framework to have a firm baseline reference. Secondly, we present the current transformed innovation management system (process) and highlight some of the key aspects' (e.g. including semi-open innovation) that have contributed to its new strong innovation capability status. Subsequently, in this paper, leveraging our results, we challenge the notion that *openness* is more important for small firms and that large firms struggle to be flexible and responsive enough to take in the full advantages of this new open-approach [2].

Subsequently, this paper confirms the assertion from at least one previous study in this area [2], that firms with a very high innovation absorptive capacity (i.e. typically large firms) are able to benefit from external knowledge and remain semi-open. Where semi-openness is defined as a firm's ability to use external (innovative) knowledge and resources, but which are not crucial for their innovation management system. Consequently, in this study we extend this assertion by illustrating how this can in fact *accelerate* and *enhance* a firms' innovation capabilities.

Furthermore – within the context of this paper, we also want to illustrate that there may exist in fact different degrees (or *shades*) of *semi-openness* in large firms (i.e. a step-wise level approach). As opposed to only three distinct categories (i.e. Open, Semi-Open & Closed). That is to say, that it's quite valid and that these *semi-openness shades* (or levels) are a natural progression over time. Which are structured (safe framework) and can themselves bring immense value (i.e. vs. going fully open in *one giant leap*). We feel that past literature in this area has been oversimplified (i.e. suffers from *binary syndrome*) and does not reflect the reality on the ground. For example, we support the argumentation from Lakhani & Tushman, 2012 [2] that organizational design literature must move beyond debates between just open vs. closed boundaries and instead embrace the notion of more dynamic organizational (semi-open) interfaces. Hence, the use of the term *semi-open* innovation and its various *shades* (levels) within our paper.

II. IMPLEMENTING AN OPEN PARADIGM

A. *Firms Background: IMS – Baseline Reference*

The firm in question was observed and studied for the period from January 2010 until the end of December 2012. Note, the firm has had a formally defined innovation management system since the year 2000. When we say formally, we mean that it included an innovation process, idea capturing tool, organization to support the idea evaluations, a rule book for awarding and a basic implementation mechanism for accepted ideas.

The innovation management process in the observed period (2010-12) is summarized in the following *IMS Metrics* table tool within Table 1 (*see* Appendix). From this table, we can see that the firm has gradually grown and developed over this three (3) year observation period in an overall positive fashion. The IMS Metrics table tool employs the use of thirteen (13) key factors that were initially suggested and then measured over time by the firm's innovation champion's forum (i.e. key innovation drivers from different verticals) on a quarterly basis. The *IMS metrics* table was established by the forum based on both the internal KPIs and on other factors that were considered important for the firm's innovation management system.

Furthermore, we present this table as a possible novel approach for other firms to follow as a framework for their respective innovation management systems.

Levels of Semi-Open Innovativeness

From the open approach perspective we can see items 10 & 11 within the IMS metrics table (*see* Table 1) depicting a gradual growth year on year. Hence, the need for the term *shades* (or levels) of semi-open innovativeness'. Initially, for example the institutional collaboration between the firm and the local university was on a smaller scale (circa 20 students plus 5 mentors). So, this could be termed the *base-low* level for the firms (first steps towards) semi-open innovativeness. Over time, the firm expanded its collaboration to what we term both a *medium* and *high* level of semi-open innovativeness.

R&D base scope & fertile terrain

Initially, innovation was reserved to the confines of the R&D organization. This was quite convenient, easy to manage and a simplistic paradigm. In essence, it involved a committee of innovation drivers from different R&D (organizational) verticals that reported to a key innovation R&D facilitator. This person than reported to a higher innovation steering (leadership) team for further support as required. Successes were heavily dependent on this key facilitator (bottle-neck) and often lead to cross company blocking barriers and frustration. For example, idea authors often felt that their valid ideas could not get a timely and satisfactory opportunity for further implementation. In contrast and perhaps to counter this negative effect, it was observed that some innovation drivers or authors were able to breach these barriers and work (horizontally) across organizational lines and bring important (valuable) synergies. In effect, by-passing this perceived bottle-neck in the current innovation process.

In terms of top level strategy and mandate, the innovation management system was on the table (refer to Table 1 – *Low Level* grading in 2010), but could be described as *hands-off* and *better left to R&D* to find the best way. In turn, this was reflected in the quality and nature of the submitted ideas within the innovation management system. Namely, being heavily dominated by the R&D side with a *sprinkle* of ideas from other departments within the firm from innovation enthusiasts. Note, an end to end idea funnel process was defined, but implementation of *good* (accepted) ideas was quite a challenge at times (significant accepted idea back-log).

Basic University co-operation

Since the year 2000, a steady stream of students (circa 20-30 yearly) formed the core of the firms Summer School program, with students primarily coming from the Universities of Zagreb, Split, and Rijeka (domestic). The aim every year was to get the students to provide and implement interesting innovative projects and ideas that had surfaced over the recent past within the firm. The companies Summer School program for senior students from local universities and a local prototype competition became one of the key enablers for some of these "good-ideas" to gain traction and be implemented in a cost effective way.

Basic Innovation Methods & Tools

In turn, when it came to the use of an innovation toolbox to help drive innovation – we saw that the basics

were in place, for example the use of a common idea box (i.e. web based, on line etc..) for capturing and sharing ideas with people both within the firm and even on the global level within the firms other parent companies. Here, in these early forming years we also witnessed the early signs and the *need* to *open-up* and search for other innovation paradigms or best practices outside the company by employees on all levels. It was driven bottom up from the idea authors towards the innovation drivers (or champions). This movement or feeling within the firm's internal innovation community was clearly dissatisfied with the *status quo*.

B. Current IMS

With the firms first two years (2010-2011) as an innovation management system baseline and backdrop (learning's from the forming years, prior to 2010), the innovation management forum of champions decided to implement several new concepts. In parallel, top management also saw the benefits of strong leadership focus (i.e. also a reaction to the dissatisfaction of the current process) and support for innovations per say. Subsequently, they decided to appoint (with mandate) an innovation manager on the total company level that served on the highest management level. This person was responsible to make the innovation management system work across the entire firm (work horizontally) with the help of all the other innovation champions from the firms different verticals (R&D, Marketing, Sales etc.)

These top-down organizational changes subsequently coincided with the bottom-up *opening-up* factor for the firm. Specifically, to the use of new innovation paradigms (tools) by the innovation champions, participating on further external innovation seminars and looking externally for best practices. Furthermore, the use of a structured idea-template or framework for innovative ideas submission within the firms idea-box (i.e. avoid idea submission *Ping-Pong*), bringing in Open-Source software tools (i.e. Perl, Google Web Technologies, GWT etc…) both by individuals (new employees/student workers) & supported by the ICT department. To compliment these changes that developed during 2010-2011 in the firms innovation culture, the innovation leadership team decided to introduce and make use of cross-functional teams within the firm (in 2012), called Innovation Circles (iCircles). This was a great luxury and strength – that was not typical in its associated companies around the world.

The formation of iCircles was organized with the philosophy that similar verticals (e.g. Back-End Operations & IT etc…) should merge within the firm that could create synergies and make common ideas of interest happen. Here – the use of the iCircles paradigm was an enabler to help reduce the existing and common problems (blockers-delays) faced by larger firms with strong organizational barriers [1]. Another element of the new innovation management system in 2012 was that the entire company was now expected to participate within the innovation process via KPIs (top-down management chain). So the prior expectation or barrier, that innovation was R&Ds sole responsibility was now removed. In turn, it was now expected from each and every employee. It subsequently became a clear KPI for all the firms' organizations and for each employee alike within their yearly performance targets.

Mature to High Level of Semi-Open Innovativeness

As previously mentioned within the baseline section of the firms' innovation management system, we now explore the close cooperation the company formed both locally and within the region (South-East Europe) with various University institutions. It's important to note that the firm had this collaboration for over ten (10) years (primarily local), but in 2012 it really was in a mature (higher level) state. With a significant expansion in terms of institutions and the number of students participating. This enabled the easy use and exploitation (taking in) of the latest and greatest *open-source* tools available at the time. One of the primary enablers for this cooperation was and still is the company's annual Summer School program for prototype implementation and execution. This relationship is further strengthened by the fact that the company has several serving professors and assistants-supervisors (PhD-Masters) that both work within the firm and directly on University campus. Furthermore, the two communities regularly share common interest lectures and presentations (IEEE) to help strengthen and foster this strong and fruitful relationship.

We use the term *Semi-Open* Innovation [2] to reflect the fact that the collaboration is well defined with a strong working framework (i.e. IPR agreements, privacy disclosures etc.). As opposed to a more generic (un-controlled) fully open innovation type relationship [1]. Again here the firm has found a mechanism to avoid common open innovation problems while working with external resource pools. Here we witnessed a healthy and vibrant flow of creative enthusiasm between both sides, which often lead to strong prototype products towards various stakeholders. In summary, reflecting on our observations and the results contained within Table 1 we can in confidence conclude that the semi-open innovativeness level can be regarded as being in a *mature shade* (*high level)* today. That has the following characteristics: controlled, safe (IPR safeguards in place), free flowing with ideas between the firm and several other external institutions.

Since 2012, a higher level and larger significant group of selected students participated (circa 80 students) within the Summer School program, with students coming from the Universities of Zagreb, Split, and Rijeka (domestic) and as far afield as Sarajevo, Tuzla (Bosnia & Herzegovina) and Budapest (Hungary). In short, the program facilitated one of the most important stages in the innovation management system, namely the *implementation* of good accepted ideas from the company and its employees. Subsequently, this has also included the launch of new innovative products (i.e. eHealth & mHealth systems & products) towards the market.

University Technology cases

In this section, we describe in more detail some of the actual technological cases which best illustrate the success of this *semi-open* innovation collaboration between the company and the university.

Case 1: Technology 1 - IMMUNES (Linux based – Telecoms IP NW Simulator)

In this case, a local presentation of the IMUNES tool [3] was given by a group of students from the *Faculty of Electrical Engineering* (FER, Zagreb) towards the R&D testing Integration & Verification (I&V). One of these I&V people in due course realized that this could be directly applied quite easily into the Mobile Core & Radio I&V activities. This tool over time (circa 3 years) has been further developed in collaboration with FER for different I&V product areas. Subsequently, this has enabled these different testing groups to strengthen and showcase there I&V activities as leaders in their respective fields. Importantly, this has been both across a local and international level.

Case 2: Technology 2 – eHealth (Secure Internet based electronic Health system)

Back in 2002 a small internal research project started in the *eHealth* arena. It wasn't until 2011 when the FER cohort improved on this prior work and produced an *eHealth* prototype. This work was then presented towards the companies' innovation stakeholders within the Summer School program framework. In turn, the Summer School program innovation management steering members took this on and further developed the concept for internal further use and development. In due course, the concept has turned into a final market product offering (i.e. eHealth system for the Croatian government).

Open source Tools

There is strong evidence to illustrate that the introduction of open source tools has had a significant and positive impact on the firms' innovation capability over the last few years. Again cultimating and bearing fruit specifically in the last two (2) years (2011-12) for the firm. Where several ideas had been implemented solely due to the existence of freely (latest) available open-source software tools. This was clearly visible and reflected via the company's innovation management system and its subsequent monetization of these open-sourced based ideas. Below, we provide evidence and cases that best illustrate these positive value adding cases. Note, we have two types of technology innovations – either from the company's innovative (pro-active, bottom-up, idea authors proactivity) contributors or from the organization itself (top-down approach, structural support).

Case 1: Technology 3 (FreeSwitch [6])

Technology Introduction: Triggered bottom-up, the ideas author showed immense pro-activeness to make this technology fit their daily use of purpose.

The author of this initial idea used this technology initially for their own I&V Test activities (i.e. increase owns daily efficiency) and once proven, then decided to take the next step and spread the value to the entire team via both the innovation management process and via local I&V management. The technology has enabled this particular organization to make significant test execution efficiency gains (Qualitative) and quantitative savings by either replacing some legacy (old-outdated-Analog based) tools or complimenting the existing I&V test-tool base.

Case 2: Technology 3 (Perl SW Development Kit)

Technology Introduction: Organizational from Innovation Champions & Drivers (top-down)

In this case an innovation management Champion in a particular organization identified that there was a clear trend that had been developing in the recent past with the use of Perl within the organization. To help spread the value of these Perl based programs (I&V script improvements) to a wider audience, it was decided to provide the team with a Perl based Software Developers Kit (SDK). This SDK was in turn purchased and made available to a wider audience, in turn, it could be used by more idea authors to make there code portable and reusable across the organization. Over time this has shown to be most beneficial in spreading the value of the ideas to a wider audience for a relatively small investment cost.

Case 3: Technology 4 (Google Web Toolkit)

Technology Introduction: by a regular contributor to innovation management process (pro-active)

The use of this technology has enabled a team within the company to come together to help them realize an attractive GUI based telecoms node viewer. Prior to this, the team was using low-level methods (i.e. Command Line Interface, CLI) to get the information they required. The efficiency gains here at present have shown not only quantitative daily efficiency gains but also qualitative gains (i.e. happy testing feeling). Making the I&V testing experience less frustrating for the easy daily mundane tasks.

Discussion: Key Innovation success factors

Some of the important key factors in this innovation metamorphosis can be summarized as follows: there was not one key area (innovation silver bullet) for the firm's innovation success, but in fact a combination of long-term commitment, effort and on-going progress in all thirteen (13) key areas. Here we observed that the firm continuously made strong efforts to measure and improve in all the innovation key areas it defined as important. This was driven top down (KPIs & awarding) by top management to middle management (Innovation Champions & drivers) down to its creative and enthusiastic people on the lower levels.

The mechanisms for measuring and tracking the innovation key factors were the company's innovation champion's forum that reflected on a bi-yearly basis. In essence, a self-assessment of the firms innovativeness. It is important to note here that some key areas where naturally taken as priorities early on (PM&Ts) in the initial set-up (kick-start) of the firms innovation management system and others at later dates once it had matured to a certain degree (like internal competitions).

It feels that the firm via its long-term innovation management system commitment had a strong base and was culturally in their DNA. Hence, this again is a strong and key element that we feel helped the firm over the last three years. In essence its work force was growing an idea bank [6] of ideas. Once combined with other key factors like semi-open-innovation and open-source factors, this innovation mosaic platform enabled the firm's innovation capability to *take-off*. This is reflected in Table 1, where we see these two factors achieving a High level by 2012. Likewise, it's important to note that items 2 & 3 within Table 1, which relate to Mandate, Leadership & secured innovation support people also achieved a strong level to help enable this to happen. In this case the firms CEO and her dedicated innovation steering forum were well in verse on how the firm was progressing in its innovation (awards) activities and aspects. This made for a strong positive and supportive top-down innovative force to emerge.

III. CONLUSION

We feel our paper contributes to a better understanding of how a blended and shaded open-approach combined with the right (dedicated) innovation (open inspired & focused) people can accelerate a firms ability to invigorate their innovativeness. Our research has uncovered that this approach, namely the use of a Semi-Open Innovation paradigm (on different levels) and the use and adoption of Open-Source tools have acted as enabling agents to make innovations happen within this firm. From our perspective, the last three years have been a highlight for the firm, as they have enabled the firm to finally reap the fruit from their long-term commitment to innovations.

For the Telecommunications firm we have studied, this has resulted in both qualitative (i.e. un-measurable, positive stakeholder feedback, new work assignments etc...) and quantitative benefits. Subsequently, a base quantitative figure can be illustrated on a yearly basis with ease, but more importantly the firm has developed this special (Innovation X-factor) aspect to its portfolio. Subsequently, leading to both internal and external stakeholders being immensely satisfied with the firms' performance and innovation capability.

It also confirms previous theories and studies in this area, namely that firms that open themselves-up (even partly/semi degree) can realize and gain immense value that goes beyond the end of year financial statement reports. Today, this firm enjoys the luxury of having open and different student thinkers all year round to enable the implementation of innovative ideas and prototypes. Hence, semi-open innovation at this firm feels more like a natural process of osmosis, no boundary talk, more like a permeable membrane that adds value in both directions.

REFERENCES

[1] RWTH-TIM, D. Luttgens, D. Antons, P. Pollok, and F. Piller, "Implementing Open Innovation Beyond the Pilot Stage: Barriers and Organizational Interventions", RWTH Aachen, Tech. & Innovation Management Group, October 2012.

[2] Andrés Barge-gil, "OPEN, SEMI-OPEN AND CLOSED INNOVATORS. TOWARDS AN EXPLANATION OF DEGREE OF OPENNESS", Imperial College London Business School, June 16 - 18, 2010.

[3] J.Dyer, H.Gregersen, C.M.Christensen, "The Innovators DNA: Mastering the five skills of disruptive innovators", Harvard Business Review & Press, **ISBN:** 1422134814, 2011

[4] Karim R. Lakhani & Michael L. Tushman, "Open Innovation and Organizational Boundaries: The Impact of Task Decomposition and Knowledge Distribution on the Locus of Innovation", Harvard Business School Working paper, January 2012.

[5] IMUNES Integrated Multiprotocol Network Emulator/Simulator, http://www.imunes.tel.fer.hr/

[6] FreeSwitch, http://www.freeswitch.org/

APENDIX

Table 1. A Telecommunications Firms - IMS Metrics Tool

Innovation Key Measures	DESCRIPTION	2010 (L-M-H) Level*	2011 (L-M-H) Level*	2012 (L-M-H) Level*
1. Define Objectives & Scope	Innovation objectives & scope need to be defined & established. The objectives should be clear & contribute to the overall business goals.	M	M	H
2. Mandate & Leadership "on-board"	Innovation mandate needs to be established with senior management & clarify what management sponsorship and support to expect.	L	M	H
3. Secure People, Funding & Time	Dedicated resources in terms of people, funding and time need to be estimated and secured in line with agreed objectives.	L	M	H
4. Define Process & Funnel	A systematic end-to-end process for innovation (i.e. scoping, ideation, screening, decision-making, business cases, prototyping, business building) needs to be established.	M	H	H

5. Identify Tools	A useful (commonly shared) innovation toolbox to drive innovation (e.g. methods, templates) needs to be established.	L	M	M
6. Define Roles & Competence Needs	Clear roles & responsibilities for the functions supporting the innovation initiative & related competence needs should be defined & established	L	M	H
7. Determine Incentives & Recognition	A system to give recognition & feedback to colleagues should be in place as well as a platform to communicate successes. Incentives can improve awareness & performance.	L	M	H
8. Promote an Innovative Culture	Plan activities to promote an innovative culture (e.g. internal openness, collaboration, participation, commitment, managed risk-taking) in your organization/unit.	L	M	H
9. Perform In sighting	Insights about markets, customers and users (e.g. systematic opportunity scanning, business intelligence, consumer research) need to be strengthened and based on the focus.	L	M	M
10. Collaboration with external institutions (University, colleges, other…)	Using University students (local international) to help implement ideas from the firms IMS (customer focused) Summer School Participation (Semi-Open Innovation)	L	M	H
11. Use of Open Source Tools	Using Open source SW to solve daily issues and challenges.	L	M	H
12. Success in (internal) co-operate innovation competitions	Internal cooperate competitions (prototype) both on a local and international level.	L	M	H
13. Patent submissions	The number of patent submissions made towards the parent firms regional patent office.	L	M	M

* (L = LOW, M = MEDIUM, H = HIGH)

Note: IMS Levels where measured on Quarterly basis by the firms innovation champions forum.

SEO in e-commerce: balancing between white and black hat methods

Sanja Duk, Dunja Bjelobrk, Mia Čarapina
Polytehnic of Zagreb/Vrbik 8, 10000 Zagreb, Croatia
sduk@tvz.hr, dbjelobrk@tvz.hr, mcarapina@tvz.hr

Abstract - Search engines are constantly improving the algorithms to bring the most relevant search results to users. Search engine optimization evolves with them. This research is based on balancing between good practices in search engine optimization and making mistakes while using particular optimization method which causes that web page is considered as less worthy or even spammy. New methods have been developed to avoid being declared as black hat and degradation by new algorithms. It has been shown which of those methods are being used in practice by several e-commerce in Croatia and which are being overlooked.

I. INTRODUCTION

Search engine queries create over 80% of all traffic on the Internet (Sullivan, 2006). Numerous studies have proven that Internet users mostly read only the first page of search engine results. Subsequently it follows that owners of mission-critical web sites should be concerned about the visibility of their websites to search engine crawlers[1]. It is important for e-commerce business to have optimized sites in that manner, in order to override competition. [5]

A number of techniques are available to increase the visibility and ranking of a website. These techniques are generally classified as SEO white hat methods, SEO black hat methods and paid placement (also referred to as PPC, or pay-per-click).

Combining web technology with business economics is of high importance for e-commerce. In the last 10 years, online visibility for such online enterprises relies on the relationship between their own online sales platform and search engines. It is important to improve traffic for customers with the intent of acquiring products or services related to the customers' needs.

In order to boost their websites of low ranking web sites into highest level search rankings, Internet spammers used search engine optimization techniques. Mostly used search engine optimizations techniques include material keyword generation, directory submission, link exchanges etc. Some methods are used in ethical manner but some of them are abused.

[1] A search engine crawler is a program or automated script for indexing web sites for search engines. It browses the World Wide Web in a methodical manner in order to provide up to date data to the particular search engine.

Only since year 2007, search engines are becoming relatively resistant to manipulation. Google claimed [7] that the ranking of Web sites is using more than 200 different parameters. Since then, SEO can't be implemented by simple changes and refinement, and it becomes real science.

The remainder of the paper is organized as follows. In Section II, we describe search engine marketing and discuss why SEO is better than PPC. Section III and IV introduce ethical and non-ethical usage of SEO methodologies, and Section V advises reader about common mistakes. Finally, we present numerical findings of current level of optimization in Croatian's e-commerce in Section VI and conclude the paper in Section VII.

II. INTERNET MARKETING

Internet behavioral analysis showed that in 2008 over 90% percent of Swedish internet users make use of search engines at least once a week [2], stating that online visibility through the use of search engines now is a crucial business marketing aspect. To improve the relationship between online e-commercial platforms and search engines several applications exists within the technical field of online marketing.

A. Search engine marketing

Internet marketing involves the methods used to perform the website marketing activities. It refers to the following: search engine marketing (SEM), search engine optimization (SEO), banner ads on specific websites, e-mail marketing and Web 2.0 strategies.

Search engine marketing, or advertising to web searchers, involves a series of procedures which sites are using via other sites, portals and search engines. Specifically, it's about promoting the website by highlighting them in a web browser. It is form of Internet marketing that seeks to promote websites by increasing their visibility in search engine result pages (SERPs) through the use of search engine optimization, paid placement, contextual advertising and paid inclusion.

On the other hand, according to SEMPO (Search Engine Marketing Professional Organization) definition, SEM area covers SEO, paid ads appear on themed websites (contextual advertising) and paid inclusion in directories or search engines.

TABLE I. A COMPARISON OF SEO AND PPC [1]

	SEO	PPC
Cost	SEO could be regarded as a free search engine marketing, so the cost will be lower.	PPC, have to pay per click cost of the keyword, the cost will be higher.
Stay of rankings	Use SEO rankings, in general use of formal behavior and without of foul under, the resulting ranking is more permanent, and will not easily change.	Using the PPC's website the exposure is relatively high, but after not using the attached fee still return to the natural sort.
Risk of invalid clicks	Do not have to bear the risk of invalid clicks.	Have to bear the risk of invalid clicks.
Time	Longer	Shorter
Uncertainty of ranking	Because of the popularity of each keyword, to ensure that each keyword appear in the high rankings, will be more uncertainty factors.	Although each keyword's cost are different, but after paying, can be seen in the keyword search, belong to the unnatural order and the relatively high certainty.
Be diversionary	Because search engine algorithms are not static, so it is easier to contain by the search engine algorithms.	Paid to obtain the ranking is not easy to contain by the search engine algorithms

The largest SEM service providers are Google AdWords, Yahoo Search Marketing and Microsoft adCenter. That is the reason why, when it is about optimizing sites for better rank, it is usually about optimizing site for Google, Yahoo and MSN search engine.

B. SEO optimization versus PPC

Two ways of internet marketing are associated with web search engines: Paid placement and search engine optimization. PPC is a non-natural website ranking, and it refers to paying for common type of keyword advertisement. The locations of their occurrence are right, above and left at the site. Quite to the contrary, SEO is known as a natural website ranking. PPC is a marketing approach that one has to pay advertising costs to after the consumer clicks. According to different keywords, there will be different spreads, and the more popular keyword cost per click would be higher.

Both of them are compared [1] as shown in Table 1. After comparing SEO and PPC, it is obvious that there are much more pros for SEO, because the use of SEO can cut down the cost, even though it might take a long time to get a good rank. However, after a good rank is obtained on the search engine, it would be affected and could work for longer times.

III. USAGE OF ETHICAL SEO METHODOLOGIES

There are many methods for optimizing websites and they are grouped into two areas: onpage and offpage. Onpage optimization is a prerequisite for offpage optimization - making structures and navigating web pages, as well as all the other elements that are essential for onpage optimization, should precede indexing by the search engines and creating links to other website.

SEO techniques can be divided into two basic groups: black hat SEO and white hat SEO. SEO search engines do not approve black hat - their effect is usually only a short-term while search engines can discover and punish them by lowering rank of site in search results or even by deletion. Overview of most common used white hat methods can be found below.

A. Directory Submission

Directory submission refers to creating incoming links to a website through related page and category. A created website needs to be ranked to get good business results and manual submission to directories is a good approach for that. Internet directory is the platform on World Wide Web for information and links to many websites.

Many directories are providing free service to website in directory. Directory submission requires some information about website, such as URL, title, description, keywords, category and email. The directory submission is used as SEO technique to promote company's business.

B. Keyword Generation

Keywords generation process is method that is used by every search engine optimization. The keywords are necessary and most important part of SEO. All internet search engines require some words to elaborate information based on these words. The keywords must be related to business domain. There are many online tools available to generate keywords relevant to website's needs. Some of them are Word tracker, Yahoo keyword selector tool, Google Ad words keyword tool, Thesaurus etc. These tools operate by entering one keyword and then tool returns huge number of keywords relevant to website. But only keywords are not providing assurance to popularity of website.

C. Link Exchanges

When starting e-commerce business on internet, it is necessary to have reciprocal link or link exchange with other websites or web shops. The link exchange is the method in SEO to place your link on other websites and

vice versa. There are many types of link exchanges used: illustrate interest directly on web pages, sending e-mails or discussion forums to show interest for link exchanges. Only related website with good page rank are required to build reciprocal links. Generation of reciprocal links requires website title, description, keywords, URL, and email address for both parties. Link exchange boosts website traffic from user who clicks on website link as well as they increase ranking in web search engine. The disadvantage of reciprocal link is that some sites do frauds. Sharing of all information, including customers, is the main disadvantage.

Every method has some positive and negative aspects. Table 2 shows a description of them according to a study of four different companies [2].

TABLE II. BASIC TECHNIQUES – PROS AND CONS

	Positive aspects	**Negative aspects**
Directory submission	• majority links to the web sites are from directories • directories provide permanent link	• time consuming technique and need extra effort to compile • free listing takes time
Keyword generation	• keyword are first and most important part of every website • keywords generation is built to attract visitor	• Some abbreviation will not provide same information as required • overfilling keywords has negative consequences • time consuming
Link Exchanges	• sharing traffic provides better visibility of business goals • provide good revenue from web	• time consuming and sometimes costly • it is difficult to find good link partner and related website on which link can be placed

IV. USAGE OF NON-ETHICAL SEO METHODOLOGIES AND BAD PRACTICES

It is easy to get lost in the process of finding out which method to use and to what extent. There are many examples of a bad SEO services that will deliver only short-term improvements, while on the long run, they will only make the damage. This chapter will provide an outline of why some methods are harmful and how to avoid them.

One of the main reasons for choosing what is essentially a bad SEO service is a promise of high ranking and number of clicks in a short period of time. That is unreasonable to expect because SEO is a process and SEO should improve the website's rankings gradually. Today, no one can offer number one ranking, simply because today's search engines offer different results to different users, depending on their geographic location or personalized results. Rankings alone are a bad metric for overall performance since the whole purpose of having a good website is to gain more customers and not to have a great number of visitors who will leave the webpage as soon as they open it.

It is possible to achieve high rankings fast, but it usually means using black hat SEO and therefore should be avoided. As mentioned in chapter III, using black hat methods usually makes more harm to the rankings than it does good.

Some of the most common black hat techniques include the following:

- keyword stuffing - putting as many keywords possible into the text of a webpage, with no attempt to give a useful information; long list of keywords or their repetition;
- doorway pages – pages created specifically to rank in search engines, but that automatically redirect visitors elsewhere;
- invisible text – text of the same color as background, intended to deceive search engine bots
- linking schemes – offering payment for inbound links or creating new sites with the only purpose to create a link back to a main site; placing a lot of inbound links on unrelated pages

Not all methods need to be considered a priori a black hat to be a bad SEO service. Already mentioned white hat method - link exchanges - can too become a bad practice: not all inbound link are valuable and some may even do harm. Using only one phrase for all the links can discount the webpage by search engines and even consider it spammy.

Another common mistake is targeting the wrong keywords. Putting too broad term or using a synonym which is not common to the target population will draw wrong visitors. Service that boosts a webpage's ranking and traffic is not worth much if the target population is wrong.

Redesigning website or creating new pages without 301 Redirect can cause loss of valuable existing rankings.

Content is the key to success - creating a content that is heavy on keywords and light on actual value is definitely a long term bad strategy. Low quality texts full with keywords can downgrade the website due to recent changes made in search algorithms. Content scraping (taking content from high-ranking websites without permission) is also considered a very bad practice, , associated with spam sites and violation of copyright law. Also, putting too much focus on metadata is not so important anymore – major search engines don't pay attention to it as they used to because it was largely misused.

V. NEW ALGORITHMS AND COMMON BELIEFS THAT ARE NO LONGER TRUE

Search engine bots never forget, so content once crawled and indexed remains indexed until its URL returns a server header that tells the bot that the page is no longer available. In most cases, old content is just useless and innocently forgettable. But in some cases forgotten content, especially content whose sole purpose was to improve SEO, may contribute to a search engine's algorithmic conclusion that a site offers little value to searchers. Google's updates of indexing services are implemented to avoid it. [8]

"Farmer/Panda Update" was designed to devalue sites that were algorithmically determined to be low value to Google searchers. The update affected 12 percent of Google's results, decreasing organic traffic for bad sites. As a result, some of companies [3] experience his Google organic traffic fall about 38 percent — a decrease of 15 percent of their total site traffic — when the algorithm launched on February 24th 2011. Some of methods that resulted in devaluation of websites were:

- Building SEO content sections - if a new section is created and filled with long, keyword-rich content, especially if that section is tenuously linked to from an obscure location and not interlinked with the primary site, the section will not be effective
- Content and links placed below the footer - site visitors are unlikely to scroll below the footer, so placing keyword-rich content and links below the footer is a spam signal
- Blog comment, wiki and answer site spam - this tactic was never ethical, and is still commonly offered by low-cost SEO outfits, no one likes a pointless "thanks so much for the great information" comment with an optimized link
- Buying paid links - paid links violate search engine webmaster guidelines

Many websites publish pages that are related as part of a series, such as products in the same categories on an e-commerce website, articles that span across more than one page and galleries that show off related images, often with descriptive text. These often include pagination that enables you to go from page to page, often with numbers as links at the pages' bottoms, and sometimes with a link to see all of the content on a single page. [4]

These paginated pages often share the same HTML title element and meta description and because of that, they could be perceived by search engines to be duplicative. Google has defined a way that websites can overcome this perception, by treating these pages as if they are related. Elements within the head sections of the pages should be linked with *rel="prev"* pointing to the previous page and *rel="next"* pointing to the next page in the series. That tells Google that this content spans more than one page.

If there is an "all" page on series of paginated pages, where all of the content is available on one page, then there is option of using canonical link elements on those pages. Good practice is to add a *rel="canonical"* link to the component pages to tell Google that the View All version is the version you want to appear in search results. Also *301 redirect* should be used to redirect to canonical version of URL.New algorithms mentioned above, like Google Panda and Penguin, influenced the world of SEO professionals and the focus on how to make a good webpage [6]. The focus is on quality content, but high-quality, key-optimized content is not the only important thing. It is just one of the ingredients for successful webpage. This chapter covers common beliefs about SEO that are usually wrong or misguiding.

One of the biggest myths is the idea that the webpage should be submitted to Google in order to appear in search results. Crawlers will find the website and index it sooner or later and submission does not guarantee anything. As already mentioned in chapter IV, ranking is also not everything. Additionally, it lost some of its value when search results started to appear with rich text, short preview and author tags.

Another common myth is that keywords need to be an exact match or that there is an ideal keyword density for every page. One should use a keyword in a way that makes the most sense. The keyword, or a variation of it, should be included in a page title, headline, URL and at least once throughout the content and there is no perfect number of keywords.

Headings used to be more important, but search engines are smarter these days and people misused this too much. SEO is not just for optimizing for search engines anymore. It became also for optimizing for users. Home page should have information about who you are, what you do, location, proposition and what to do next. As mentioned in chapter V, onpage SEO is important, but to have a high rank, a webpage needs to be optimized onpage, offpage and in every other way possible.

Another aspect which should be considered today is influence of social media. Social media plays a greater role every day and should not be considered separately from SEO. Intersection of SEO and social media even has its own name - social search. It is a growing field and major companies are investing into its research and development. Content that has social connection is prioritized.

Myths related to the links and pages are that the more indexed pages one has, the better. But not everything publish gets indexed and sometimes the pages that do, do not remain so. Furthermore, some of the link related problems were solved by Google, as mentioned in chapter **Error! Reference source not found.**. Domains by the same owner that have links to each other will not be counted towards higher rank. Search engines are smart enough to know who the registered owner is, and registering under a different name is considered spamming.

Geographical location also became more important. Search engines now know where users are located and often show results that are specific to their location, which

means that information about a business location should be included in SEO.

Common misconception is that SEO is something that can be hand off to IT section. It may have been so some time ago, but with all the new algorithms and methods, SEO became an expertise for itself and to gain professional results, one needs a professional to it.

VI. SEO IN CROATIAN E-COMMERCE

In order to validate the use of SEO optimization in e-commerce, research was provided to detect optimization level of some well-known Internet shops in Croatia. It was analyzed which SEO methods were used while creating pages for electronic sales and which methods were left out. The study was conducted over six internet stores that sell clothing, footwear, technical devices, commercial goods, equipment for cars and groceries. Optimization level was checked by using available online tools. These tools have predefined fields, and for each is checked whether it exists and if yes, to what extent is it satisfied. The results are presented as a percentage, where 80% or more is considered a great result. Each page was tested by a several optimization tools in order to obtain relevant information.

Percentage of optimization level of pages amounted to an average of 49.2%, ranging from 44% to 55%. Although the importance of optimizing a site for search engines was highlighted in the last few years, the Croatian websites specializing in e-commerce still do not pay enough attention to adapting to a search engine. A possible reason is less competition in the world of e-commerce. We also found that the world famous sites for Internet commerce are more optimized. Examples of some well-known web shops that sell items in same categories, give us optimization levels mounted to an average of 65%, ranging from 62% to 68%, which is, compared to Croatian sites, almost 25% better result.

The basis of presuppositions during optimization is that the site should have particularly good onpage SEO, but our results show drawbacks even here. Page titles are well structured and are of good length, but a way of defining metadata is not good. Although entering keywords and page description in the "meta description" part of the HTML code is one of the oldest known methods in general, this rule is not respected. It is shown that the keywords and description of the page, either do not exist, or are too lengthy set - assuming 'the more the better'. The recommended amount of keywords in the meta tags is 5-10, while the studied sites contain three to four times more words. Recommended length of description meta value is up to 160 characters (two sentences), but studied sites contain average 3.5 times more characters.

Internal linking to internal pages is, as mentioned in previous chapters, a good method of optimization. Given the nature of the internet site for sale, and because of higher volumes of products that have to be linked, in all observed cases, we found a very large number of internal links. Disadvantage is that over 90% of the links had no defined value for the title attribute, which is considered bad. Similar failure occurred in the omission of the "alt" attribute for images. Less than 35% of images have specified description.

In Section III, we argued that directory submission is considered good and effective white hat method. On the other hand, it is shown that to submit a web page to appropriate directory or catalog requires investment of lots of resources. The above statement is here confirmed since only one of six studied sites is listed in the catalogs. Connection site with other sites is variable and it depends on case study. Some of the researched site has very good connections from outside (over 100 links), but other sites have less, for example only 30.

The site should also have enabled access through its IP address using the 301 redirection. If IP address of one's site is entered in the browser, user should be automatically redirected with 301 redirection to the address of that site, and the site should open in a browser. It was noticed none of the tested sites support IP canonicalization and that none of those web pages can be accessed via its IP address.

Each of the researched sites are indexed by Google, but are not indexed by Bing and Yahoo search engines. Furthermore, none of the sites are optimized for mobile devices and don't have applied responsive web design.

Internet shops are mostly linked from social networks. Linking from social networks is the popular white hat method and is noted that the e-commerce sites often use it. The most posted links on the sites are links to like, share and comment, and most links to sites are realized through Facebook and Google plus.

VII. CONCLUSION

SEO is a process that takes time and effort to achieve its final objective, which is a good, easy-to-find and optimized web page. This is especially important in expanding field of e-commerce, where the essence is to draw potential customers and offer them exactly what they are looking for.

The web shop page should always be up to date and with every update, there is a possibility of unwanted bad consequences in form of devaluation by a search engine or targeting the wrong population. To avoid that, several methods exist and well optimized web pages should always use them. In e-commerce, the special emphases are on individualization of each paginated page and the right usage of keywords – not too few nor too many, and words that are often used by targeted customers.

Our researched e-commercial web sites show that, compared to the popular worldwide sites, Croatia is still well behind and that there is a lot of room for improvement – mostly regarding the usage of keywords, optimization for indexing by other search engines and optimization for mobile devices.

REFERENCES

[1] C.Y. Chen, B.Y. Shih, Z. S. Chen, and T. H. Chen, "The exploration of internet marketing strategy by search engine optimization: A critical review and comparison", African Journal of Business Management Vol. 5(12), pp. 4644-4649, 18 June, 2011, ISSN 1993-8233 ©2011 Academic Journals

[2] M. Akram, I. Sohail, S. Hayat, M. I. Shafi, and U. Saeed, "Search Engine Optimization Techniques Practiced in Organizations: A Study of Four Organizations", Journal of computing, volume 2, issue 6, June 2010, ISSN 2151-9617

[3] J. Kocher, „SEO case study: "When past SEO efforts hurt today's traffic", Practical eCommerce, April 2011.

[4] B. Slawski and W. Critchlow, "What The Heck Is SEO? A Rebuttal", Smashing magazine, December 2012.

[5] M. Weideman, "Use of Ethical SEO Methodologies to Achieve Top Rankings in Top Search Engines", Cape Peninsula University of Technology Digital Knowledge, Informatics & Design Papers and Report, January 2007.

[6] HubSpot, "17 SEO Myths That You Should Leave Behind in 2013", eBook

[7] S. Hansell; "Google Keeps Tweaking Its Search Engine", The New York Times, January 2009.

[8] HubSpot, "How to Spot Bad SEO Services", eBook

Izazovi razvoja odgovarajućih poslovnih procesa neophodnih za migraciju telekomunikacijskih usluga

Kristina Musa, M.Sc., Darko Markulin, M.Sc.
Odjel za usluge i poslovanje s korisnicima
HT d.d.
Kupska 2, 10000 Zagreb, Hrvatska
Telefon: (385) 1-4912 721 Fax: (385) 1-4917 777
E-mail: darko.markulin@t.ht.hr, kristina.musa@t.ht.hr

Sažetak – Migracija usluga je za telekomunikacijske operatore postala svakodnevnica. Naročito se to odnosi na operatore koji su na telekomunikacijskom tržištu prisutni već duže vrijeme. Uz sve izazove koje donosi implementacija novih tehnologija kroz koje će se podržavati već postojeće telekomunikacijske usluge vrlo je važno i nezaobilazno definirati poslovne procese kroz koje će se migracija realizirati.

Osim što poslovni procesi trebaju osigurati sve potrebne korake i akcije za realizaciju usluge na novoj platformi vrlo je važno prilikom migracije osigurati neprekinutost usluge ili minimalni prekid, ukoliko je on nezaobilazan. Kako se kod telekomunikacijskih operatora s dužim prisustvom na telekomunikacijskom tržištu radi o većoj bazi korisnika poslovne procese koji osiguravaju potrebne akcije i aktivaciju mrežnih resursa potrebno je realizirati kroz sustave programske podrške, a sustav tijeka poslova ima središnju ulogu u koordiniranju tih koraka.

U radu je prikazan primjer razvoja programske podrške i poslovnih procesa potrebnih za podržavanje migracije govorne usluge s PSTN na VoIP tehnologiju. Prikazani model korišten je prilikom modeliranja procesa migracije u HT-u.

Ključne riječi – migracija usluga, poslovni procesi, neprekinutost usluga, programska podrška

I. UVOD

Telekomunikacijski operatori koji su duže prisutni na telekomunikacijskom tržištu suočeni su s neminovnošću zamjene pojedine tehnologije koja se koristi za potrebe realizacije telekomunikacijske usluge. Na to su prisiljeni jer postojeća tehnologija više ne može zadovoljavati korisničke potrebe za što složenijim i sofisticiranijim uslugama i/ili je održavanje postojeće tehnologije postalo preskupo i neisplativo. S obzirom da takvi telekomunikacijski operatori imaju veliku bazu korisnika migracija usluga s jedne tehnologije na drugu vrlo je složeno i osjetljivo pitanje. Pogotovo se to odnosi na pristupnu mrežu gdje zamjena tehnologije ima utjecaj na svakog pojedinog korisnika. Primjer takve migracije je migracija s bakrene pristupne mreže na optičku ili migracija s PSTN-a na VoIP telefoniju. Dodatnu težinu i složenost migracije usluga s jedne na drugu tehnologiju daju složenost telekomunikacijskog tržišta na kojem korisnici očekuju sve veću kvalitetu i pouzdanost usluga te prisutnost sve brojnije i jače konkurencije tako da problemi prilikom pružanja usluga mogu rezultirati nezadovoljsvom korisnika i prelaskom drugom Operatoru.

Provesti migraciju usluga s jedne tehnologije na drugu u ovakvim uvjetima sigurno nije lako. S jedne strane potrebno je osigurati dobru komunikaciju s korisnicima upoznavajući ih na vrijeme da se s migracijom usluge koju koriste s jedne tehnologije na drugu osigurava bolja kvaliteta usluge i osigurava platforma za daljnji razvoj usluga i kvalitete pružanja usluga. Ovo je vrlo osjetljiv segment migracije jer telekomunikacijskom operatoru sigurno nije u interesu promjenom tehnologije uzrokovati otpor i nezadovoljstvo korisnika. Ili bar to svesti na najmanju moguću mjeru. S druge strane samu migraciju potrebno je provesti na takav način da se prilikom migracije osigura neprekinutost ili, ukoliko je prekid ipak neizbježan, minimalan prekid pružanja usluge. Da bi se to osiguralo potrebni su dobro dizajnirani poslovni procesi. Vrlo je važno istaknuti da nije dovoljno prije migracije implementirati poslovne procese i programsku podršku za sami proces migracije, već je bitno osigurati poslovne procese koji osiguravaju suživot različitih tehnologija jer proces migracije nije trenutačan i traje duži niz godina. Posebno ako se radi o migraciji između tehnologija u pristupnoj mreži. Naravno, preduvjet za uspješnu migraciju je da performanse nove tehnologije zaista osiguravaju kvalitetno i pouzdano pružanje usluga.

Postojanje programske podrške za upravljanjem procesa migracije na području pristupne mreži jedan je od odlučujućih faktora za uspješnu migraciju [1], [2], [5]. Kako se radi o velikoj bazi korisnika i kako se radi o vrlo složenom slijedu akcija koje se izvršavaju na različitim platformama za pružanje usluga, različitim mjestima u pristupnoj mreži i upućene su na različite izvršioce dobro dizajniran i upravljan poslovni proces je preduvjet za uspješnu migraciju. U nastavku rada detaljno su opisani postupci, poteškoće i rezultati dobiveni prilikom razvoja programske podrška za migraciju govorne usluge s PSTN na VoIP tehnologiju. Programska podrška je razvijena za potrebe upravljanja konfiguracijama mrežnih elementa i ljudskim radom na području telekomunikacijske pristupne mreže.

II. DEFINIRANJE MOGUĆIH SCENARIJA I POSLOVNIH PROCESA

Da bi se dobro pripremila programska podrška za migraciju potrebno je definirati sve moguće scenarije koji

se mogu pojaviti tijekom i nakon migracije [3]. Scenariji ovise o složenosti usluga i poslovnih procesa i bitno je definiranje mogućih scenarija sprovesti kroz jedan od odgovarajućih načina koji omogućuju jednostavni i pregledni prikaz mogućih scenarija. Jedan od načina je jasno definiranje slijeda aktivnosti u obliku strukturiranog teksta kroz koji se navode mogući scenariji. Taj se popis može dodatno obogatiti definiranjem osnovnih akcija koje kroz svaki scenarij treba osigurati. Primjer takvog popisa scenarija prikazan je u nastavku rada. Obuhvaćena su samo dva scenarija.

Scenariji za migraciju HALO (govorne) usluge

1. Korisnik ima samo HALO uslugu
   Migracija HALO usluge izvodi se na MSAN tehnologiju
   - Konfiguracija porta na MSAN-u, port ostaje blokiran
   - Konfiguracija na IN centrali ukoliko broj korisnika već ranije nije portiran na područje druge centrale
   - Ožičenje MDF-a
   - Konfiguracija na lokalnoj centrali (preusmjeravanje dolaznih poziva na IN centralu)
   - Konfiguracija na IN centrali ukoliko je broj korisnika već ranije bio portiran na područje druge centrale
   - Deblokiranje porta na MSAN-u
   - Raspajanje broja na lokalnoj centrali
2. Korisnik ima HALO i jednu od širokopojasnih usluga (MAXadsl (pristup na Internet) i/ili MAXtv (IPTV))
   Migracija HALO usluge izvodi se na ADSL tehnologiju, instalaciju izvodi tehničar
   - Konfiguracija port-a (postavljanja novog profila) na DSLAM-u, port ostaje blokiran
   - Konfiguracija na IN centrali ukoliko broj korisnika već ranije nije portiran na područje druge centrale
   - Odlazak kod korisnika na migraciju
   - Konfiguracija na lokalnoj centrali (preusmjeravanje dolaznih poziva na IN centralu)
   - Konfiguracija na IN centrali ukoliko je broj korisnika već ranije bio portiran na područje druge centrale
   - Deblokiranje porta na DSLAM-u
   - Raspajanje broja na lokalnoj centrali

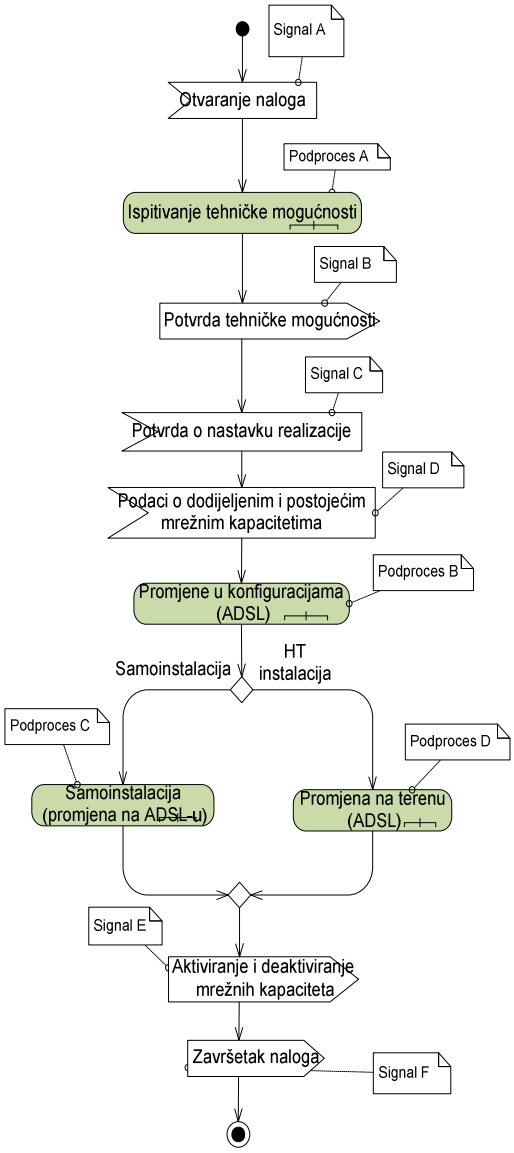
Među scenarijima potrebno je uvrstiti i scenarije koji se odnose na poslovne procese koje je potrebno osigurati nakon implementacije nove tehnologije i migracije usluga. Primjer je preseljenje ili promjena broja migriranog korisnika. Pogrešno je razmišljanje da je za uspješnu realizaciju migracije potrebno samo dobro definirati procese koji to omogućuju. Potpuno uspješna migracija je onda kada se nakon uspješno migrirane usluge na novu tehnologiju osigura njezin neometan suživot s ostalim uslugama koje se nude kroz različite tehnologije. Dodatnu složenost uvode i regulatorne obaveze te definirani rokovi realizacije i otklanjanja smetnji koje je potrebno uskladiti s aktivnostima migracije.

Naravno da je gotovo nemoguće unaprijed definirati baš sve moguće scenarije i potrebne poslovne procese koji se mogu pojaviti prilikom i nakon migracije. Posebno je teško predvidjeti sve moguće situacije koje se mogu pojaviti prilikom i nakon migracije. Dodatnu nesigurnost u definiranje poslovnih procesa migracije unosi ponašanje korisnika koje uvelike može utjecati na potrebu izmjene primijenjenog poslovnog procesa i programske podrške. No, dobro definirani scenariji i poslovni procesi znatno smanjuju mogućnost pojavljivanja nepredviđenih situacija i potrebe za naknadnim izmjenama u implementiranim poslovnim procesima.

III. MODELIRANJE PROGRAMSKE PODRŠKE

Nakon definiranja scenarija za njihovo uspješno provođenje potrebno je modelirati poslovne procese. Model treba definirati sve akcije koje je u određenom vremenskom slijedu potrebno poduzeti kako bi se usluga uspješno migrirala s jedne tehnologije na drugu. Kako je telekomunikacijska mreža, usluge i tržište vrlo složeno, posljedično su i poslovni procesi za migraciju vrlo složeni. Zbog toga je vrlo bitno odabrati model kojim se na uspješan i što jednostavniji način mogu modelirati tako složeni poslovni procesi. U [6] su definirani osnovni elementi modela uz pomoć kojeg je moguće modelirati poslovne procese za realizaciju telekomunikacijskih usluga. Kako poslovni procesi za realizaciju telekomunikacijskih usluga mogu biti i vrlo složeni procesi za realizaciju više različitih usluga tako se ovaj model može uspješno primijeniti i na modeliranje poslovnih procesa za migraciju usluga.

Model predstavlja dinamičko ponašanje programske podrške kojom se upravlja odvijanjem migracije. Model omogućuje prikaz poslovnog procesa kroz više razina složenosti [6]. Na slici 1 model definira osnovni scenarij. Poslovni proces je prikazan kroz najvišu razinu apstrakcije u kojoj se daje osnovni prikaz procesa. Definirana je komunikacija osnovnih sustava koji sudjeluju u osiguravanju programske podrške i osnovni dijelovi procesa koji čine cjeline u sklopu kojih se poduzimaju potrebne akcije. Tako cjelina „Promjene u konfiguracijama (ADSL)“ obuhvaća potrebne akcije koje se izvršavaju prije odlaska korisniku ili obavijesti korisniku da migrira svoju uslugu. „Samoinstalacija (Promjena na ADSL-u)“ obuhvaća akcije koje se odnose na komunikaciju s korisnikom i završne konfiguracije u mreži prije uspješnog završetka migracije. „Promjene na terenu (ADSL)“ obuhvaća akcije kroz koje tehničar migrira uslugu kod korisnika i pokreće završne konfiguracije. Uz svaki element modela nalazi se zabilješka kroz koju se dodatno obogaćuje izražajna snaga modela. Prolaskom kroz ovakav dijagram dobivaju se osnovne informacije o procesu. Te su informacije kroz grafički prikaz mnogo jasnije i bliže ljudskoj percepciji nego je to slučaj kod definiranja scenarija i njihovih osnovnih akcija.



Signal A - Iniciranje otvaranje naloga na strani mreže

Podproces A - Ispitivanje tehničke mogućnosti na strani mreže za tražene usluge. Rezervacija potrebnih mrežnih kapaciteta i vraćanje podataka o postojanju tehničke mogućnosti Prodaji

Signal B - Potvrda postojanja tehničke mogućnosti

Signal C - Potvrda da se može krenuti s realizacijom usluge na strani mreže

Signal D - Dobivaju se podaci o rezerviranim i postojećim mrežnim kapacitetima

Podproces B - Rekonfiguracija ACS-a, DSLAM-a, IMS-a. Kod ACS-a i DSLAM-a radi se o promjeni profila

Podproces C - Korisnik sam radi konfiguraciju modema. Automigracija u slučaju ako korisnik ima odgovarajući modem. Samomigracija ako korisnik nema odgovarajući modem koji je potrebno zamijeniti. Rekonfiguracija porta na PSTN centrali radi se u sklopu podprocesa «Konfiguracija centrala (VoIP – teren)»

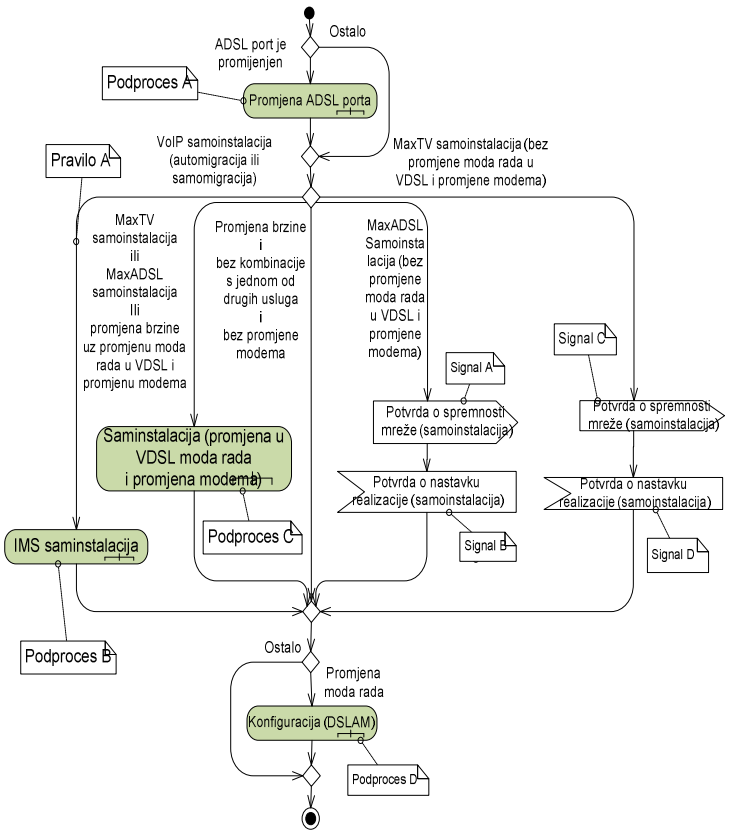
Podproces D - Odlazak tehničara kod korisnika i uključenje usluge na HT mrežu. Rekonfiguracija porta na PSTN centrali radi se u sklopu podprocesa «Konfiguracija centrala (VoIP – teren)»

Signal E - Aktiviranje mrežnih kapaciteta za usluge koje se uključuju. Deaktiviranje mrežnih kapaciteta za usluge koje se isključuju

Signal F - Slanje informacija Prodaji o realizaciji naloga. Potrebni podaci: Due date za ADSL (HT instalacija) - vrijeme realizacije zadnjeg od slijedećih zadataka: «Instalacija kod korisnika», «FC - ADSL popravak (s korisnikom)» ili «Realizacija PSU-a i instalacija kod korisnika (VI)». Due date za POTS - vrijeme automatske rekonfiguracije porta na PSTN centrali (u sklopu VoIP rekonfiguracije (teren))

Slika 1: Osnovni prikaz migracije HALO usluge

Kroz slijedeće razine prikaza modela modeliraju se cjeline definirane u osnovnom prikazu. Ovo je niža razina apstrakcije i prikaz je puno složeniji. Kroz cijeli model se koristi sintaksa UML dijagrama aktivnosti [4]. Time se dobiva na većoj točnosti modela i pridonosi boljem razumijevanju svih uključenih aktera u definiranju i razvoju traženih poslovnih procesa. Kako ovaj model služi za specifikaciju poslovnih procesa na temelju koje se pristupa razvoju programske podrške sintaksa modela treba biti vrlo precizna i jasno treba modelirati i definirati tijek poslovnog procesa i sve detalje bitne za razvoj programske podrške. Na slici 2 prikazan je podproces „Samoinstalacija (promjena na ADSL-u)“. Ovaj se podproces ne koristi samo u sklopu migracije, već i u ostalim procesima uključenja usluga u kojima instalaciju opreme radi korisnik. Ovo dodatno pojačava potrebu za preciznošću modela.



Pravilo A - Iz Donata će uvijek kod IMS samomigracije dolaziti preferirani mod rada: ADSL

Podproces A – MDF ožičenje i postavljanje profila na novom DSLAM portu

Podproces B – poruka o spremnosti mreže i rekonfiguracija mrežnih elemenata nakon ostvarenih potrebnih preduvjeta

Podproces C – distribucija odgovarajućeg modema korisniku i promjena moda nakon ostvarenih potrebnih preduvjeta

Podproces D - Pokreće se promjena moda rada DSLAM porta (iz ADSL moda u VDSL)

Signal A - Šalje se poruka Prodaji o spremnosti mreže za uključenje usluge; Poruka «ADSL samoinstalacija – DSLAM spreman»

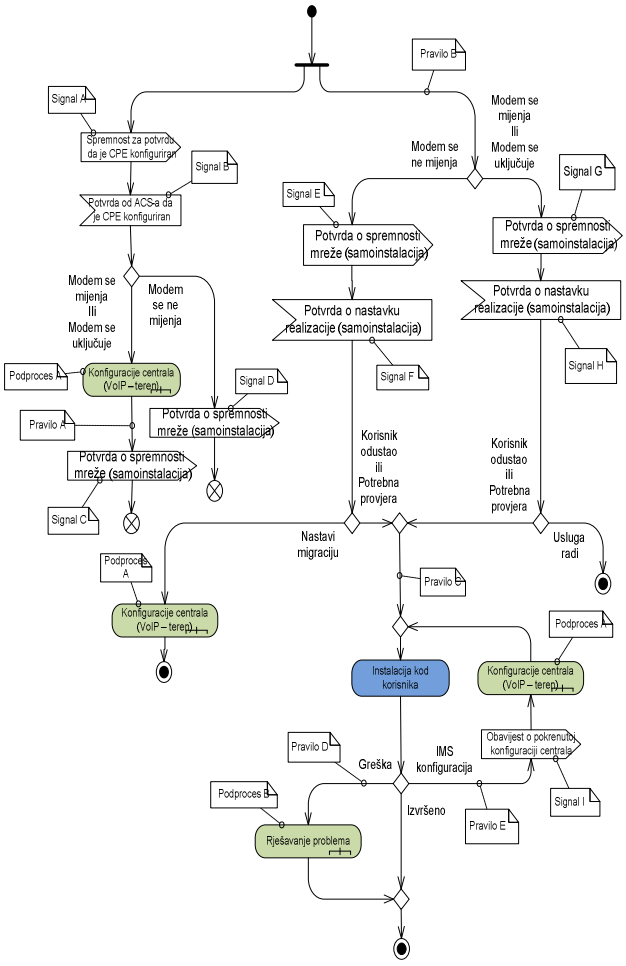
Signal B - Potvrda da se može završiti realizacija usluge

Signal C - Šalje se poruka Prodaji o spremnosti mreže za uključenje usluge; Poruka «MAXTV samoinstalacija –DSLAM spreman»

Signal D - Potvrda da se može završiti realizacija usluge

Slika 2: Niža razina prikaza poslovnog procesa: „Samoinstalacija (promjena na ADSL-u)“

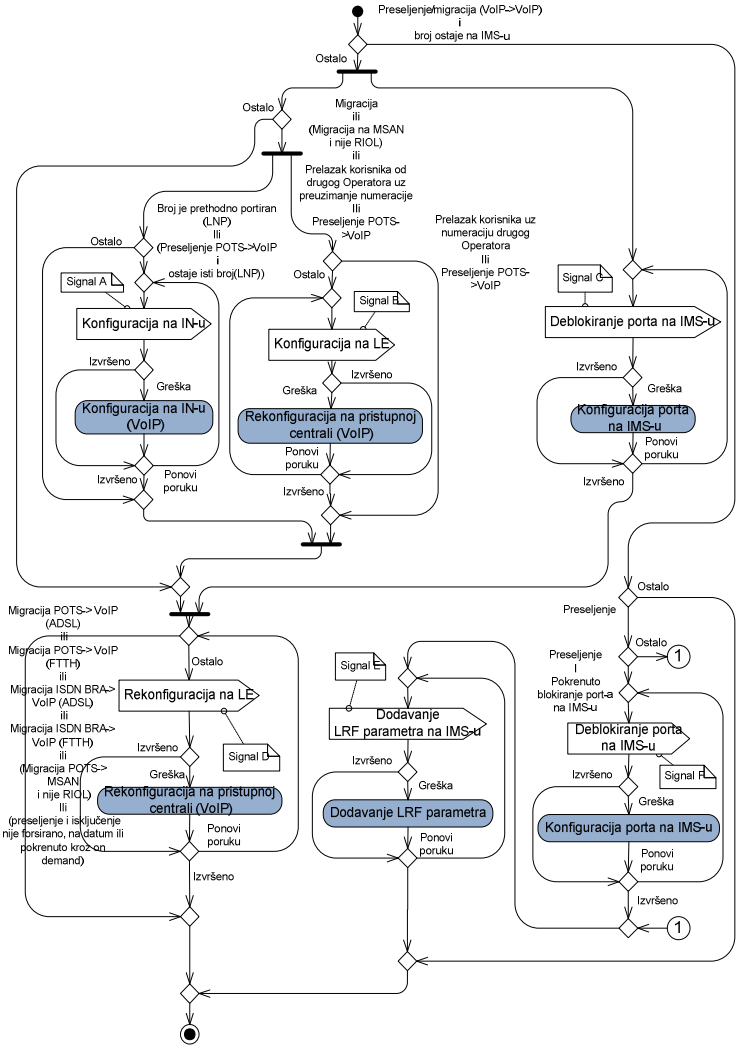
Na slici 3 prikazana je slijedeća razina prikaza u sklopu kojeg je kroz podproces „IMS samomigracija“ modeliran dio procesa koji se odnosu na komunikaciju s korisnikom i konačnu konfiguraciju mrežnih elemenata.



Signal A - Šalje se poruka ACS-u da je moguće primiti potvrdu o tome da je CPE konfiguriran
Signal B - Potvrda da je CPE konfiguriran
Signal C - Šalje se poruka Prodaji o spremnosti mreže za uključenje usluge
Poruka «IMSsamomigracija, modem konfiguriran»
Signal D - Šalje se poruka Prodaji o spremnosti mreže za uključenje usluge
Poruka «IMSautomigracija, modem konfiguriran»
Signal E - Šalje se poruka Prodaji o spremnosti mreže za uključenje usluge
Poruka «IMSautomigracija – mreža spremna, modem još nije konfiguriran"»
Signal F - Potvrda da se može nastaviti realizacija usluge
Nastavak realizacije ovisi o odgovoru od CS-a u potvrdi.
Signal G - Šalje se poruka Prodaji o spremnosti mreže za uključenje usluge
Poruka «IMSsamomigracija – mreža spremna, modem još nije konfiguriran"»
Signal H - Potvrda da se može nastaviti realizacija usluge
Nastavak realizacije ovisi o odgovoru od CS-a u potvrdi
Signal I - Šalje se poruka Prodaji o tome da su pokrenute konfiguracije na centrala
Poruka «IMSsamomigracija, pokrenuta konfiguracija na centralama»
Podproces A - Završne konfiguracije na lokalnoj, IN centrali i VoIP centrali
Podproces B – pokretanje procedure za ispravljanje pogreške koja onemogućuje nastavak procesa
Pravilo A - Ako se grana u kojoj se čeka potvrda iz Donata o nastavku samoinstalacije završi (kroz potvrdu iz Donata ili odabir IMS konfiguracije iz zadatka «Instalacija kod korisnika») automatski se završava grana s čekanjem potvrde da je CPE konfiguriran)
Pravilo B - Ako se u grani s čekanjem potvrde da je CPE konfiguriran odradi konfiguracija centrala (modem se ne mijenja) potrebno je u zadatku «Instalacija kod korisnika» onemogućiti izbor «IMS konfiguracija»
Pravilo C - Ukoliko iz zadatka «Instalacija kod korisnika» nije pokrenuta opcija «IMS konfiguracija» moguće je iz Donata inicirati nastavak procesa tako da se zatvara zadatak «Instalacija kod korisnika» i pokreće konfiguracija na centralama («Konfiguracija centrala (VoIP – teren)» te nakon toga završava proces automigracije
Pravilo D - Onemogući izbor Izvršeno ukoliko nije izvršena rekonfiguracija na centralama (ACS trigger ili izbor «IMS konfiguracija»)
Pravilo E - Onemogućiti ovaj izbor ukoliko se ne radi o VoIP uključenju.

Slika 3: Niža razina prikaza poslovnog procesa: „IMS samoinstalacija“

Dio procesa koji se odnosi na komunikaciju s korisnikom u ovom modelu nije prikazan. Taj se dio inicira porukom CRM sustavu o spremnosti mreže i završava odgovorom CRM sustava koji šalje informaciju o uspješnoj ili neuspješnoj migraciji od strane korisnika. U slučaju uspješne migracije pokreću se završne konfiguracije u mreži. U slučaju neuspješne migracije pokreće se zadatak za tehničara koji izvršava migraciju i nakon toga pokreće završne konfiguracije. I u ovoj se razini koriste zabilješke koje osim što dodatno opisuju elemente modela dodatno opisuju i poslovna pravila koje je potrebno ugraditi u poslovne procese. Završne konfiguracije modelirane su kroz podproces „Konfiguracija centrala (VoIP – teren)“. Ovaj podproces prikazan je na slici 4.



Signal A - Konfiguracija na IN-u zbog preusmjeravanja numeracije na VoIP centralu ili prema mreži drugog Operatora
Signal B - Konfiguracija na lokalnoj centrali zbog preusmjeravanja dolaznog poziva na IN
Signal C - Omogućavanje odlaznih poziva na VoIP centrali
Signal D - Rekonfiguracija na lokalnoj centrali (POTS ili ISDN BRA) – odspajanje broja s porta i preusmjeravanje dolaznih poziva na IN
Signal E- Dodavanje LRF parametra u slučaju kada se kod presljenja ne mijenja broj na IMS-u («VoIP preseljenje»)
Signal F - Deblokiranje porta na IMS-u slučaju preseljenja («VoIP preseljenje»)

Slika 4: Niža razina prikaza poslovnog procesa: „Konfiguracija centrala (VoIP) - teren“

Ovo je najniža razina modela u kojem više nema pozivanja podprocesa i u kojem su definirane aktivacije

mrežnih elemenata. Prikazani model korišten je prilikom modeliranja procesa migracije u HT-u. Model se pokazao kao uspješan alat za definiranje poslovnih procesa i osnova za razvoj programske podrške za provođenje migracije. Kako je broj i raznolikost potrebnih poslovnih procesa vrlo velik prije pokretanja migracija potrebno je bilo osigurati realizaciju usluga („suživot“) kako za maloprodaju tako i za veleprodaju sa svim mogućim scenarijima migracije. Bilo je potrebno definirati poslovne procese kroz koje se povezuju aktivnosti za realizaciju migracije i aktivnosti vezane uz realizaciju maloprodajne i/ili veleprodajne usluge. Ukoliko ne postoje takvi poslovni procesi potrebna je primjena različitih workaround rješenja što nikako nije efikasno i dobro rješenje.

IV. ZAKLJUČAK

Migracija usluga s jedne tehnologije na drugu vrlo je izazovan i složen zadatak. Migracija većeg broja korisnika, a to je najčešće slučaj kod migracija koje provode telekomunikacijski Operatori s dužim prisustvom na tržištu, gotovo je nemoguća bez efikasne programske podrške koja osigurava poslovne procese za provođenje migracije. Da bi se uspješno sprovela migracija uz pomoć programske podrške potrebno je odraditi vrlo dobru pripremu za početak migracije uz definiranje mogućih scenarija i situacija koje se mogu pojaviti tijekom i nakon migracije. Svaki scenarij treba imati dobro modeliran poslovni proces kroz koji se mogu na neometan način provoditi migracija ili poslovni procesi koji obuhvaćaju i migriranu uslugu. Konačno, potrebno je kontinuirano pratiti odvijanje migracija te ostale procese koji obuhvaćaju migrirane usluge. Uočene nedostatke treba odmah ispravljati.

LITERATURA

[1] Musa K., Markulin D., “Realizacija složenih procesa uključenja kroz sustave za podršku”, *Mipro 2009, HT d.d.*, 2009.

[2] Musa K., Markulin D., ”Način implementacije novih usluga u integriranom IT okruženju”, *Mipro 2007, HT d.d.*, 2007.

[3] Markulin D., Kunštić M., ” Formal modeling business management of public telecommunication network”, *Mipro 2010, HT d.d.*, 2010.

[4] Markulin D., Musa K., Kunštić M., “Using UML 2 activity diagram for visual business management modeling”, *Mipro 2011, HT d.d.*, 2011.

[5] Markulin D., Musa K., “ Proposal for specification of complex processes for telecommunication service provisioning in existing IT environment”, *Mipro 2010, HT d.d.*, 2010.

[6] Markulin D., Musa K., “Proposal for business process modelling and development in condition of complex telecommunication market ”, *Mipro 2012, HT d.d.*, 2012.

Povezanost društveno odgovornog poslovanja i poslovnih performansi

Nataša Pivačić*, Nenad Kusić **
* Ericsson Nikola Tesla d.d.
Poljička cesta 39, Split, Hrvatska
E-mail adresa: natasa.pivacic@ericsson.com

Sažetak - Svaka organizacija bi trebala u svoje poslovanje uključiti strateško i aktivno praćenje društveno odgovornog poslovanja kako bi izgradila održivu konkurentsku sposobnost, dugoročan rast i opstanak te unaprijedila svoje poslovne performanse.

U radu su analizirane i proučene teorije i istraživanja iz područja društveno odgovornog poslovanja poduzeća, s osvrtom na financijska mjerila kao sredstvo mjerenja poslovnih performansi. Kao rezultat teorijskih spoznaja o navedenim pokazateljima, osmišljen je i konstruiran model povezanosti društveno odgovornog poslovanja i poslovnih performansi poduzeća. Empirijska istraživanja provedena su formuliranim anketnim upitnikom na uzorku od 76 velikih hrvatskih poduzeća. Rezultati istraživanja su ukazali na pozitivnu međuovisnost ovih dvaju pokazatelja.

I. UVOD

Društvena odgovornost pomaže poduzeću u izgradnji ugleda, osigurava motiviranost i zadovoljstvo zaposlenika kao i širu društvenu prihvaćenost, jača konkurentnost i dugoročno omogućava brži poslovni razvoj poduzeća. U tom smislu uvijek poduzeća uz financijsku, moraju imati i visoko osviještenu odgovornost za društvenu zajednicu te održivi razvoj.

Primarni cilj istraživanja bila bi analiza povezanosti društveno odgovornog poslovanja i poslovnih performansi u poduzećima. Temeljem ovog cilja može se postaviti hipoteza: Implementacija društveno odgovornog poslovanja u strategiju poduzeća donosi poduzeću bolje performanse. Koncept društvene odgovornosti poduzeća je primjenjiv na poduzeća svih veličina, premda će se najviše istražiti utjecaj velikih poduzeća jer je njihov utjecaj najvidljiviji, a moć najveća.

Svrha istraživanja ovog rada bila bi podići razinu svijesti o društveno odgovornom poslovanju i njegovom utjecaju, istražiti i mogućnosti i izazove, te dati niz preporuka za učinkovite buduće aktivnosti i promicanje društveno odgovornog poslovanja u Hrvatskoj.

II. REZULTATI TEORIJSKOG RAZMATRANJA O DRUŠTVENO ODGOVORNOM POSLOVANJU

U literaturi društvena odgovornost je opisana kroz mnoštvo definicija (Howard Bowen 1953, Keith Davis i Robert Blomstrom 1960, definicija Svjetskog poslovnog savjeta za održivi razvoj, predstavnici Evropske unije u dokumentu „Zelena knjiga“ 2001, razne organizacije koje djeluju na području društveno odgovornog poslovanja (DOP-a). Jedna od najviše korištenih definicija govori da „Društveno odgovorno poslovanje je koncept u kojem poslovni subjekt odlučuje na dobrovoljnoj osnovi doprinositi boljem društvu i čišćem okolišu, u interakciji s ostalim dionicima.“

Postoje mnogobrojne koristi od DOP-a poput povećanja prodaje i udjela na tržištu, poboljšanja konkurentnosti kompanije i tržišnog pozicioniranja, smanjenja troškova i unapređenja kvalitete upravljačkih procesa i proizvoda, jačanja brenda, reputacije i imidža, povećanja zapošljavanja i zadržavanja kvalitetnih radnika, jačanja motivacije, povećanja privlačnosti za ulagače i većeg povrata, povećanja učinkovitosti poslovanja, boljeg odnosa sa svim interesnim grupama [1].

Društveno odgovorno poslovanje se može prikazati kroz ključna područja (osnovne poslovne aktivnosti, investiranje u društvo i filantropija, dijalog o politikama, zagovaranje i jačanje institucija), kroz internu i eksternu dimenziju te kroz internu i eksternu razinu [2]. Postoji nekoliko sredstava u provedbi politike DOP-a poput vrijednosti, načela, kodeksa ponašanja; praćenja, mjerenja i izvještavanja; vodstva; inicijativa i normi; ključnih aktera; partnerstva [3]. Društvena odgovornost poduzeća ima svoju hijerarhiju u kojoj se na najnižoj razini nalazi odgovornost za ostvarenje profita, zatim slijedi poštivanje zakona, nakon toga etičnost u poslovanju i konačno doprinos poduzeća društvu. Te su odgovornosti poredane odozdo prema gore u ovisnosti o njihovoj veličini i frekvenciji kojom im menadžer pristupa [4].

U Europi i širom svijeta u porastu je broj kompanija koje promoviraju svoje strategije društvene odgovornosti. Opća je procjena da u Hrvatskoj djeluje oko 200 poduzeća koja su prihvatila DOP kao dio svoje vrijednosne i strateške orijentacije. Prakse DOP-a u Hrvatskoj relativno su dobro razvijene, učestalije na područjima kao što su kvaliteta i sigurnost radnog mjesta, zadovoljstvo potrošača, zaštita okoliša, ulaganje i partnerstvo u zajednici, a manje učestale u područjima kao što su korporacijsko upravljanje, integriranje DOP-a u procjenu rizika i razvoj poslovne strategije, upravljanje dobavljačkim lancem i društveno odgovorno ulaganje.

III. MODEL STUPNJA RAZVIJENOSTI DRUŠTVENO ODGOVORNOG POSLOVANJA

Klasifikacija DOP-a je predstavljena u obliku matrice pokazatelja inicijalnih (početnih) ciljeva koji se mogu iskoristiti kao polazna točka za razvoj pokazatelja

održivosti i metoda procjene zadovoljenosti ciljeva i načela DOP-a. Osnovna klasifikacija područja DOP-a koja odražavaju ciljeve i načela DOP-a je dana u SFA1A dokumentu " An introduction to CSR": Ekologija; Društvo (ljudska prava i prava radnika, zdravstvo, društvena pitanja) te Ekonomija (etičko poslovanje, upravljanje).

Navedena područja DOP-a predstavljena su od strane eminentnih udruga koje se bave razvojem DOP-a kao što su "Business in the Community", "CSR Europe", "Business Leader Forum" itd. te su postali dio raznih nacionalnih razvojnih programa. DOP više predstavlja univerzalni skup regulativa i potrebnih znanja vezanih za svako poduzeće pojedinačno. Potrebno je definirati taj skup osnovnih pokazatelja za svako područje DOP-a.

Osnovna matrica DOP-a se sastoji od ekonomske, društvene te ekološke matrice. Svaka matrica definira osnovne tematike DOP-a koje predstavljaju najbolja iskustva koja vežemo uz odgovarajuće pokazatelje DOP-a.

A. Ekonomska matrica DOP-a

Ekonomska matrica u prvom redu podrazumijeva etičko poslovanje, transparentno upravljanje i sudjelovanje u globalnom partnerstvu.

Etičko poslovanje

Etičko poslovanje obuhvaća široko područje poslovnih aktivnosti koje definiraju vezu poduzeća s svojim dobavljačima i klijentima. Neki od primjera su etičko oglašavanje, cijena, naplata, privatnost korisnika, zaštita proizvoda, kvaliteta i sigurnost proizvoda itd. Osnovna postavka DOP-a je podržavanje jednakih mogućnosti uključujući raznovrsnost u procesu nabavu i zaštitu djece.

TABLICA I. EKONOMSKA MATRICA DOP-A [5]

Ekonomska matrica DOP-a	
Zadatak	Akcije
Sigurnost proizvoda	• Integracija proizvodnje i distribucije • Sigurnost proizvoda • Sigurnost ambalaže • Minimizacija negativnih utjecaja proizvoda i usluga • Uvažavanje posebnih zahtjeva kupaca
Odgovoran odnos prema dobavljačima	• Jednake mogućnosti i dostupnost kupcima • Promicanje raznovrsnosti dobavljača • Suradnja sa ugovornim stranama, agentima i dobavljačima u svrhu implementacije načela DOP-a • Sudjelovanje u lancu odgovornosti u svrhu promicanja ljudskih prava • Podmirenje potraživanja • Transfer tehnologije i znanja • Suradnja sa javnim tijelima za promicanje raznovrsnosti dobavljača
Zaštita osobnih podataka	• Politika zaštite privatnosti dobavljača i kupaca
Upravljanje kompanijom	• Transparentnost • Kompenzacije i kontrola izvršnih tijela
Etički standardi	• Poslovna etika i nadgledanje kretanja cijena, pranja novca, utaje poreza, mita, monopola

Ekonomska matrica DOP-a (nastavak)	
Zadatak	Akcije
Pitanja vezana uz kupce	• Omjer cijena/vrijednost • Kvaliteta proizvoda • Informiranje i iskreno označavanje proizvoda • Istinito oglašavanje • Odgovoran marketing i prodaja, osobito kada su u pitanju djeca • Aktivnosti vezane za zadržavanje kupca
Podrška pravednoj globalizaciji	• Privrženost diplomaciji • Aktivno sudjelovanje u procesu pravedne globalizacije

B. Društvena matrica DOP-a

TABLICA II. DRUŠTVENA MATRICA DOP-A [5]

Društvena matrica DOP-a	
Zadatak	Akcije
Suzbijanje prisilnog i dječjeg rada	• Prevencija svih oblika prisilnog i dječjeg rada • Podrška akcijama/inicijativama za eliminaciju prisilnog i dječjeg rada
Osiguranje jednakopravnosti zaposlenika	• Jednake mogućnosti za žene, manjine i fizički nemoćne • Otklanjanje bilo kakvih oblika diskriminacije • Suzbijanje uznemiravanja na radnom mjestu • Politika jednakih prava i obveza za sve zaposlenike
Obuka i profesionalni razvoj	• Razvoj kompanijske politike planiranja obuka i razvoja kompetencija zaposlenika • Programi kontinuirane zaposlenosti i razvijanje dugoročne karijere zaposlenika u kompaniji • Tečajevi usvajanja novih znanja tehnologija,profesionalno usmjeravanje • Podupiranje postdiplomskih studija • Odobravanje dopusta za obrazovanje • Razvoj upravljačkih struktura • Programi obuke o DOP-u • Posebne programe obuke za zaposlenike
Dobri unutrašnji odnosi	• Omogućiti slobodu izražavanja • Transparetnost u pitanjima zaposlenja i radnih uvjeta • Pravodobno i istinito informiranje zaposlenika • Poticanje otvorene komunikacije • Podrška zaposlenicima u balansiranju radnih obveza, osobnog zalaganja i obiteljskih potreba • Dobra suradnja sa sindikatima • Zbrinjavanje radnika i podrška u pronalaženju posla
Nagrađivanje zaposlenika	• Pohvale i promocije • Nagrađivanje zaposlenika kroz politiku podjele dobiti • Dodatno životno/zdravstveno osiguranje • Mirovinsko osiguranje • Zbrinjavanje invalidnih osoba • Podrška realokaciji zaposlenika • Regulacija prekovremenog rada • Druge beneficije
Optimizacija utjecaja proizvoda na zdravlje	• Od kompanija čiji proizvodi mogu utjecati na zdravlje očekuje se da minimiziraju negativni utjecaj svojih proizvoda i usluga
Sponzorstva	• Sponzorstvo • Filantropija

Društvena matrica DOP-a (nastavak)	
Zadatak	Akcije
Zdravlje i sigurnost na radnom mjestu	• Podizanje svijesti o važnosti zaštite na radu i očuvanja zdravlja na radnom mjestu kroz razne oblike edukacije • Dostupnost obučenih radnika/savjetnika o zdravstvu i zaštiti na radu • Upute o sučeljavanju sa stresom na radnom mjestu • Dostupnost zdravstvenih ustanova • Provođenje akcija zaštite zdravlja zaposlenika
Podrška lokalnim inicijativama za zaštitu zdravlja	• Ohrabrivanje zaposlenka na aktivno sudjelovanje u životu zajednice i volontiranje • Suradnja sa lokalnom zajednicom i zdravstvenim organizacijama na zdravsteim pitanjima
Podrška lokalnim inicijativama za edukaciju	• Podrška obuci mladih • Podrška edukacijskih središta za obuku nezaposlenih • Podupiranje edukacije ljudi s posebnim potrebama
Sudjelovanje u programima Global Partnership on Health	• Pristup lijekovima i osnovnoj zdravstvenoj zaštiti u prvom redu u zemljama u razvoju • Razvoj osnovne zdravstvene infrastrukture uključujući izgradnju vodovodne i sanitarne mreže u zemljama u razvoju • Smanjenje smrtnosti djece • Borba protiv HIV/AIDS-a, malarije i ostalih zaraznih bolesti.

Ljudska prava i radne norme

Od odgovornih multinacionalnih kompanija koja posluju širom svijeta očekuje se poštivanje ljudskih prava i socijalne pravde te uklanjanje bilo kakvih oblika prisilnog rada i iskorištavanja djece. Uvjeti rada moraju biti utemeljeni na međunarodno opće prihvaćenim normama "International Labour Organisation“. Temeljni dokument o ljudskim pravima je Međunarodna deklaracija o ljudskim pravima donešena od strane UN-a, 10. prosinca 1948. Tekuće odrednice propisane su rezolucijom 2003/16 UN-ove Komisije za promociju i zaštitu ljudskih prava. Radi se o dokumentu "United Nations Norms on the Resposibilities of Transnational Corporatons and other business Enterprises with Regard to Human Rights" odobrenom 13. kolovoza 2003.

Zaštita zdravlja

Jedan od zadataka poslovnog sektora je buđenje svijesti radnika o zdravstvenim pitanjima i sigurnosti na radnom mjestu. Svjetska Zdravstvena Organizacija djeluje na rješavanju ključnih zdravstvenih problema te stvaranju regulative i osvještavanju privatnog sektora kao važnog dionika u razvoju tehnologija i usluga potrebnih da bi se s ovim problemima stalo na kraj.

Društvena pitanja

Tipično djelovanje DOP na području društvenih pitanja kanalizirano je kroz: podršku razvoju društvenih resursa kroz know-how pristup; podršku zavodima za zapošljavanje te doprinos zapošljavanju osoba s posebnim potrebama te kroz poticanje zaposlenika na aktivno sudjelovanje u zivotu zajednice, poticanje volontiranja i pružanje podrške akcijama lokalne zajednice za promicanje zdravog načina života, suradnja za zdravstvenim organizacijama.

C. Ekološka matrica DOP-a

Pregled najvažnijih odgovornosti na području ekologije je predstavljen tablicom III.

TABLICA III. EKOLOŠKA MATRICA DOP-A [5]

Ekološka matrica DOP-a	
Zadatak	Akcije
Proaktivna podrška u pitanjima zaštite okoliša	• Obuka zaposlenika i podizanje svijesti o okolišu • Suradnja sa lokalnom zajedicom na razvojnim programima zaštite okoliša • Istraživanje na području obnovljivih izvora energije i smanjenje štetnih emisija
Upravljanje rizicima u ekologiji	• Politike i procesi upravljanja rizicima • Odgovornost, sprečavanje ekoloških katastrofa
Zaštita okoliša	• Učinkovita uporaba sirovina • Upravljanje energijom • Upravljanje utjecajem na vode • Upravljanjem biodiverzitetom • Minimizacija otpada/emisija/otpadnih voda • Gospodarenje otpadom • Minimizacija štetnog utjecaja proizvoda na okoliš

U današnje vrijeme od poduzeća se očekuje integracija odgovornosti za okoliš na svim razinama poslovanja: iznalaženjem održivih rješenja u iskorištavanju prirodnih resursa; smanjenjem otpada, zagađenja i štetnih emisija; poboljšanjem učinkovitosti i produktivnosti na svim poljima, pogotovo na području upravljanja vodom, energijom i sirovinama.

IV. POSLOVNE PERFORMANSE

Poslovne performanse se mogu najednostavnije definirati kao radni učinci bilo pojedinca, tima ili cjelokupne oragnizacije; mjerenjem kojih dolazimo do ocjene jesu li i u kolikoj mjeri ostvareni ciljevi i planovi definirani strategijom organizacije.

Mjerilo performansi i rezultati do kojih se tim mjerenjem dolazi, su osnova za ocjenu uspješnosti poslovanja organizacije u proteklom razdoblju, kako za management, vlasnike i zaposlenike organizacije, tako i za vanjske stakeholdere. Organizacijska mjerila performansi se mogu posmatrati iz više perspektiva; (Armstrong, 2006) s obzirom na područja mjerenja koja pokrivaju, mogu se podijeliti u tri temeljne skupine: financijska mjerenja performansi (kao npr. profiti, prodaja, povrat na investirani kapital, dodana ekonomska vrijednost, zarada po dionici,...), operativna mjerila performansi (npr. razina usluge kupca, zadovoljstvo kupaca, kvaliteta, kontrola zaliha, broj isporuka na vrijeme,...) te mjerila performansi vezane za zaposlenike (npr. profit, prodaja ili dodana vrijednost po zaposleniku, troškovi plaća kao postotak prodaje, produktivnost zaposlenika, zadovoljstvo zaposlenika,...) [6]. U literaturi se može naći i jednostavnija podjela na financijska i nefinancijska mjerila poslovnih performansi [7].

FINANCIJSKA MJERILA

Neto operativna marža je omjer profitabilnosti koji se izračunava dijeljenjem neto dobiti sa ukupnim prihodima. Pokazuje nam koliki je neto profit ostvaren od ukupno realiziranog posla na tržištu, a kojim možemo slobodno raspolagati. Taj dio prihoda koji poprima karakter dobiti poslije oporezivanja poduzeće može isplatiti vlasnicima ili pak može dio iznosa ostaviti u bilanci kao zadržani dobitak.

Neto profitna marža = neto dobit / ukupni prihodi

ROA (Return on Assets) ili povrat na ukupnu imovinu je izražena u postocima i jednostavno rečeno pokazuje kako efikasno neko poduzeće zarađuje novac. ROA može pomoći i kod procjene profitabilnosti nekih novih projekata koje poduzeće planira, naime novi projekt je profitabilan ako je ROA veći od iznosa kamatnih stopa za kredite.

ROA= neto dobit / ukupna imovina

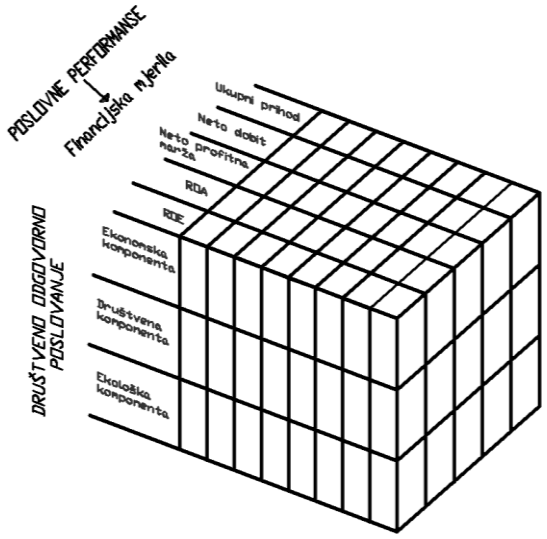
ROE (Return on Equity) ili povrat na vlastiti kapital pokazuje kolika je dobit po kuni uloženog kapitala. ROE je i dobar pokazatelj brzine rasta poduzeća jer se smatra da ukupni prihodi ne mogu rasti po stopi većoj od trenutnog iznosa ROE, osim zaduživanjem kod banaka. Obično se smatra da je ROE od 15 % prosjek za dobro poduzeće.

ROE= neto dobit / vlastiti kapital

ROA, ROE, ROI, EPS smatraju se klasičnim mjerilima performansi poduzeća koja više nisu adekvatna pa im se traže supstituti. Smatra se da ti pokazatelji zasnovani na računovodstvenim podacima nisu pouzdani indikatori vrijednosti korporacijske ekonomije. Stoga mnoga poduzeća raspolažu sa novim mjerilima poput: dodana vrijednost, ROVA i ROVA/ROI, MVA i EVA te bogatstvo dioničara. Zbog dostupnosti financijskih podataka sa web stranice (http://rgfi.fina.hr/JavnaObjavaweb/jsp/prijavaKorisnika.jsp), gdje su bez naknade dostupni godišnji financijski izvještaji te druga dokumentacija koju su, prema članku 20. Zakona o računovodstvu, poduzetnici dužni dostaviti FINI radi javne objave, donesena je odluka o iskazivanju performansi poduzeća preko klasičnih mjerila performansi poduzeća (ukupni prihod, neto dobit, neto profitna marža, ROA, ROE).

V. MODEL POVEZANOSTI STUPNJA RAZVIJENOSTI DRUŠTVENO ODGOVORNOG POSLOVANJA I POSLOVNIH PERFORMANSI PODUZEĆA

Istražit ćemo kako društveno odgovorno poslovanje, s obzirom na svoju razvijenost u poduzeću utječe na poslovne performanse samog poduzeća. Do odgovora na ovo pitanje se dolazi međusobnim križanjem navedenih elemenata razvijenosti društveno odgovornog poslovanja sa pojedinim elementima poslovnih performansi poduzeća uzimajući u obzir financijska mjerila. Rezultat međusobnog križanja navedenih elemenata je 3x5 matrica koja će kad se ubace rezultati istraživanja dati jasnu sliku traženog utjecaja. Navedenu matricu su sami izradili autori.



Slika 1. Model (matrica) povezanosti stupnja razvijenosti društveno odgovornog poslovanja i poslovnih performansi poduzeća

VI. REZULTATI EMPIRIJSKOG RAZMATRANJA

Kao uzorak istraživanja uzeta su velika hrvatska poduzeća budući da je u praksi pokazano da društveno odgovorno poslovanje najčešće primjenjuju pretežno velika poduzeća koja su najčešće motiviranija i posjeduju resurse za strukturiranje, praćenje i izvještavanje o njihovoj društveno odgovornoj praksi. Kao kriterij za ulazak u osnovni skup „velikih poduzeća“ koristio se Zakon o računovodstvu (NN br. 109/07, članak 3. stavak 3.,4.) te registar poslovnih subjekata Hrvatske gospodarske komore [8] koji sadrži bazu podataka svih hrvatskih trgovačkih društava registriranih na području Republike Hrvatske. Iz istraživanja su isključena sva poduzeća nad kojima je tijekom istraživanja otvoren postupak stečaja ili likvidacije, što je utvrđeno na temelju podataka Hrvatske gospodarske komore. Uzorak nije obuhvatio javna lokalna i komunalna poduzeća.

Kao glavni instrument za prikupljanje primarnih podataka odabran je anketni upitnik. Istraživanje je provedeno na uzorku od 436 velikih poduzeća u Hrvatskoj (obuhvaćena su poduzeća koja su predala godišnje financijsko izvješće). Anketni upitnik je oblikovan na 2 stranice, sastoji se od 20 pitanja, grupiranih u dvije logičke cjeline: osnovni podaci o poduzeću te društveno odgovorno poslovanje- cjelina u kojoj se tražilo da anketirana osoba pruži uvid u kolikoj mjeri je razvijen DOP u anketiranom poduzeću kroz sve elemente triju matrica DOP-a. Pitanja su zatvorenog tipa s ponuđenim odgovorima inteziteta, kod kojih je korištena diskretna Likertova skala sudova s pet stupnjeva inteziteta, koja se i teoretski smatra najboljim rješenjem prilikom izrade anketnih upitnika (Zelenika, 1998).

Primarni podaci, dobiveni preko anketnog upitnika, te samostalno obrađena financijska mjerila korištenjem financijskih izvještaja dostupnih na FINI, obrađivani su preko računalnog programa za statističku obradu podataka SPSS 17.0. Korišten je i Microsoft Excel.

Rezultat istraživanja elektronskim putem odaslanih anketnih upitnika, je vraćenih 76 ispravno popunjenih upitnika, ili 17,4% vraćenih od ukupnog broja odaslanih anketnih upitnika što je očekivana ili nešto iznad očekivane stope povrata za ovakvu vrstu istraživanja. Analizom je obuhvaćeno 69 poduzeća iz istraživanog uzorka dok za 7 poduzeća nije bilo dostupno ime poduzeća pa se zbog tog razloga nije mogao dobiti uvid u financijsko izvješće iz kojeg bi se mogla izračunati financijska mjerila.

Donijet je finalni zaključak je li i koliki je stupanj razvijenosti društveno odgovornog poslovanja u istraživanim poduzećima. U slijedećoj tablici rezimirani su svi sumarni rezultati razvijenosti DOP-a po pojedinim elementima u istraživanim poduzećima.

TABLICA IV. UKUPNA RAZVIJENOST DRUŠTVENO ODGOVORNOG POSLOVANJA U ISTRAŽIVANIM PODUZEĆIMA

	N		Std. Error				Std.	Percentiles		
	Valid	Missing	Mean	of Mean	Median	Mode	Deviation	25	50	75
Ekonomska komponenta DOP-a	74	2	3.52	0.10	3.5	3	0.872	3	3.5	4.2
Društvena komponenta DOP-a	73	3	3.78	0.101	4	4.17	0.866	3.33	4	4.33
Ekološka komponenta DOP-a	71	5	3.61	0.101	3.6	3,4 (a)	0.854	3	3.60	4.2
Ukupna razvijenost DOP-a	69	7	3.63	0.099	3.66	4.8	0.824	3.1167	3.6556	4.2444

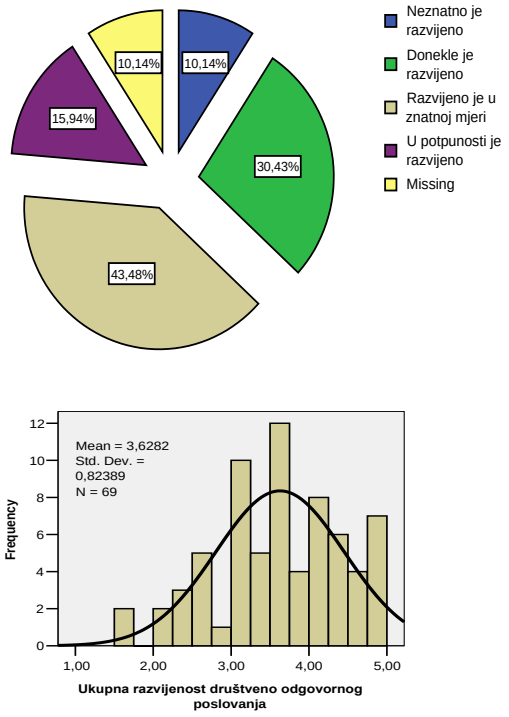
a. Multiple modes exist. The smallest value is shown

Iz tablice IV je vidljivo da sve komponente društveno odgovornog poslovanja imaju sve srednje vrijednosti [9] približno iste. Od spomenutih triju komponenata, s obzirom na njihove srednje vrijednosti, ekonomska komponenta je ta koja se ističe nešto nižim vrijednostima (aritmetička sredina 3,52, medijan = 3,5 te mod = 3). Iz svega navedenog može se zaključiti da je društvena komponenta na nešto višoj razini u odnosu na ostale dvije komponente, ekonomsku i ekološku komponentu te da je ekološka komponenta nešto više razvijenija od ekonomske. Može se dati zaključak da su sve tri komponente prije svega razvijene te da su ekonomska i ekološka komponenta razvijene između donekle do u znatnoj mjeri dok je društvena komponenta razvijena u znatnoj mjeri. Sve to govori o tome da poduzeća vrlo dobro ulažu u razvijanje svih komponenata društveno odgovornog poslovanja pridajući im donekle do znatnu važnost.

Sveukupno gledajući razvijenost društveno odgovornog poslovanja u istraživanim poduzećima, njena aritmetička sredina iznosi 3,63. Medijan, druga srednja vrijednost iznosi 3,66, tj. prosječno gledano društveno odgovorno poslovanje je između donekle do razvijeno u znatnoj mjeri, s tim da ova srednja vrijednost je ipak puno bliže razvijenosti u znatnoj mjeri. Treća srednja vrijednost, tj. mod ima vrijednost 4,8, odnosno 4, ako gledamo ovu varijablu kao grupiranu varijablu, što bi značilo da najveći broj istraživanih poduzeća u uzorku ima znatnu razvijenost društveno odgovornog poslovanja.

Slika koja prikazuje distribuciju istraživanih poduzeća, s obzirom na sveukupnu razvijenost društveno odgovornog poslovanja ukazuje na distribuciju koja je ipak blago pomaknuta prema desno, odnosno prema vrijednosti 4 (razvijenosti u znatnoj mjeri). Sa slike je također vidljivo da oko 43,48% istraživanih poduzeća ima razvijen DOP u znatnoj mjeri. Naime, skoro 60% istraživanih poduzeća je razvilo DOP u znatnoj mjeri do u potpunosti, u odnosu na oko 40% istraživanih poduzeća koji su razvili DOP neznatno do donekle.

Imajući u vidu upravo iznesene tvrdnje pojedinih komponenata, srednje vrijednosti ukupne razvijenosti društveno odgovornog poslovanja, te strukturu distribucije istraživanih poduzeća po pojedinim intezitetima razvijenosti društveno odgovornog poslovanja, možemo zaključiti da je društveno odgovorno poslovanje u istraživanim poduzećima znatno razvijeno. To smatramo dobim rezultatom za poduzeća u Republici Hrvatskoj koja na kraju rezultatski gledano ulažu u razvoj društveno odgovornog poslovanja u zadovoljavajućoj (znatnoj) mjeri.



Slika 2. Ukupna razvijenost društveno odgovornog poslovanja

Slijedeća analiza prikazat će rezultate temeljene na korelacijskim testovima. Korelacija označava vezu između dvaju obilježja (varijabli) kod koje je past/pad jednog obilježja popraćen u većoj ili manjoj mjeri rastom/padom drugog obilježja, ovisno o predznaku i veličini koeficijenta korelacije [10]. Korelacijski testovi, osim što daju odgovor da li postoji korelacija i koliki je koeficijent korelacije, također odgovaraju na pitanje da li je ta korelacija statistički značajna. [10]. S obzirom na činjenicu da je analiza obuhvaćala parove redoslijednih i numeričkih obilježja, korišten je Spearmanov koeficijent korelacije, primjereniji za parove obilježja ovakvog tipa. [10]. Ukoliko je $\alpha \geq 0,05$ to znači da ne postoji statistički značajna korelacija između dviju varijabli, dok ukoliko je $\alpha < 0,05$ to znači da postoji statistički značajna korelacija između dviju varijabli. U ekonomiji koeficijenti

korelacije (r) govore o: 0< r < 0,5 => slaba korelacija; 0,5< r < 0,8 => srednje jaka korelacija; 0,8< r < 1 => jaka korelacija.

Analizom podataka iz tablice V., korištenjem Spearmanova koeficijenta korelacije, može se zaključiti da između svih komponenti društveno odgovornog poslovanja (kao i ukupne razvijenosti društveno odgovornog poslovanja) i svih poslovnih performansi (financijskih mjerila) u istraživanim poduzećima su skoro sve korelacije savršene što se tiče statističke značajnosti ($\alpha \approx 0{,}000$), odnosno statistički značajne na razini povjerenja od 99% osim savršenosti kod povezanosti ekološke komponente i ukupne razvijenosti društveno odgovornog poslovanja sa ukupnim prihodom gdje postoji statistički značajna korelacija ($\alpha \leq 0{,}05$) na razini povjerenja od 95%.

Koeficijenti korelacije statističkih značajnih korelacija su pozitivni, te se većinom kreću od 0,270 do 0,505 (0,270 < r < 0,505) što govori o intezitetu slabo pozitivne korelacije. Možemo zaključiti da u ovisnosti o rastu ukupne razvijenosti društveno odgovornog poslovanja, kao i njenih komponenata, kod istraživanih poduzeća, rastu i performanse istih tih poduzeća.

TABLICA V. POVEZANOST RAZVIJENOSTI DRUŠTVENO ODGOVORNOG POSLOVANJA SA PERFORMANSAMA PODUZEĆA

			Ukupni prihod	Neto dobit	Neto profitna marža	Povrat na ukupnu imovinu	Povrat na vlastiti kapital
Spearman's rho	Ekonomska komponenta društveno odgovornog poslovanja	Correlation Coefficient	**0,287****	**0,438****	**0,398****	**0,426****	**0,384****
		Sig. (1-tailed)	**0.009**	**0.000**	**0.000**	**0.000**	**0.001**
		N	67	67	67	67	67
	Društvena komponenta društveno odgovornog poslovanja	Correlation Coefficient	**0,288****	**0,488***	**0,488****	**0,505****	**0,505****
		Sig. (1-tailed)	**0.010**	**0.000**	**0.000**	**0.000**	**0.000**
		N	66	66	66	66	66
	Ekološka komponenta društveno odgovornog poslovanja	Correlation Coefficient	**0,270***	**0,422****	**0,358****	**0,458****	**0,407****
		Sig. (1-tailed)	**0.016**	**0.000**	**0.002**	**0.000**	**0.000**
		N	64	64	64	64	64
	Ukupna razvijenost društveno odgovornog poslovanja	Correlation Coefficient	**0,278***	**0,429****	**0,394****	**0,452****	**0,407****
		Sig. (1-tailed)	**0.014**	**0.000**	**0.001**	**0.000**	**0.001**
		N	62	62	62	62	62

** Correlation is significant at the 0.01 level (1-tailed).
* Correlation is significant at the 0.05 level (1-tailed).

VII. ZAKLJUČAK

Rezultati provedenog teorijskog i empirijskog istraživanja ukazuju slijedeće:

- Društveno odgovorno poslovanje je u istraživanim poduzećima znatno razvijeno. Gledajući po pojedinim dimenzijama razvijenosti društveno odgovornog poslovanja, prednjači društvena komponenta koja je razvijena u znatnoj mjeri, dok ostale dvije dimenzije (ekonomska i ekološka komponenta) imaju nešto nižu, odnosno djelomičnu (donekle) do znatnu prisutnost u istraživanim poduzećima.
- Postoji statistički značajna korelacija (povezanost) između ukupne razvijenosti društveno odgovornog poslovanja (i svih komponenata) u istraživanim poduzećima i svih poslovnih performansi (mjerenih kroz financijska mjerila) u navedenim poduzećima.
- Slaba pozitivna korelacija je prisutna između ukupne razvijenosti društveno odgovornog poslovanja (i svih komponenata) u istraživanim poduzećima i svih poslovnih performansi.

Sveukupno gledajući rezultate teorijskog i empirijskog istraživanja razvijenosti društveno odgovornog poslovanja i poslovnih performansi poduzeća dokazano je da razvijenost društveno odgovornog poslovanja u poduzećima, sa svojim pojedinim dimenzijama (i podelementima) pozitivno utječe na poslovne performanse tih istih poduzeća čime je dokazana hipoteza.

Rezultat ovog istraživanja (uz ograničenja s obzirom na veličinu uzorka, vrijeme provedbe istraživanja, nedostatke metode anketnog upitnika, nedostatka definiranja sveobuhvatne literature o društveno odgovornom poslovanju) će, nadamo se, biti od koristi svima koji se bave problematikom implementiranja društveno odgovornog poslovanja, s obzirom da je napravljeno vrlo malo istraživanja u praksi koja bi mogla potvrditi traženu povezanost sa poslovnim performansama. Iako je u Hrvatskoj prepoznata važnost društveno odgovornog poslovanja malo je onih koji ga pravilno integriraju u poslovne procese. U vremenu kada je sve više dokaza kako društveno odgovorno poslovanje nije uzaludan trošak nego direktna investicija u konkurentnost, kada on prestaje postojati samo na dobrovoljnoj bazi, poduzeća sve više prepoznaju da nešto moraju učiniti, ali ne znaju i što. Zbog toga se nadamo da će naš rad biti korak u smjeru prema prepoznavanju nužnosti implementacije društveno odgovornog poslovanja, ali i upozorenje svima nama da i sami možemo sudjelovati u stvaranju klime koja će pogodovati svim dionicima društva.

LITERATURA

[1] P. Kotler, i N. Lee, *Društveno odgovorno poslovanje: suvremena teorija i najbolja praksa*. Zagreb, M.E.P Consult, 2009.

[2] J. Nelson, *Building Competitiveness and Communities*. London: Prince of Wales Business Leaders Forum, 1998.

[3] A. Bagić, M. Škrabalo, i L. Narančić, *Pregled društvene odgovornosti poduzeća u Hrvatskoj*. Zagreb: Academy for Educational Development, 2004.

[4] Archie B. Carroll, "*The Pyramid of Corporate Social Responsibility: Toward the Moral Management of Organizational Stakeholders*," Business Horizons, July-August, 1991.

[5] P. Katsoulakos, T. Sapouna, i D. Zevgolis, *CSR Classification and Corporate Sustainability Indicators*, 2004 [White Paper CA1B, Version 1.0].

[6] M. Armstrong, *Performance management: key strategies and practical guidelines*, 3rd ed. London: Kogan Page Ltd, 2006.

[7] M. Buble, *Osnove menadžmenta*. Zagreb: Sinergija, 2006.

[8] Registar poslovnih subjekata. Dostupno na: http://www1.biznet.hr/HgkWeb/do/extlogon

[9] S.Pivac, i A. Rozga, *Statistika za sociološka istraživanja*. Split: Sveučilište u Splitu, Filozofski fakultet, 2006.

[10] A. Rozga, *Statističke analize u znanstvenom istraživanju*. Split: Sveučilište Splitu, Ekonomski fakultet, 2010. [Predavanje]

WEB APPLICATIONS & M2M COMMUNICATIONS

Web application functional size estimation based on COSMIC method and UWE approach

Denis Čeke*, Marijan Đurek **, Suad Kasapović***
* University of Tuzla/Technical Department, Tuzla, Bosnia and Herzegovina
** University of Zagreb, Faculty of Electrical Engineering and Computing/Department of Applied Computing, Croatia
***University of Tuzla, Faculty of electrical engineering/Department of communications, Tuzla, Bosnia and Herzegovina
*denis.ceke@untz.ba , **marijan.djurek@fer.hr, ***suad.kasapovic@untz.ba

Abstract - In the last years measuring functional size of software applications became one of the most popular methods when there is need for creating models for effort, costs and budget estimation in software development. Functional Size Measurement (FSM) methods quantify the software from its user's perspective, disregarding quality and technical criteria. In context of conceptual models where models represent application at higher level of abstraction, functional size of developed software application can be measured on the basis of conceptual model and requirements analysis. Requirements analysis is usually conducted thru usage of use case, sequence, activity and other UML diagrams. In this work, we have investigated possibility of using COSMIC FSM method for functional size estimation of web applications whose basic model is created using UML-based Web Engineering (UWE) approach. An example of developed web application with UWE requirements model used for COSMIC measurement is presented and discussed.

I. INTRODUCTION

Web application development is moving toward application development based on model. This kind of development brings new advantages to project managers to make early estimation of cost and size of future web applications. Nature of web projects forces managers to focus primarily on time variable to achieve the best possible solution that have best ratio of cost and project requirements.

To succeed in this, in the last couple of years, researchers have tried to find the best solution to calculate web applications size and development effort based on calculating number of web pages, links, animations, pictures etc [1]. These solutions required projects to be finished to be able to calculate some parameters on whose bases they have than calculated development effort or created budget models.

It is important to state that there is no "*silver bullet*" method for development effort estimation in web development.

To this day, the main challenge that remains in this area is how to estimate web applications size or development effort before application is even developed.

The possible answer to this question lies in the assumption that web application development effort depends from the functional size of web application. This may be the truth, because, the more complex application is, the more effort is needed to invest to develop it. This possible hypothesis was already tested by some researchers [2]–[6] , and they showed that this dependency exists.

Functional size of software is defined as "size of the software derived by quantifying the Functional User Requirements" [7]. Functional User Requirements (FUR) describes what the software is expected to do for its users [8]. Examples are data transfer, data transformation, data storage, and data retrieval. Functional size is independent of software language and development methods.

Several Functional Size Methods (FSM) exists that could be used for functional size estimation of software. Among others, the most popular are: IFPUG FPA [9], MK II FPA [10], NESMA FPA [11], and COSMIC FFP [7]. These FSM have been successfully used in software application development estimations, but there are only few examples of their implementation in web projects. A literature survey covering several methods can be found in [12].

COSMIC (Common Software Metrics International Consortium) [7] measurement method can be applied to any kind of software and it allows application measurement on multi-layer bases, that other functional size methods are not able to do. Because of this fact, COSMIC FSM is used for specifying measurement approach that could be applied on conceptual models.

For development of web application conceptual models number of methods exists (OOHDM [13], WebML [14], OO-H [15], W2000 [16] and UWE [17]). UWE approach was one of the first projects which were used specially for web applications. It is UML based tool with stereotypes specially created for web applications. UWE is supported thru CASE tool like ArgoUWE [18],

and now exists and MagicUWE [19] as plug-in for MagicDraw tool [20].

In [21] there is an exhausting survey of methods for estimating functional size of software applications based on models and COSMIC measurement method.

In research by [2] approach for effort estimation in web application development was presented based on COSMIC method. This approach is based on [22], for identifying rules that were used for COSMIC measurement procedure in context of web hypermedia applications.

COSMIC-FFP-like approach for calculation of functional size of web applications based on COSMIC method and hypermedia elements was presented in [23].

Our approach goes one step ahead. We have developed measurement procedure that uses COSMIC method for calculating data movements and UWE tool for creating user requirements model based on use case and activity diagrams. Using this approach we have investigated possibility of calculating sum of the COSMIC Function Points (CFP), which are basic numerical value in COSMIC method, in the early phase of software development life cycle. Most of approaches that we have analyzed, beside use case and sequence diagrams, had to include class diagrams to present mapping of data that is entered in the system and that were used for mapping to persistent storage. This was necessary because all diagrams used in analyzed approaches had no stereotypes with enough semantics to describe data movement as inputs from users forms to persistent storage and that were required in use case, sequence or other diagrams in the analyses or design phase. On the contrary, UWE approach has possibility to use this kind of stereotypes in the early requirements phase thru activity diagrams.

II. COMPONENTS USED FOR CREATING MODEL

Our model is composed from two major components: UWE approach for modeling web applications and COSMIC measurement method.

A. UWE approach

UWE was one of the first projects that were created especially for modeling web applications that uses pure UML profile, extended with stereotypes for web applications. Another important extension is possibility to create navigation and presentation design, which are one of the major elements in web applications. Basic types of artifacts in UWE are: Requirements model, Conceptual model, Navigation model, Presentation model and Process model.

UWE is domain specific modeling language (DSML), which provides the annotations needed to enrich standard requirements models with web features and to reduce model-complexity. The models specified with the DSML are used in the model transformations of the model-driven engineering (MDE) process to generate the design models of the web applications [24].

This approach comprises a notation, a method and tool support. The notation is defined as a UML profile [25], i.e. using the extension mechanism provided by the UML itself, which allows for the refinement of UML in a strictly additive manner by stereotypes, tag definitions and constraints, providing the required additional annotations [24]. The cornerstones of the UWE method are the principle of separation of concerns and a model-driven approach. Tool that supports UWE approach is the MagicUWE [19], plug-in implemented for Magic-Draw [20] and the UWE4JSF eclipse plug-in for the code generation.

In the design phase, UWE modeling approach follows the principle of "separation of concerns" building separate models for views of the navigation, content, presentation, processes, etc. in the same way other web development methods do, such as OOHDM [26], OOHRIA [27] and WebML [28], among others. The set of model types is a highly flexible and modular modeling framework providing the basis for the model-driven engineering (MDE) development process.

A set of model transformations is implemented in MagicUWE to benefit from the efforts invested in the requirements models and to produce initial versions of all design models, i.e. content, navigation and presentation models [24]. The set of transformations that are exploited thru this tool, are:

- requirements to content,
- requirements to navigation,
- requirements to process, and
- requirements to presentation.

Detailed and well explained usage of elements and transformation is presented in [24] and also can be find on the project web page [19].

B. COSMIC method

COSMIC measurement method is widely used for software size estimation and it is approved as ISO standard (ISO/IEC 19761) since March 2003. This method was originally designed for estimation of business systems and real-time applications, but it is applicable in any other context where in some software solution dominates large number of data movements [7]. COSMIC method measures data movements from/to software process across boundary that is defined by the application and COSMIC rules. To be able to apply COSMIC method, it is necessary to separate a software application, which is to be measured, from its operational environment. Potential of COSMIC method for being applied accurately in the requirements specification phase compared to other FSM methods is demonstrated by the study [29]. COSMIC does not rely on subjective decision by the functional size measurer during the measurement process [7]. Thus, its measurements, taken from well specified requirements, tend to be the same among multiple measurers [30].

In COSMIC, size is measured in terms of the number of data movements, which accounts for the movement of one or more data attributes belonging to a single data group. A data group is an aggregated set of data attributes. A functional process, in COSMIC, is an independently executable set of data movements that is triggered by one or more triggering events. A triggering event is initiated

by an actor (a functional user or an external component) that occurs outside the boundary of the software to be measured. Thus, a functional process holds the similar scope of a use case scenario, starting with the triggering event of a user request and ending with the completion of the scenario [30]. Fig. 1 illustrates the generic flow of data groups from a functional perspective, presented in the COSMIC standard [7].

COSMIC recognizes four types of data movements:

- Entry - moves data from functional users into the functional process;
- Exit - moves data from the functional process, across boundary to the functional users;
- Read - moves data from persistent storage to the functional process;
- Write - moves data from the functional process to the persistent storage;

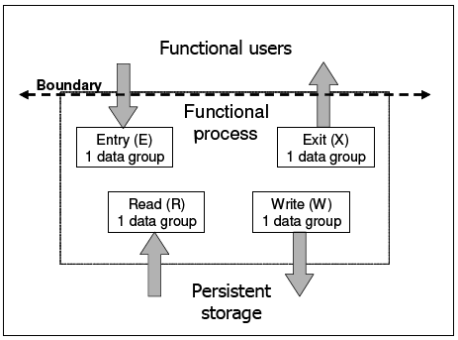


Figure 1. Four types of data movement and their relationship with functional processes and users [7]

COSMIC counts each of these data-movements as one CFP of functional size, and measures the size of each of the functional processes separately. It then adds the sizes of all the functional processes to compute the total size of the system to be measured. COSMIC offers an objective method of measuring functional size.

To use COSMIC measurement method, it is necessary to go thru tree major phases: measurement strategy phase, mapping phase and measurement phase [7].

After that, it is possible to determine functional size of software and calculate number of CFP for project.

For some functional process *i*, total functional size is a sum of all data movements that occurs in that process:

$$\text{Size}(FP_i) = \Sigma\text{size}(\text{Entries}_i) + \Sigma\text{size}(\text{Exits}_i) + \Sigma\text{size}(\text{Reads}_i) + \Sigma\text{size}(\text{Writes}_i) \quad (1)$$

Size of software in terms of COSMIC is then sum of sizes of all functional processes that occurs in measured software application:

$$\text{Size}(\text{software}) = \Sigma\text{Size}(FP_i) \quad (2)$$

III. COSMIC MEASUREMENT FOR WEB APPLICATIONS

Dynamic web applications are mainly characterized by a lot of data movement (from users to persistent storage and vice versa). Because of this fact, they are potential candidate suitable to be measured with COSMIC approach [31]. In order to apply method, software and context model of COSMIC method need to be adapted like it is presented in [2] in accordance with UWE approach. Software context model adapted for COSMIC for sizing hypermedia systems used in [22] is presented on Fig. 2. In this model Web Server is considered as data store.

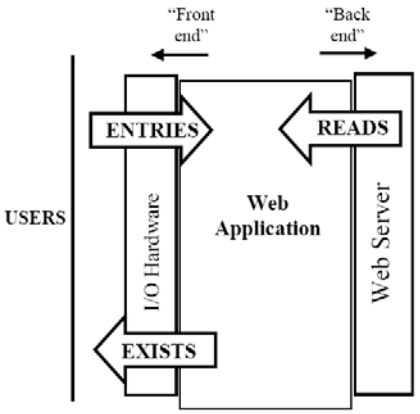


Figure 2. The functional flow of data attributes through Web applications used in [22]

Our approach is based on generic software model presented in [2] except that instead of use case and class diagrams for defining requirements and *Conalen's* model for presentation and process flow diagrams, we have used use case and activity diagrams for requirements with special stereotypes for web applications. These stereotypes are main elements that will be used for defining rules and counting procedure.

The general COSMIC measurement process consists of three phases:

- Measurement Strategy, in which the Software Context Model is applied to the software to be measured,
- Mapping Phase in which the Generic Software Model is applied to the software to be measured,
- Measurement Phase, in which actual size measurements are obtained.

The result of applying the measurement process to a piece of software is a functional size measure of the FUR of the piece of software expressed in 'COSMIC Function Points' (or 'CFP') [7].

A. *Measurement Strategy phase*

This phase includes four steps: purpose and scope of measurement, identification of functional users and level of granularity. The purpose and scope of measurement would be e.g. a measurement of size of the web-based applications to support effort estimation in web development. Functional users of web applications can be

human users or other peer components (identified as actors thru use case diagrams) that can interact with web application. For software measurement, there is only one standard level of granularity that it is possible to define unambiguously. That is the level of granularity at which individual functional processes have been identified and their data movements defined [32].

B. Mapping phase

There are two major elements that have to be identified in this phase: Identification of functional processes and data groups. In web-based applications, functional process can be defined as elementary if it is triggered by an event sent by a user of the system (actor) or by the occurrence of temporal event [23].

Functional processes in our approach in accordance with UWE are defined thru use case diagrams which can have two types of stereotypes: *<<browsing>>* and *<<processing>>*. These use case diagrams are then mapped into activity diagrams that perform desired actions in accordance with FUR. Data groups in our approach are defined as attribute values that are send to or received from activity diagram stereotypes *<<user action>>*, *<<system action>>*, *<<display action>>* or *<<central buffer node>>*. Stereotype *<<user action>>* presents actions that have been initialized by the user of software. Stereotype *<<system action>>* presents actions that are executed by the server side processes. Stereotype *<<central buffer node>>* presents persistent storage which enables a functional process to store a data group beyond the life of the functional process and/or from which a functional process can retrieve a data group stored by another functional process, or stored by an earlier occurrence of the same functional process, or stored by some other process.

C. Measurement Phase

In this phase we perform identification of data movements and aggregation of measurement results.

Identifying data movements; this step consists of identifying data movements for each functional process on the basis of arrows that present object or control flow in activity diagrams. If we have object flow arrow, that means that we have potentially some kind of data movement. On the other hand, control flow arrow presents commands execution sequence inside of functional process. In UWE approach, these kinds of movements are identified based on the source/destination elements of UWE activity diagram elements:

- If the source element if *<<user action>>* stereotype element and destination element is *<<system action>>*, than we identify 1 Entry data movement.
- If *<<system action>>* is a source element and destination stereotype element is *<<central buffer node>>* than we identify 1 Write data movement.
- If *<<central buffer node>>* is source element and destination element is *<<system action>>* than we identify 1 Read data movement.
- If *<<system action>>* is source element and destination stereotype element is *<<user action>>*, we identify 1 Exit data movement.
- If *<<system action>>* is source element and destination stereotype element is *<<display action>>*, we identify 1 Exit data movement.
- If *<<system action>>* is source element and destination element is *<<activity final node>>*, we identify 1 Exit data movement.

Aggregating of measurement results; in this step we sum the results of all data movements in all functional processes (equation 2). Functional process is presented with use case diagram explained thru activity diagram.

Functional size of each identified object is equal to the sum of all data movements identified inside that object (equation 1).

IV. EXAMPLE APPLICATION

In our example we have simple web application for management of university curriculum register. FUR is presented with use case and activity diagrams. This system has one user who is presented as actor in use case diagram and system to be measured is separated from the user by the boundary (Fig. 3). Basic use case diagrams are: *Login*, *Logout*, *View Study Programs*, *View Curriculum*, *View Curriculum Details*, *View Subjects* and *View Subject Details*. Some of the use cases have extension points which allow us detailed explanation of FUR. Process flow inside use case diagrams is presented with activity diagram, who presents functional process, on whose bases we calculate CFP (Fig. 4).

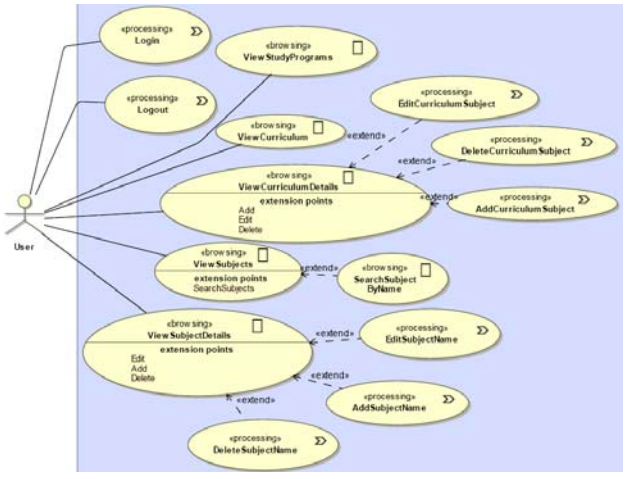


Figure 3. Functional requirements modeled with use cases

User of this web application is first authorized thru *Login* process. After successful authorization, he has possibility to e.g. view curriculums, edit them, add new etc. As example, we will show actions for *Login* process. User send his user name and password thru UWE stereotype *<<user action>> Input* toward system into UWE stereotype *<<system action>> CheckUserPass*. We identify 1 Entry. This UWE stereotype *<<system action>> CheckUserPass* checks if the user credentials are correct against persistent storage element presented with *<<central buffer node>> Users*. We identify 1

Read. If the user is authorized, <<*system action*>> *WriteLogForUser* writes log for logged user to persistent storage <<*central buffer node*>> *LogUsers*. We identify 1 Write. And finally, we identify 1 Exit for error or confirmation message about unsuccessful of successful completed functional process, respectively. At the end, for *Login* process, we identify 4 CFP.

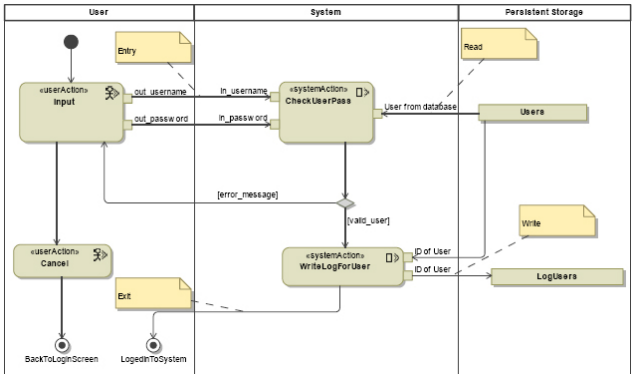


Figure 4. Example of workflow represented as UML activity diagram for Login process

V. EXPERIMENTAL RESULTS FOR CFP COUNT

After analyze of use case and activity diagrams in accordance with proposed procedure, in Table I results of counting data movements and calculated CFP for example web application are presented.

From the preformed analyze, number of CFP for each functional process and percentage of influence of every functional processes CFP in total sum of CFP, we can conclude that the most complex functional process are *Add Curriculum Subject* and *Edit Curriculum Subject*. They are characterized with a lot of data movements which is consequence of a large number of attributes which forms one curriculum subject. Operations like add or edit are usually one of the most complex operations especially when there is need for consistency checking before they are conducted. On the basis of this requirement analysis we can conclude that e.g. the most development effort for developing of this simple web application will be invested in the creation of these two functional processes.

TABLE I. FUNCTIONAL PROCESSES WITH CFP COUNT AND PRECENTAGE OF INFLUENCE

No.			
	Functional process	***CFP***	***%***
1	Login	4	3.95
2	Logout	3	5.26
3	View Study Programs	7	9.21
4	View Curriculum	7	9.21
5	Add Curriculum Subject	12	15.79
6	Edit Curriculum Subject	10	13.16
7	Delete Curriculum Subject	7	9.21
8	View Subjects	4	3.95
9	Search Subject by Name	4	3.95
10	Add Subject Name	4	3.95
11	Edit Subject Name	7	9.21
12	Delete Subject Name	7	9.21
		76	100

VI. PROPOSED EFFORT ESTIMATION MODEL SCHEMA AND VALIDATION PROCEDURE

CFP counting and UWE approach are part of the integrated effort estimation model we plan to develop whose schema is presented in Fig. 5. This model, beside mentioned two basic components, has other components, which are:

- XMI parser.
- UWE to COSMIC rules engine.
- Regression model.

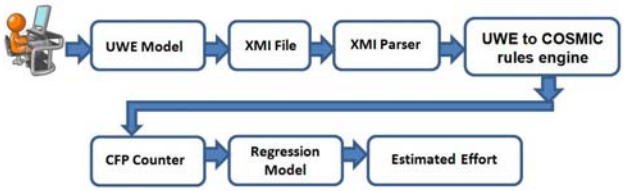


Figure 5. Web development effort estimation models schema

To perform empirical validation of the proposed approach, we will applie Ordinary Least-Squares Regression (OLSR) analysis ([2],[3],[33],[34],[4]). Independent variable CFP will be calculated from the past projects documentation. Linear regression analysis determines equation line which interpolates data and can be used to predict development effort [2]. To assess the acceptability of the derived effort prediction model, we will use some summary measures, namely MRE, MMRE and Pred(0.25) ([2], [3],[34],[4]).

MMRE and Pred(0.25) are considered the *de facto* standard evaluation criteria to assess the accuracy of software prediction models as stated in [35], [36], [37] and they have been used for many years in many comparisons ([2], [34], [4]). The Magnitude of Relative Error [34] is defined as:

$$MRE = |EFHreal — EFHpred| / EFHreal \quad (3)$$

where EFHreal and EFHpred are the actual and the predicted efforts, respectively [2]. We will calculate the Mean of MRE (MMRE) to measure the aggregation of MRE over the observations.

The prediction at level 0.25 [34] is defined as:

$$Pred\ (0.25) = k/N \quad (4)$$

where *k* is the number of observations whose MRE is less than or equal to 0.25, and *N* is the total number of observations of the sets. A good effort prediction model

should have a MMRE≤0.25 and Pred(0.25)≥0.75 (i.e., at least 75% of the predicted values should fall within 25% of their actual values) [2], [34].

VII. Advantages and Drawbacks of the Proposed Approach

Most of the proposed procedures in past works for counting CFP were using older versions of COSMIC measurement manual, mostly versions 2.1 or 2.2 and none of them used new version of manual, version 3.0.1 or COSMIC Business Application Guidelines v1.1 [32] which addresses instructions how to count CFP for on-line business applications. Our opinion is that newer versions of COSMIC manual and especially COSMIC Manual for Business Application Guidelines [32] provide improvements and clarifications that help practitioners to better understand the measurement method and to obtain accurate measures. Therefore, we think that using latest versions (and Method Updates) of manual provide us with instructions for the correct development of measurement procedures which were implemented thru our approach.

UWE conceptual model have enough semantic formalization to specify all the functional requirements, thus, allowing the measurement of the functional size to be more precise.

A. *Advantages of using this approach*

- Usage of UWE tool that exploits notation of UML.
- Easy to use for people familiar with standard design methodology using UML.
- Possibility of using developed application XMI file for purpose of analyze and creating tool for automatic CFP counting on the basis of UWE model. Most of other approaches don't have tool for automated counting of CFP.
- The tool support is guaranteed by the OpenUWE model-driven development environment that comprises, at the current implementation state, two CASE tools: ArgoUWE and MagicUWE.
- Using COSMIC method who is ISO/IEC standard 19761:2011, free and massive support (manuals, examples, case studies etc.).
- COSMIC have improved project control on scope, effort and quality.
- Can be used for web application development effort estimation based on model that is created with usage of OLSR analysis.
- Independent from programming language used for creating applications based on UWE models.

B. *Drawbacks of using this approach*

- Not tested with enough sample applications or practitioners.
- Used small data set of developed applications.
- Small number of Web projects that use this approach (COSMIC for Web Applications).
- Very hard to compare results from our models to some other approaches.
- No repository of finished projects based on this approach.
- With COSMIC method we can't measure: software which is characterized by complex mathematical algorithms, games, streaming software (audio, video).
- Request basic knowledge of UWE process for creating models.
- Request knowledge about UML in compliance with UWE.

Most of the existing methods and procedures are based on some approaches that partially use some standard notations and standards. Most of them use some kind of the rules that are clearly known only to authors of method. Methods and models must be clearly defined in accordance with some internationally accepted standards and modeling languages to provide ease of use for all practitioners with minimal effort for usage and understanding. For new method, it should be very easy to access manuals, case studies, examples, free of charge on whose bases method will be created. In presented approach, this was the main idea and direction for developing this method.

VIII. Future Work

Future work will be focused on further collecting data from finished projects with models, creating database with information's about projects (duration, development team, type of applications, effort, and number of CFP).

Evaluation of model will be conducted thru the usage of data from finished projects, creating effort model with the usage of OLSR analysis and validation of models thru calculation of estimation error (MRE, MMRE, and Pred(n)).

IX. Conclusion

In this paper we have presented possibility of using UWE approach, which is used for modeling web applications, and COSMIC measurement method for estimation of functional size of web applications based on the FUR. FUR has been identified with the usage of use case and activity diagrams in the requirements analyze phase.

COSMIC measurement was conducted with support of UWE stereotypes which can clearly define on which part of system the action is performing: user, system or persistent storage. This is very important for identification of data movements and their direction because they are one of the main elements used for COSMIC measurement method.

Estimation of functional size on the basis of the proposed approach has been demonstrated with simple web application for management of university curriculum register with the simple functionality. As we could saw, two functional processes had the biggest number of CFP,

Add Curriculum Subject and *Edit Curriculum Subject*. Based on these values, project manager could conclude that the biggest effort for development should be invested on the implementation of these two functional processes, e.g. if we have been used this method for development effort estimation purposes. The proposed method requires further experimental research.

At the end, we presented proposed model for development effort estimation of web applications on the basis of our approach with some advantages and drawbacks of this idea.

ACKNOWLEDGEMENT

This work is supported in part by the Croatian Ministry of Science, Education and Sports, under the research project "Computer Support for Education", project number 036-0361983-2019 and WUS project "Information System for Support of Internal Quality Assurance at the Universities of Tuzla and "Džemal Bijedić" of Mostar", project number QA_04/TZ-DzB/09.

REFERENCES

[1] E. Mendes, N. Mosley, and S. Counsell, "Investigating Web size metrics for early Web cost estimation," *Journal of Systems and Software*, vol. 77, no. 2, pp. 157–172, Aug. 2005.

[2] G. Costagliola, S. Di Martino, F. Ferrucci, C. Gravino, G. Tortora, and G. Vitiello, "A COSMIC-FFP approach to predict web application development effort," *J. Web Eng.*, vol. 5, no. 2, pp. 93–120, Jun. 2006.

[3] S. D. Martino and C. Gravino, "Estimating Web Application Development Effort using COSMIC-FFP method," *International Journal of Computers & Applications*, vol. 31, no. 3, pp. 153–158, Jan. 2009.

[4] F. Ferrucci, C. Gravino, and S. D. Martino, "A Case Study Using Web Objects and COSMIC for Effort Estimation of Web Applications," in *Proceedings of the 2008 34th Euromicro Conference Software Engineering and Advanced Applications*, Washington, DC, USA, 2008, pp. 441–448.

[5] L. Buglione, F. Ferrucci, C. Gencel, C. Gravino, and F. Sarro, "Which COSMIC Base Functional Components are Significant in Estimating Web Application Development? - A Case Study," in *Proceedings of the International Conference on Software Process and Product Measurement*, 2010, pp. 205–224.

[6] F. Ferrucci, C. Gravino, and S. D. Martino, "Estimating Web Application Development Effort Using Web-COBRA and COSMIC: An Empirical Study," in *Proceedings of the 2009 35th Euromicro Conference on Software Engineering and Advanced Applications*, Washington, DC, USA, 2009, pp. 306–312.

[7] "The COSMIC Functional Size Measurement Method Version 3.0.1 Measurement Manual," *COSMICON*, 2009. [Online]. Available: www.cosmicon.com/dl_goto.asp?id=73. [Accessed: 13-Dec-2012].

[8] K. Lind and R. Heldal, "A model-based and automated approach to size estimation of embedded software components," in *Proceedings of the 14th international conference on Model driven engineering languages and systems*, Berlin, Heidelberg, 2011, pp. 334–348.

[9] "IFPUG. Function Points Counting Practices Manual (release 4.2)," Westerville, Ohio, 2004.

[10] "MkII Function Point Analysis Counting Practices Manual, v 1.3.1," Sep-1998.

[11] "The application of function point analysis in the early phases of the application life cycle a practical manual: Theory and case study, Version 2.0," 2005.

[12] C. Gencel and O. Demirors, "Functional size measurement revisited," *ACM Transactions on Software Engineering and Methodology*, vol. 17, no. 3, pp. 1–36, Jun. 2008.

[13] D. Schwabe and G. Rossi, "The object-oriented hypermedia design model," *Commun. ACM*, vol. 38, no. 8, pp. 45–46, Aug. 1995.

[14] S. Ceri, P. Fraternali, and A. Bongio, "Web Modeling Language (WebML): a modeling language for designing Web sites," *Computer Networks*, vol. 33, no. 1–6, pp. 137–157, Jun. 2000.

[15] J. Gómez, C. Cachero, and O. Pastor, "Conceptual Modeling of Device-Independent Web Applications," *IEEE MultiMedia*, vol. 8, no. 2, pp. 26–39, Apr. 2001.

[16] R. Paiano and A. Pandurino, "From the Design to the Development: a W2000 Based Framework, Issues and Guidelines," in *Proceedings of Information Resources Management Association (IRMA) International Conference*, Philadelphia, PA, 2003.

[17] G. Rossi, O. Pastor, D. Schwabe, and L. Olsina, Eds., *Web Engineering: Modelling and Implementing Web Applications*, 1st Edition. Springer, 2007.

[18] "ArgoUWE," *UWE–UML-based Web Engineering*. [Online].Available: http://uwe.pst.ifi.lmu.de/toolargoUWE.html. [Accessed: 13-Dec-2012].

[19] "MagicUWE," *UWE – UML-based Web Engineering*. [Online]. Available: http://uwe.pst.ifi.lmu.de/toolMagicUWE.html. [Accessed: 13-Dec-2012].

[20] "MagicDraw," *No Magic*. [Online]. Available: http://www.nomagic.com/products/magicdraw.html. [Accessed: 13-Dec-2012].

[21] B. Marín, G. Giachetti, and O. Pastor, "Measurement of Functional Size in Conceptual Models: A Survey of Measurement Procedures Based on COSMIC," in *Proceedings of the International Conferences on Software Process and Product Measurement*, Berlin, Heidelberg, 2008, pp. 170–183.

[22] E. Mendes, N. Mosley, and S. Counsell, "Comparison of Web size measures for predicting Web design and authoring effort," *IEE Proceedings - Software*, vol. 149, no. 3, pp. 86–92, Jun. 2002.

[23] X. Gu, G. Song, and Q. Li, "An Improved FSM Method for Web-based Applications," in *International Conference on Computational Intelligence for Modelling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce*, 2006, p. 21.

[24] N. Koch and S. Kozuruba, "Requirements Models as First Class Entities in Model-Driven Web Engineering," in *Current Trends in Web Engineering*, M. Grossniklaus and M. Wimmer, Eds. Springer Berlin Heidelberg, 2012, pp. 158–169.

[25] "OMG Unified Modeling Language (OMG UML), Superstructure,Version 2.4.1," Aug-2011. [Online]. Available: http://www.omg.org/spec/UML/2.4.1/Superstructure/PDF/. [Accessed: 13-Dec-2012].

[26] G. Rossi and D. Schwabe, "Modeling and Implementing Web Applications with Oohdm," in *Web Engineering: Modelling and Implementing Web Applications*, G. Rossi, O. Pastor, D. Schwabe, and L. Olsina, Eds. London: Springer London, 2008, pp. 109–155.

[27] S. Meliá, J. Gómez, S. Pérez, and O. Díaz, "A Model-Driven Development for GWT-Based Rich Internet Applications with OOH4RIA," in *Proceedings of the 2008 Eighth International Conference on Web Engineering*, Washington, DC, USA, 2008, pp. 13–23.

[28] S. Ceri, M. Brambilla, and P. Fraternali, "The History of WebML Lessons Learned from 10 Years of Model-Driven Development of Web Applications," in *Conceptual Modeling: Foundations and Applications*, A. T. Borgida, V. K. Chaudhri, P. Giorgini, and E. S. Yu, Eds. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 273–292.

[29] C. Gencel, O. Demirors, and E. Yüceer, "Utilizing Functional Size Measurement Methods for Real Time Software Systems," in *11th IEEE International Software Metrics Symposium (METRICS 2005)*, Como, Italy, 2005.

[30] I. Hussain, L. Kosseim, and O. Ormandjieva, "Approximation of COSMIC functional size to support early effort estimation in Agile," *Data & Knowledge Engineering*, no. In Press, 2012.

[31] T. Rollo, "Sizing e-commerce," in *Proceedings of the Australian Conference on Software*, Sydney, 2000.

[32] "Guideline for Sizing Business Application Software 1.1," *COSMICON*, May-2008. [Online]. Available:

www.cosmicon.com/dl_goto.asp?id=71. [Accessed: 08-Jan-2013].
[33] L. De Marco, F. Ferrucci, C. Gravino, F. Sarro, S. Abrahao, and J. Gomez, "Functional versus design measures for model-driven Web applications: A case study in the context of Web effort estimation," in *2012 3rd International Workshop on Emerging Trends in Software Metrics (WETSoM)*, 2012, pp. 21 –27.
[34] A. Abran, *Software Metrics and Software Metrology*, 1st ed. Wiley-IEEE Computer Society Press, 2010.
[35] E. Stensrud, T. Foss, B. Kitchenham, and I. Myrtveit, "A Further Empirical Investigation of the Relationship Between MRE and Project Size," *Empirical Softw. Engg.*, vol. 8, no. 2, pp. 139–161, Jun. 2003.
[36] L. C. Briand and I. Wieczorek, "Resource Estimation in Software Engineering," in *Encyclopedia of Software Engineering*, John Wiley & Sons, Inc., 2002.
[37] E. Mendes, N. Mosley, and S. Counsell, "Do adaptation rules improve web cost estimation?," in *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, New York, NY, USA, 2003, pp. 173–183.

Značajke i mogućnosti web orijentiranih radnih sučelja

Željko Širanović, Nikola Majstorović i Tomislav Tolić
Tehničko veleučilište Zagreb, Zagreb, Hrvatska
e-mail: zsiranovic@tvz.hr, nmajstor@yahoo.com, ttolic@tvz.hr

Sažetak - Današnje tehnologije za razvitak dinamičkih web aplikacija omogućuju izradu web orijentiranih radnih sučelja. Ona koriste tehnologije računalstva u oblaku, s gotovo istim funkcionalnostima i karakteristikama kao i radne površine klasičnih stolnih računala. Svrha ovakvih web orijentiranih radnih sučelja je objedinjavanje većeg broja usluga i aplikacija iz tzv. oblaka, a koje se mogu koristiti u svakodnevnom poslovanju. U ovom radu analiziraju se značajke i mogućnosti nekoliko takvih web orijentiranih radnih površina.

I. UVOD

Korištenje naprednih tehnologija za izradu dinamičkih web sadržaja omogućilo je razvojnim programerima web aplikacija izradu web rješenja koja vjerno imitiraju radne površine operacijskih sustava klasičnih stolnih računala. Osim što izgledaju gotovo isto, web orijentirane radne površine imaju i solidnu razinu funkcionalnosti kao i radne površine klasičnih stolnih računala. Sve što korisniku treba kako bi pristupio u računalni oblak i pokrenuo svoju osobnu web orijentiranu radnu površinu je Internet veza i neki od standardnih Internet preglednika. Pomoću web orijentiranih radnih sučelja, osobno računalo koje je povezano sa Internetom, postaje dostupno bilo kada i s bilo kojeg mjesta, neovisno o platformi preko koje se korisnik spaja.

Zbog specifičnosti tehnologije računalstva u oblaku, tehnički zahtjevi za rad s web orijentiranim radnim površinama su minimalni. S druge strane, oblak pruža sigurnost od raznih vrsta klasičnih napada virusima, crvima i sl. Ono što je također karakteristično za primjenu web orijentiranih radnih površina je sustavna i konstantna izrada sigurnosnih kopija podataka i manja osjetljivost na kvarove u odnosu na stolna računala, s obzirom da se podaci nalaze u oblaku te se mogu po potrebi u bilo kojem trenutku restaurirati. Broj web orijentiranih radnih površina svakog dana raste, zahvaljujući inovativnosti autora koji ih razvijaju, kao i funkcionalnošću koje su u njih ugrađuju.

II. CLOUD RJEŠENJA I WEB ORIJENTIRANA RADNA SUČELJA

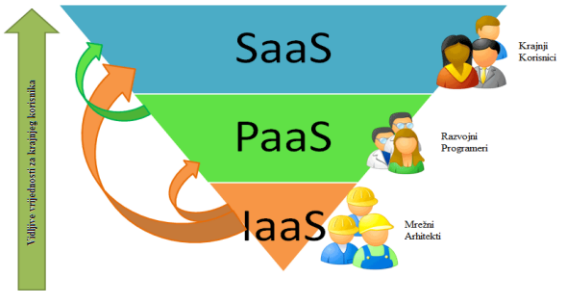
Računalstvo u oblaku (eng. Cloud computing) jedan je od novijih oblika računalstva, za koje postoje više definicija. Nekako najprihvatljivija definicija ona je koja ga opisuje kao koncept podjele programskog okruženja koje koristi Internet kao platformu i omogućuje dohvaćanje aplikacije, usluga i dokumenata iz bilo kojeg dijela svijeta, uz pohranjivanje i čuvanje tih aplikacija, usluga i dokumenata na unaprijed predviđenim poslužiteljima. Područje računalstva u oblaku je u fazi snažnog razvitka, a jedna od sastavnica i značajki njegova razvitka je kombinirano korištenje virtualnih strojeva i aplikacija [1].

Od trenutka pojave tehnologije računalstva u oblaku, na tržištu se pojavilo mnogo pružatelja usluga iz oblaka (Google, Amazon, GoGrid i dr.). Trend svakodnevne pojave novih davatelja ovakvog tipa usluga nastavlja se i dalje, pri čemu svaki od davatelja nudi rješenja koja imaju svoje prednosti i nedostatke, a odluka o izboru najprikladnijeg rješenja je na samim korisnicima, ovisno o njihovim potrebama [1].

Računalstvo u oblaku kao model promiče dostupnost usluga i aplikacija, a sastoji se od tri modela pružanja usluga te četiri modela implementacije.

Ključne karakteristike računalstva u oblaku su: pružanje usluge na zahtjev korisnika, širok mrežni pristup, udruživanje resursa i brza odmjerena usluga [1] [2].

Isporuka usluga računalstva u oblaku podijeljena je u tri arhitekturalna modela, kao i kombinacije izvedene iz njih. Tri osnovne klasifikacije često se nazivaju SPI model (eng. Software, Platform, Infrastructure), odnosno softver tj. program, platforma i infrastruktura. Mogući modeli pružanja usluga su SaaS (eng. Cloud Software as a Service), PaaS (eng. Cloud Platform as a Service) i IaaS (eng. Cloud Infrastructure as a Service) [2] (Slika 1)[1].



Slika 1. Arhitektura Cloud-a

Neovisno o modelima, postoje četiri različita modela provođenja usluga računalstva u oblaku, a to su [2][3]: javni oblak, privatni oblak, zajednički oblak i hibridni oblak. Sve četiri vrste usluga računalstva u oblaku međusobno se prožimaju i nadopunjuju čineći jednu kompaktnu cjelinu.

[1] Preuzeto sa http://astoundcom.com/how-perceptions-of-cloud-computing-have-changed-over-time-part-2-larry-ellison-has-it-right/ datuma 29.1.13

Ovakva međusobna povezanost daje mogućnost korištenja različitih rješenja unutar svih dostupnih oblaka. Zahvaljujući toj povezanosti, korisnik može putem javnog oblaka pristupiti privatnom oblaku svoje tvrtke te putem udaljenog sučelja raditi kao da se nalazi na svom radnom mjestu ili biti dio tima, radeći na projektu kao da se nalazi unutar prostorija svoje tvrtke.

Osnovna ideja web orijentirane radne površine je povezivanje svih usluga računalstva u oblaku: kalendara, kontakt upravitelja, editora teksta, prezentacija, kalkulatora, preglednika slika, klijenta elektroničke pošte i slično, u jednu smislenu i zaokruženu cjelinu koja, bi bila dostupna korisniku bilo kada i od bilo kuda [4] (Slika 2)[2].

Slika 2. Ideja iza pojma web orijentiranih radnih sučelja

Web orijentirane radne površine u stvarnosti samo imitiraju radne površine operacijski sustava klasičnih stolnih računala i zbog toga ne mogu u potpunosti zamijeniti operacijski sustav stolnih računala. On je stoga i dalje potreban kako bi korisnik uopće mogao pokrenuti bilo kakvu internetsku aplikaciju ili uslugu, koristeći internetski pretraživač preko kojeg se povezuje sa Internetom, odnosno web orijentiranom radnom površinom ili web radnim sučeljem.

Koristeći web orijentirane radne površine, korisnikovo personalizirano sučelje zajedno sa svim aplikacijama i dokumentima koje korisnik koristi, postaju dostupni s bilo kojeg računala koje ima pristup Internetu. Usluge računalstva u oblaku daju mogućnost dijeljenja dokumenata i podataka s drugim korisnicima gdje god oni bili, također daje mogućnost zajedničkog timskog rada i drugim korisnicima bez potrebe za fizičkim kontaktom. Ovim širenjem tehnologije na druge platforme korisnik više nije vezan samo za klasično stolno računalo. Programskim rješenjima i uslugama u tzv. oblaku može se pristupiti i s drugih mobilnih uređaja koji imaju pristup Internetu, poput pametnih telefona, tablet računala i sl.

U početku, usluge i aplikacije koje su nuđene kroz računalni oblak bile su, kao i danas, distribuirane na mnogo mjesta i do njih se dolazilo posebnom individualnom registracijom na svaku od njih. Uvođenjem web orijentiranih radnih površina, korisnik se registrira na samo jednom mjestu nakon čega može pristupati svim svojim uslugama. Dobar primjer jednostruke registracije je Google, jedan od vodećih primjera na području razvoja računalstva u oblaku i oblak orijentiranih rješenja poput Gmail-a, Google kalendara, Google Docs. Ono što bi se moglo u budućnosti očekivati od Google-a je implementirana web orijentirana radna površina, koja bi bila puno praktičnija poveznica svih tih pojedinačnih rješenja.

III. ZNAČAJKE I MOGUĆNOSTI WEB ORIJENTIRANIH RADNIH SUČELJA

Kao i svaka tehnologija, web orijentirana radna površina ima svoje prednosti i nedostatke. S obzirom da se radi o rješenju napravljenom za računalstvo u oblaku, osnovni je problem kod svih usluga računalstva u oblaku potreba za korištenjem Internet veze. U slučaju kada korisnik nema dostupan Internet svaki daljnji rad, a time i pristup do web orijentiranog sučelje je onemogućen. Trenutno zbog slabe informiranosti, a dijelom i zbog limitiranih mogućnosti preglednika, postoji mali broj korisnika koji koriste web orijentirana radna sučelja. Zbog toga nema dovoljno prikupljenih korisničkih iskustava, kao ni informacija o prednostima i nedostatcima korištenja web orijentiranih radnih površina. Očekuje se da bi daljnji razvitak i korištenje web orijentiranih radnih površina kao i povećanje brzine i dostupnosti Interneta, pridonijelo kvaliteti i povećanju broja korisnika. Bez obzira na to, navode se neke od glavnih prednosti web orijentiranih radnih površina kao što su [5]:

- pogodnost
- mobilnost
- sesijsko upravljanje
- softver za upravljanje
- sigurnost
- visoka dostupnost.

Kao nedostaci web orijentiranih radnih površina obično se navode [5]:

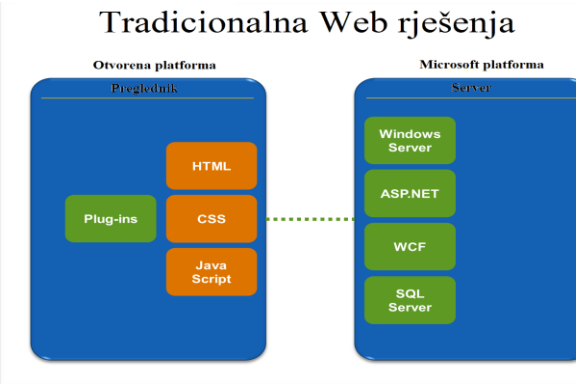
- sigurnost
- brzina
- značajke aplikacija
- mrežni pristup
- kontrola pristupa
- centralna kontrola.

S obzirom da je web orijentirano radno sučelje derivat računalstva u oblaku, cijela infrastruktura web orijentiranog radnog sučelja leži na infrastrukturi računalstva u oblaku. Ono što je važno za web orijentiranu aplikaciju je to da krajnji korisnik ima pristup Internetskoj mreži, bez obzira da li je to zemaljska linija, ADSL, optika ili mobilna mreža HSDPA i sl. Iskustva korisnika najviše proizlaze iz svojstva i mogućnosti internetskog preglednika preko kojeg korisnik pristupa svojoj web orijentiranoj radnoj površini. Većina web orijentiranih radnih površina napravljena je na način da

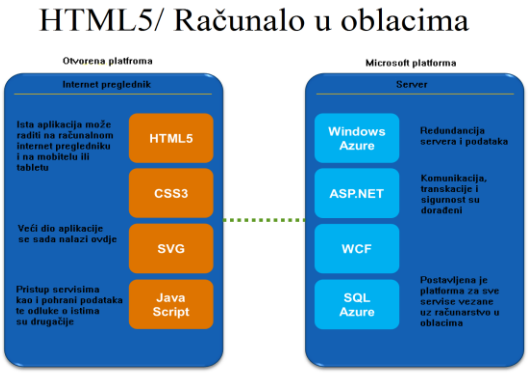
[2] Preuzeto sa http://www.alchemysys.net/solutions/hosted-cloud-computing/windows-hosted-desktop/how-it-works/ datuma 29.1.2013.

funkcionira kroz danas najzastupljenije preglednike kao što su Internet Explorer, Firefox, Safari, Opera, Crome.

Korištenje HTML-a ne daje dovoljno mogućnosti kod izrade web orijentiranih radnih sučelja, naročito u postizanju dinamike i omogućavanje potrebne interaktivnosti. U svrhu razvitka dinamičkih i interaktivnih rješenja koja bi unaprijedila komunikaciju između servera i klijenta, koriste se druge tehnologije i alati za izradu web orijentiranih radnih površina. Jedna od pogodnih platformi za razvoj takvih aplikacija je AJAX. Uz korištenje AJAX-a, vrlo se često pri izradi web orijentiranih radnih površina, koristi i PHP. Tehnologije koje se još susreću u radu s web orijentiranim radnim površinama su Flash, Java, Javascript i HTML5. Sve prethodno navedene tehnologije omogućuju izradu interaktivne komunikacije sa serverom, kao i postizanje gotovo svih funkcionalnosti kakve nalazimo i kod klasičnih desktop operacijski sustava i aplikacija. Osobitost HTML5 nalazi se u mogućnosti izrade dinamičkih web rješenja čime se nadopunjuju svi nedostaci koji su otežavali izradu takvih aplikacija kod njegovih prethodnika. Na slikama je dan grafički pregled načina na koji HTML5 funkcionira unutar tradicionalnog web sustava (Slika 3)[3] ili rješenja napravljenih za računalstvo u oblaku (Slika 4)[4].



Slika 3. Tradicionalno pristupanje i komunikacija sa serverom



Slika 4. Komunikacija klijent-server koristeći HTML5 i cloud rješenje

Kao što je vidljivo iz grafičkog prikaza (Slika 4), broj tehnologija koje su ugrađene unutar HTML5 je popriličan, za razliku od standardnog HTML-a. Dio koda ili cijelu aplikaciju sada preuzima Internet preglednik, kako bi ubrzao rad, dok server prepoznaje da li se radi o desktop računalu, pametnom telefonu ili nekom drugom uređaju. S druge strane, server je više orijentiran prema radu u oblaku, povećana je redundancija kao i pristup podacima uz dodatnu sigurnost pristupa.

Razvitak web orijentiranih radnih sučelja u većini slučajeva se oslanja na „open source“ zajednicu, a to je ujedno i razlog zbog kojeg je većina aplikacija u konstantnoj razvojnoj fazi. Velik dio rješenja i funkcionalnosti je preuzet iz već postojećih aplikacija, dok se razlike jedino očituju u grafičkom djelu. Neki od proizvođača web orijentiranih radnih sučelja nude mogućnost kupnje ili ograničeno trajanje. Drugi daju kompletnu besplatnu aplikaciju koju korisnik može postaviti na vlastiti server te na taj način privatno pokrenuti vlastito interno oblak rješenje i tako omogućiti svojim internim korisnicima privatne web orijentirane radne površine.

IV. PRIMJERI WEB ORIJENTIRANIH RADNIH SUČELJA

U ovom radu je prezentirano nekoliko najraširenijih web orijentiranih radnih sučelja. Njihove mogućnosti se međusobno razlikuju. Neka od tih sučelja nude samo osnovne funkcionalnosti, a neka nude puno više funkcija koje se mogu naći i kod stolnih računala. Jedan dio njih je potpuno besplatan, a dio njih ima ograničenu funkcionalnost u okviru probne verzije.

1. Schmedley

Schmedley [6] je zamišljen kao početna stranica sastavljena od mnoštva dodataka (eng. widgets). Kao što je moguće vidjeti na slici (Slika 5), na dnu ekrana nalazi se alatna traka s ikonama, gdje svaka ikona predstavlja jednu widget funkcionalnost, koja se može postaviti na radnu površinu. Radna površina je zamišljena baš kao i tab-ovi u Internet preglednicima, tako da krajnji korisnik može imati aktivirano više radnih površina tj. tab-ova. Preko widget-a korisnik se može povezati na svoj email, društvene mreže poput Facebook-a, Twitter-a, Myspace-a i sl. Od drugih aplikacija, korisnik ima na raspolaganju kalendar, sat i listu zadataka (eng. To-do list). Na radnoj površini također postoji mogućnost pregledavanja Youtube filmova, Google maps i sl. Pretraživanje Interneta je omogućeno preko ugrađenih tražilice koje se mogu povezati na Google i Yahoo, a postoje i drugi standardni widgeti kao što su prognoza vremena, rss feeder, sat. Za zabavu postoji set od nekoliko jednostavnih flash igrica.

Iako Schmedley predstavlja web orijentirano radno sučelje, za poslovne korisnike nije toliko interesantan jer nema mogućnost izrade i uređivanja tekstova, slika, prezentacija i tabličnih kalkulacija. Također, ne postoji mogućnost postavljanja prava na kreirane dokumente, kao ni njihovog dijeljena drugim korisnicima. Cijeli projekt je

[3] Preuzeto sa http://davidpallmann.blogspot.com/2011/08/when-worlds-collide-html5-meets-cloud.html (29.1.2013.)
[4] Preuzimanje isto kao i slika 3.

još uvijek u fazi razvoja, stoga se može očekivati daljnji napredak i povećanje funkcionalnosti ovog rješenja.

Slika 5. Schmedley sučelje

2. *Glide OS 4.0*

Glide OS [7] predstavlja jedno potpuno funkcionalno web orijentirano sučelje (Slika 6), kao besplatno rješenje koje u svojoj osnovnoj inačici uključuje i 30 GB besplatnog diskovnog prostora. Uz to Glide OS sadrži popularne uredske alate kao što su: uređivač teksta, program za izradu prezentacija i program za uređivanje slika. Kod ovog rješenja standardno dolazi kalendar, klijent elektroničke pošte, kalkulator te program za jednostavno crtanje nalik na Windows Paint.

Glide OS daje mogućnost dijeljenja dokumenata s drugim korisnicima ili grupama korisnika te njihovo administriranje. Jedna od zanimljivih mogućnosti Glide-a je i ta što omogućuje korisniku kreiranje vlastitih web stranica. Također Glide OS ima mogućnost sinkronizacije korisnikovih podataka na lokalnom računalu s podacima spremljenim u oblaku, uz mogućnost njihove automatske sinkronizacije (podatci, kontakt lista i slično).

Slika 6. Glide OS 4 sučelje

3. *ZeroPC*

ZeroPC [8] je zamišljen kao multimedijalno središte (Slika 7). Izgled sučelja zajedno s izgledom i rasporedom elemenata je uvelike slično Microsoft Windows sustavu. Zbog toga je rad u ZeroPC jednostavan, jer se krajnjem korisniku uglavnom sve čini poznatim i obično se sve nalazi tamo gdje i očekuje, kao i kod Windows-a. ZeroPC nudi integraciju s mnogim uslugama koji se nalaze na Internetu kao što su: Dropbox, Box.net, Evernote, Facebook, Flickr, Google Docs, Instagram, SkyDrive, Picasa. U besplatnoj verziji, korisnik ima na raspolaganju 1GB diskovnog prostora, a s prethodno navedenim uslugama, može ga koristiti za spremanje i pregledavanje teksta, slika, glazbe, videa i sl. Navedene materijale također može dijeliti i s drugim korisnicima, administrirajući prava i kontrole nad tim dokumentima. Osim multimedijalnih playera, druge aplikacije nisu instalirane, već se ZeroPC oslanja na rješenja drugih, kao što je npr. Google Docs za dokumente. Pritiskom na gumb APPS (donji lijevi kut)[5] otvara se izbornik koji omogućuje aktiviranje dodatnih programa, kao npr. jednostavan tekst editor, messenger i web browser i sl. Ova web orijentirana radna površina nije usmjerena ka poslovnim korisnicima i to joj je nedostatak. Taj nedostatak premošćuje time što poslovne aplikacije nasljeđuje od drugih proizvođača i ne oslanja se na vlastito integrirano rješenje. Upravljanje datotekama i mogućnosti njihova dijeljenja drugim korisnicima solidno je izvedena.

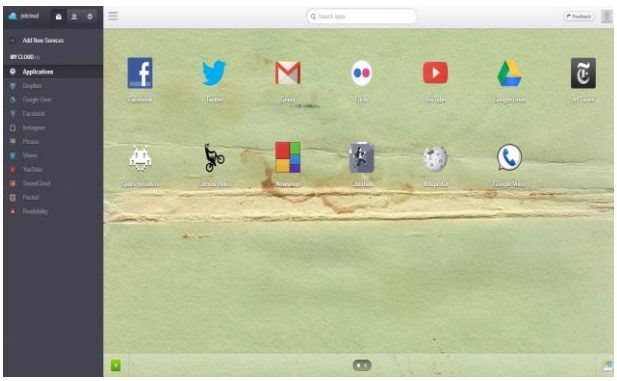
Slika 7. ZeroPC sučelje

4. *Jolicloud*

Jolicloud [9] je po funkcionalnosti sličan Schmedley-u (Slika 8). Sučelje je sastavljeno od nešto većih ikona nego što je to uobičajeno, što daje naslutiti da je orijentiran prema uređajima s ekranom na dodir (eng. touchscreen). Kao i kod većine web orijentiranih radnih površina, ovo rješenje ima mogućnost personalizacije izgleda pozadine i rasporeda elemenata. Za upravljanje datotekama Jolicloud koristi Dropbox, za elektroničku poštu koristi Gmail, a od uredskih alata Microsoft Office live.

[5] Slično kao gumb Start na Microsoft Windows OS

Iako Jolicloud predstavlja svojevrsnu web orijentiranu radnu površinu prije ga se može okarakterizirati kao zbirno mjesto za različite aplikacija iz oblaka.



Slika 8. Jolicloud sučelje

5. *CloudTop*

CloudTop [10] svoje grafičko sučelje temelji na izgledu Microsoft Windows-a 7 OS-a (Slika 9). Kao i kod Joliclouda, ekranom dominiraju veće ikone, dok se pritiskom na ikonu oblačića dobivaju funkcije slično kao kod pritiska na start ikonu kod Windows OS-a. CloudTop pruža punu funkcionalnost koja postoji i kod desktop operativnog sustava. Sučelje je pregledno, grafički dobro organizirano, a zbog dizajna sličnog Windows-ima sve djeluje jako poznato i intuitivno, dok je upravljanje i dijeljenje datoteka dobro riješeno. Od ostalih aplikacija na raspolaganju je sve kao i kod konkurentskih programa, od kalendara, kalkulatora, kontakt upravitelja, klijenta elektroničke pošte, preglednika slika, multimedijskog play-era, preglednika i uređivača slika. Jedna od zanimljivijih aplikacija je marketplace, kod koje postoji mogućnost dodavanja (instaliranja) drugih aplikacija kreiranih za CloudTop.

Još jedna posebnost ovog web orijentiranog radnog sučelja je razina personalizacije koju korisnik može ostvariti. Desnim klikom miša unutar prozora omogućen je izbornik kao kod Windows-a gdje je moguće promijeniti veličinu ikona, raspored elemenata i njihovo sortiranje. Korištenjem besplatne verzije na raspolaganju je i 3GB diskovnog prostora kojeg je, uz nadoplatu, moguće i povećati. Ovo sučelje djeluje kao pravi web operativni sustav (iako to nije) koji ima i tzv. DOS konzolu koja omogućuje upravljanje s diskom i podacima. Jedan od dodataka je i Sidebar preko kojeg se mogu instalirati i dodavati različiti widgeti.

CloudTop u potpunosti zadovoljava većinu zahtjeva i potreba velikog broja korisnika standardnih desktop računala, gdje mogućnosti, jednostavnost i brzina rada ne zaostaju za stolnim rješenjima.

Slika 9. CloudTop sučelje

6. *EyeOS*

EyeOS[6] također predstavlja cjelokupno i zaokruženo rješenje, kako za privatne tako i za poslovne korisnike (Slika 10) [11]. EyeOS raspolaže s naprednim aplikacijama za kreiranje grupa korisnika, kao i dijeljenje datoteka korisnicima grupe. Napredni upravitelj datoteka osim upload-a, nudi dodatnu aplikaciju koja se integrira s računalom korisnika, pružajući lakšu sinkronizaciju podataka. Osim klasičnih podatka mogu se prenijeti i sinkronizirati kontakti i obveze iz kalendara. EyeOS raspolaže sa standardnim aplikacijama kao što su kalkulator, upravitelj kontaktima, interaktivni razgovor (eng. chat), klijent elektroničke pošte, jednostavni notepad, preglednik slika i uređivač teksta.

Posebnost EyeOS-a je i njegova velika „open source“ zajednica. Kroz EyeOS open source zajednicu razvijen je veliki broj dodatnih aplikacija i multimedijalnih mogućnosti. Zbog toga EyeOS raspolaže multimedijalnim programima kao što su preglednik video datoteka, svirač glazbenik datoteka, program za uređivanje slika. Također, zbog velikog broja ljudi koji su uključeni u razvitak ovog rješenja, EyeOS radan površina je lokalizirana na više od 20 jezika.

Razvitak i postavljanje ovog rješenja na vlastiti server tvrtke prilično je dobro dokumentiran, a njegove prednosti prepoznale su i velike računalne tvrtke poput IBM, Orange, Vodafone.

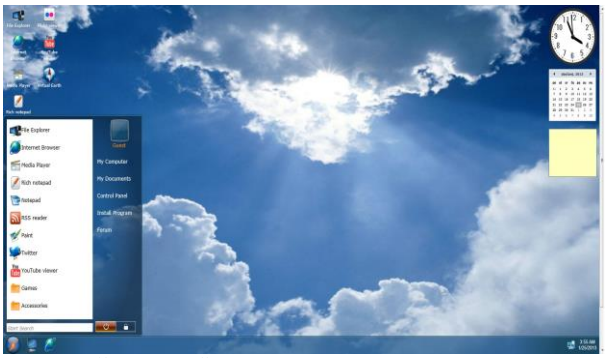
Slika 10. EyeOS sučelje

[6] Projekt je u međuvremenu iz otvorenog koda prešao u komercijalni proizvod bez mogućnosti probnog rada.

7. *SilveOS*

SilveOS [12] jedan je od novijih proizvoda (Slika 11). Već nakon prve prijave u sustav, korisnik nailazi na poznato sučelje, vrlo slično sučelju Microsoft Windows 7. Tehnologija kojom je SilveOS izrađen je Microsoft Silverlight. Od aplikacija, sustav posjeduje jednostavan uređivač teksta, jednostavan program za crtanje te ugrađeni preglednik slika. Tu su još neke jednostavne računalne igrice koje standardno dolaze uz Windows platforme.
Unutar sučelja je integriran Microsoft Bing preko kojeg se može pristupiti mapama, kao i pregledavati cijelu kartu svijeta. SilveOS posjeduje i jednostavan Internet preglednik, kao i multimedijalni svirač.

SilveOS za sada nema neku veću praktičnu primjenu, osim kao demo primjer mogućnosti Silverlight tehnologije i prikaz njezine primjene u svrhu web orijentiranog sučelja. Ovo rješenje ne donosi nikakvu aplikaciju za upravljanje datotekama, kao ni aplikacije za međusobnu suradnju korisnika.



Slika 11. SilveOS sučelje

TABLICA 1. Pregled navedenih web orijentiranih radnih površina

Naziv	**Mogućnost javnog testiranja**	**Besplatni diskovni prostor**	**Mogućnost preuzimanja serverske aplikacije**	**Mogućnost kreiranja grupa i dijeljene dokumenata**	**Uredske aplikacije (word, excel, powerpoint)**	**Multimedijske mogućnosti**
Schmedley	Da	0 MB	Ne	Ne	Ne	Ne, widget od Youtube-a
Glide OS 4.0	Da	30 GB	Ne	Da	Uređivač teksta i program za izradu prezentacija	Preglednik slika
ZeroPC	Da	1 GB	Ne	Da	Nema vlastito rješenje koristi ThinkFree Office[7]	Audio svirač, video player, preglednik slika
Jolicloud	Da	0 MB	Ne	Nema vlastito rješenje nego se koristi DropBOX[8]	Nema vlastito rješenje, nudi linkove na aplikacije drugih proizvođača	Nema vlastito rješenje, nudi linkove na aplikacije drugih proizvođača
CloudTop	Da	3 GB	Ne	Da	Nema vlastito rješenje, nudi Zoho proizvode	Preglednik slika, audio player
EyeOS	Ne (postojala javna beta verzija koja je ugašena)	1 GB	Da	Da	Uređivač teksta i program za izradu prezentacija, mogućnost povezivanja s Open Office[9]	Audio/video player, preglednik slika
SilveOS	Da	0 MB	Ne	Ne	Samo jednostavan uređivač teksta	Audio/video svirač, preglednik slika, Bing mape

[7] http://www.thinkfree.com/main.jsp java orijentirano, cloud rješenje za uređivanje teksta, tabličnih proračuna i izradu prezentacija
[8] https://www.dropbox.com/about usluga za upload i download, te dijeljenje datoteka.
[9] http://www.openoffice.org/hr/ projekt otvorenog koda sa motivom zajedničkog stvaranja skupa uredskih alata koji će raditi na svim značajnijim platformama

V. ZAKLJUČAK

Današnje tehnologije za razvitak dinamičkih web aplikacija i računalstvo u oblaku omogućile su izradu web rješenja koja imitiraju sučelja klasičnih stolnih računala, koja nazivamo web orijentirane radne površine ili sučelja. Web orijentirana radna sučelja imaju gotovo iste funkcionalnosti i karakteristike kao i radne površine klasičnih stolnih računala. Web orijentirane radne površine imaju za cilj objedinjavanje što većeg broja usluga i aplikacija iz tzv. oblaka, a koji se mogu koristiti kako za privatne tako i za poslovne svrhe. Većina web orijentiranih radnih površina je razvijena na način da se pokreću kroz danas najzastupljenije preglednike kao što su Internet Explorer, Firefox, Safari, Opera i Crome. Uvođenjem web orijentiranih radnih površina korisnik se registrira samo na jednom mjestu, nakon čega može pristupati svim svojim uslugama koje su u tzv. računalnom oblaku.

U ovom radu je prezentirano nekoliko najvažnijih web orijentiranih radnih sučelja, uz prikaz njihovih osnovnih značajki i mogućnosti. Neka od rješenja nude samo osnovne funkcionalnosti, a neka nude znatno više funkcija koje se mogu naći i kod stolnih računala. Jedan dio njih je potpuno besplatan, a dio njih je u okviru probne verzije s ograničenim funkcionalnostima. Svaki korisnik, sukladno vlastitim potrebama i zahtjevima, može za sebe pronaći ono rješenje koje mu najbolje odgovara.

Web orijentirana radna sučelja još su u razvitku i dijelom ograničena svojim mogućnostima u odnosu na klasična stolna računala. Tehnologije koje danas stoje na raspolaganju za njihov razvitak oslanjaju se na HTML5, AJAX, PHP, Flash, Java, Javascript, kao i veliku „open source“ zajednicu iz koje dolazi najveći broja programera koji ih razvijaju, zbog čega je i većina rješenja u konstantnoj razvojnoj fazi. Tehnologije koje stoje na raspolaganju obećavaju, a uz napredak i povećanja mogućnosti Internet preglednika, veće brzine mrežnog prijenosa i daljnji rast broja korisnika, web orijentirana radna sučelja mogla bi imati u skoroj budućnosti dobru perspektivu.

VI. LITERATURA

[1] http://sigurnost.lss.hr/Dokumenti/cloud-computing.html (3.1.2013.)

[2] http://www.cert.hr/sites/default/files/NCERT-PUBDOC-2010-03-293.pdf (3.1.2013.)

[3] http://www.itcitaliste.rs/index.php (3.1.2013.)

[4] http://en.wikipedia.org/wiki/Cloud_computing (3.1.2013.)

[5] http://en.wikipedia.org/wiki/Web_desktop (3.1.2013.)

[6] http://schmedley.com (24.1.2013.)

[7] https://desktop.glidesociety.com (24.1.2013.)

[8] http://www.zeropc.com (24.1.2013.)

[9] http://my.jolicloud.com (24.1.2013.)

[10] http://cloudtop.com (24.1.2013.)

[11] http://www.eyeos.com (24 1.2013.)

[12] http://www.silveos.com (29.1.2013.)

Abstraction and Semantics support in M2M communications

Vanesa Čačković, Željko Popović
Ericsson Nikola Tesla d.d.
Krapinska 45, Zagreb, Croatia
e-mail: vanesa.cackovic@ericsson.com, zeljko.popovic@ericsson.com

Abstract - In line with the global horizontalization trends in M2M solutions, it is predicted that considerable percentage of the M2M service platforms will be designed to accommodate multiple verticals. Data and information management tools will be of increasing importance. This includes capabilities to access and capture data, but more prominently how to handle data and information in a structured way. Semantic annotation of data and means to add contextual information will be needed. Data models will evolve both as domain specific models and as common models. Instrumental will be tools to link these models, and also technologies to perform reasoning and knowledge e.g. based on evolved semantic web technologies. This paper will give an overview of current work in standardization bodies on semantic support for M2M data: how to discover, interpret and use the M2M data from different sources, without any kind of prior knowledge of that. This is essential to offer high-level M2M horizontal services and to develop open markets for M2M data.

I. INTRODUCTION TO M2M

Over the last few years there has been much talk about Machine-to-Machine (M2M) communications. Machine-to-Machine (M2M) communications by the definition are technological solutions and deployments allowing Machines and Devices to communicate with each other, with no (or little) human interactions [1]. Although the concept is not new, M2M is now at a rapid expansion. The reason for that is mainly in the technological advancement resulting with availability of applicable communication technologies, affordable devices and embedded computing with sensing capabilities. There is also an emerging need for M2M applications from across industries and society, driven mainly from enterprise or government needs. Enterprises focus on cost reduction and efficiency while government focus is on sustainability, safety and socio-economic impacts. The telecom industry is increasingly looking at M2M as a new source for revenue when on many markets revenue growth is stalling [2].

Although the M2M market is much alive and growing, there are still a lot of open questions like fragmentation of solutions, network misalignment, security, privacy, service capabilities, testing and certification of devices, etc. that need to be overcome before the M2M market can reach its full potential. Therefore, the next logical step is the standardization of the M2M domain. [3]. Without the proper standards there is no interoperability both within and between domains. Within a domain, standards provide cost efficient realizations of solutions. Between domains, the interoperability ensures cooperation between the engaged domains [3].

Current M2M markets are still highly fragmented. Various vertical M2M solutions have been designed independently and separately for different applications, which inevitably impacts or even impedes large-scale M2M deployment. Standardization bodies have the long term intent to drive the M2M domain towards a horizontal system and business orientation. The reason for belief in horizontalization is that it provides the most efficient structure for the involved industries [7].

The European Telecommunications Standards Institute (ETSI) formed an M2M Technical Committee responsible for developing standards for Machine to Machine Communications. The group aim is to provide an end-to-end view of Machine to Machine standardization, and to co-operate closely with ETSI's activities on Next Generation Networks, and also with the work of the The 3rd Generation Partnership Project (3GPP) standards initiative for mobile communication technologies. The Release 1 of the M2M functional architecture has been published in October 2011 [7]. Release 1 of ETSI M2M defines a Service Capability Layer (SCL) which is enabling transport of M2M data between devices or gateways and network applications. Release 1 provides an abstraction layer hiding the heterogeneity of M2M access networks and provides means for secure data transport. By design choice, the ETSI M2M Rel.1 SCL is handling only data containers without any knowledge of the data contained. The advantages of this approach are a clean separation of data transport from data handling and focus on the generic, commonly needed functions of an SCL - thus avoiding applications-specific functionality to be included into the ETSI M2M standards.

While Rel. 1 of ETSI M2M already opens a lot of opportunities in the M2M area, there are a number of limitations:

- The common-place vertically integrated, but isolated M2M applications are now replaced by M2M applications which are re-using a common data transport, but which are still vertically integrated and isolated from each other

- Device and application need to agree beforehand on a common definition of the exchanged containers as well as on the contained data. This makes re-use of M2M data across different applications difficult.

• There are only very limited functions in Rel.1 to discover which data are available in an SCL

• There is no support in the SCL to enable an open market of data, e.g. in which data owner publish (sell) their data and independent data users provide applications that make use of the data

• Limited chances for ETSI M2M compliant platform providers to enable value-added services re-using M2M data

• Limited opportunities for treating different kinds of M2M data with different Quality-of-Service or by charging differently for them

For operators and providers of an ETSI M2M compliant platform this is limiting their ability for offering new and innovative business models [4].

This paper will provide an overview of current work on semantic support for M2M data: how to discover, interpret and use the M2M data from different sources, without any kind of prior knowledge of that. Section two will provide the description of key features of M2M communication and existing technology solutions, section three will provide the overview of standardization activities on semantics in M2M, section four will provide an explanation what constitutes semantic information and how can it be used in M2M communication whilst the last fifth chapter will give a conclusion.

II. M2M TECHNOLOGY SOLUTIONS

The main capability of any M2M system is the remote monitoring and control of real world properties of different environments and places (air quality, mines, cities) or of a diverse set of assets (buildings, vehicles, goods, other infrastructures). Although the general purpose is the same, the service requirements of M2M services can vary significantly [6]. Connected consumer electronics will have similar (demanding) QoS requirements as mobile multimedia services for mobile broadband. On the contrary, sensors, meters and connected devices have low QoS requirements, but may have very stringent requirements on energy efficient operation and cost. Other M2M services may impose other requirements, like high reliability and low delay for industrial control applications and smart grid; high reliability, low delay and high mobility for automotive applications, etc. Since the M2M services can be so diverse, today's most M2M solutions are highly specialized stove-pipes with hardcoded integration of devices into the applications.

A. *Sensor and Actuator service functionality*

It is predicted that in following years a considerable percentage of the M2M service platforms will be designed to accommodate multiple verticals, a trend that we already see today. In order to accommodate multiple verticals there have to be certain horizontalization points [2].

One horizontalization point is at the level of M2M device access (device access reuse by multiple applications). The M2M Service layer is expected to become a resource management layer: The web resource abstraction will be extended to sensors, actuators, processing software and sensor data storage (current examples include Pachube and SENSEI, [9], [10]).

The next horizontalization step is to enable sensor data reuse by multiple applications and the ability of multiple applications to send commands to certain actuators. This requires a common data model or linked data models for heterogeneous sensor data or actuator commands. So by 2016 it is predicted that there will be data transformation functions in the sensor and actuator service layer since it is expected that there will still be competing sensor technologies for the same application area (e.g. ZigBee, Bluetooth for Health) or different sensor technologies with different sensor models for different application areas that the service layer has to accommodate. However by 2016 there will only be a few cases of verticals that share information from the same sensors (e.g. electric vehicle car sensors for utility and automotive) and therefore their data models need to be at least linked [4].

The sensors collect information and the actuators provide control capabilities for certain real world entities of interest (a room is a real world entity of interest and the sensors and actuators are the resources). By 2016 the information model that describes the entities of interest (people, places, things) will be decoupled from the information model of the information sources (sensors) and actuation targets (actuators) related to an entity of interest. Due to mobility of resources and entities in certain application scenarios, the relationship between the entities and the relevant resources (sensor S is in room R) as well as entities themselves (e.g. person P is in room R) will be dynamic and their relationship explicitly declared using semantic technologies such as Resource Description Framework (RDF). Nevertheless application logic that uses the resource and entity of interest information models will be hardcoded, i.e. there will be no artificial intelligence to process semantically annotated relationships. We might also see some standardized relationships for resources and entities the same way as there are standardized representations for relationship between people (e.g. friend-of-a-friend ontology). The Semantic Sensor Network (SSN) incubator group of The World Wide Web Consortium (W3C) already has started this effort.

B. *Data and information management*

Data and information can come from any conceivable source. They include sensors, measurements and observations made by various ICT systems, explicit and implicit user inputs from e.g. social network sites, or other static or dynamic databases. All in all, data and information processing is about efficient tools for handling, understanding and making use of data and information as increased knowledge of for making decisions or executing actions. It should be noted that tools are needed both for handling individual unique data items (“small data”) as well as large amounts of data (“big data”).

The creation of a data source and development of any application making use of the data already is or will be decoupled. This is in particular true for sensor networks

where deployment traditionally has been done with a single application in mind. This decoupling of usage from data source requires a semantic description of the data.

Raw sensor data (e.g. temperature) will not be specific to a vertical application. A sensor data model should provide a stepwise mapping to higher abstractions. The first step is to make a solely sensor related semantic tagging of the data. Secondly, domain specific information should be added explaining the context of the data, e.g. health. Subsequent steps could include specifications of application specific information. Work in this direction has been conducted by e.g. Open Geospatial Consortium in Sensor Web Enablement as well as the Semantic Sensor Network incubator group in W3C [11], [12].

In addition to domain specific information models, there will also be general ontologies and common vocabularies which can link together information from different domains. The Linked Open Data project [13] provides information and tools for linking data from different domains. The basic technologies used Language to describe the ontologies are RDF and Web Ontology (OWL). In the development of domain specific ontologies common vocabularies will be increasingly accepted and used.

The current semantic web technologies build on the notion of adding new information will not change previous conclusions. However, for dynamic knowledge creation, extensions to the current Semantic Web Technologies are needed.

III. STANDARDIZATION NEEDS

As already mentioned in the introduction, the current M2M related standards and technologies landscape is highly fragmented. The fragmentation can be seen across different applied domains where there is very little or no reuse of technologies beyond basic communications or networking standards. Even within a particular applied sector, a number of competing standards and technologies are used and promoted.

The standards landscape related to M2M is very large. The Global Standards Collaboration Machine-Machine Task Force (GSC MSTF) identifies 143 organizations with a direct or indirect interest in M2M standardization [4]. The entire ecosystem of solution providers and users would greatly benefit from less fragmentation and should strive towards use of a common set of basic tools. This would provide faster time to market, economy of scale and reduce overall costs.

To avoid creation of competing M2M standards Global One M2M Initiative has been formed [15]. Its goal is to develop one globally agreed M2M specification with initial focus on Service Layer and consolidation of current M2M Service Layer standards activities into the One M2M initiative. The work in One M2M will be based on ETSI M2M vision that can be simplified shown as in Figure 1. There are essentially three elements of the M2M value chain: the device, the network and the application. The goal is to connect multiple devices through technology independant network to common horizontal application infrastructure which is built reusing existing standards and can serve multiple applications.

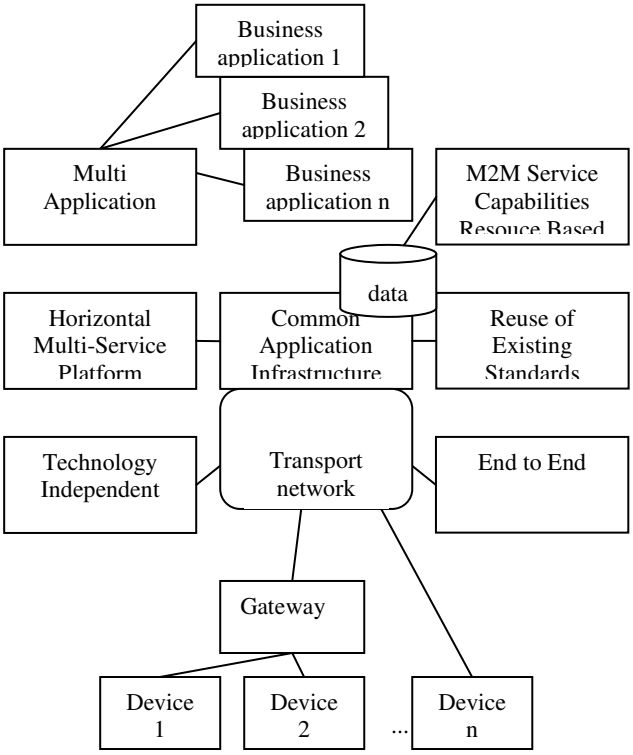


Figure 1. ETSI M2M vision

ETSI Release 2 M2M standard should also propose some default way how syntactic and semantic information could be documented by vertical industries. Work on support of semantics in M2M is still in state of early draft: „ETSI TR 101 584 V0.4.0 Machine to Machine Communications (M2M); Study on Semantic support for M2M Data“[5]; however it is announced that the standard will be available in next few months time frame.

This Study on Semantic support for M2M Data will analyze benefit, feasibility and potential requirements for the support of semantic information on application related M2M Resources in the M2M system. It will create use cases that illustrate provisioning and usage of such semantic information and that demonstrate the benefit for the M2M ecosystem.

It will also investigate on the kind and amount of semantic information that would become available in the M2M system and investigate discovery mechanisms for semantic information in the ETSI M2M System [5].

IV. SEMANTIC IN M2M COMMUNICATION

At the core of providing M2M applications is the collection, storage and processing of data and information from different M2M devices for subsequent use by the M2M applications themselves. Those functionalities are implemented in M2M Service Enablement (SE) layer which can simplified be shown as in Figure 2. The main assumption behind the architecture in Figure 2 is that it follows a service oriented design. That translates to the fact that different services can be used by the service users as stand alone services. The minimum set of functions

needed for a meaningful M2M SE Platform contains the Data Access and Messaging, B2B Service Access and the Security and Policies.

Information Services are part of SE (as opposed to raw data services) and include services that offer high level information about devices, entities of interest or processed raw data. Information Services contain domain specific data services that facilitate the processing of raw data into information (e.g. context information) meaningful to the service users.

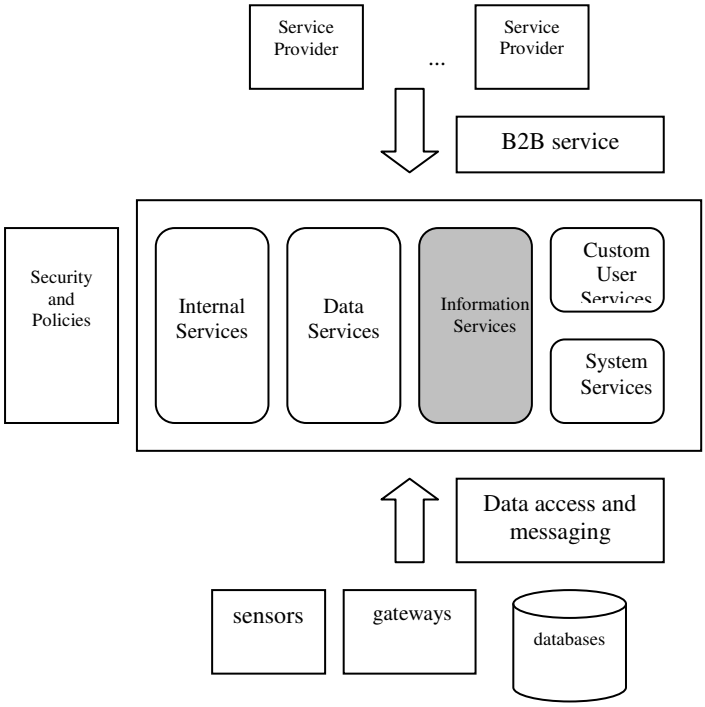


Figure 2. M2M Sevice Enablement architecture

The measured or observed information about physical phenomena and the context information about entities of the real world needs to be modelled in order to enable the interpretation of this information in the M2M SE as well as by consumer applications and services.

Semantic annotation of the data (for example, with domain knowledge) can provide machine-interpretable descriptions on what the data represents, where it originates from, how it can be related to its surroundings, who is providing it, and what are the quality, technical, and non-technical attributes.

Data collected from sensors normally have different formats and do not include semantic information in the data. To be able to efficiently use the data by various applications, the data need to be mapped to a standard format and tagged with semantic information.

Data mapping function could look like shown on figure 3:

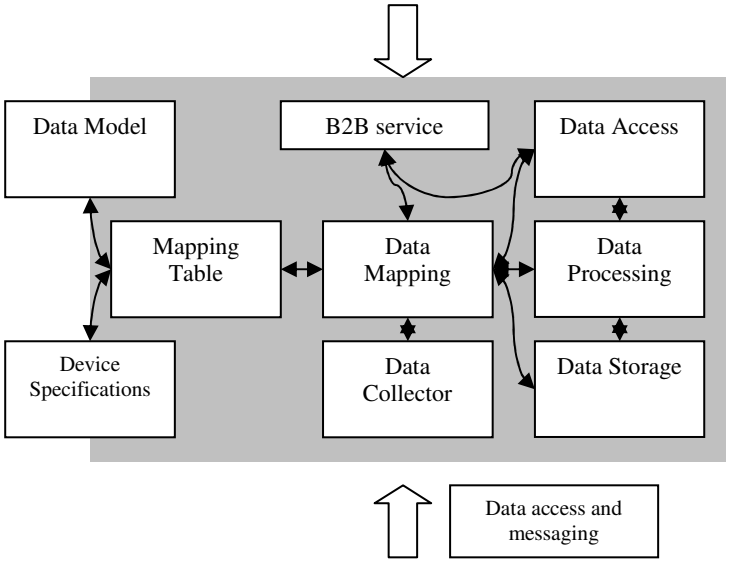


Figure 3. Data mapping function

To support the data mapping, a mapping table needs to be created to link the unified data model with specific sensor and device descriptions. To do the mapping, the semantics of the collected data need to be understood. The original sensor data or device descriptions can be obtained either manually or automatically fetched from manufacturers. It is then compared with unified data model to create a mapping table for each sensor and device. This step is done at the device setup phase.

When the data are collected at runtime, the data mapping function identifies the data source and find the relevant mapping information from the mapping table to map the data into our unified data model.

From the application side, it only needs to know the defined data model with specified interface to query for information. The applications will not need to know the device specifications and format used by each sensor and device manufacturers. When the sensors are replaced by other models, only the mapping table needs to be updated. The applications are not affected.

Example of using semantic technologies in M2M communication would be a sensor measuring and setting a room temperature on demand of an end user. Devices and sensors register through network gateway which writes their data into semantic database. Every time the sensor sends the temperature, the gateway writes it into the repository, and matches it with desired temperature inside the room where sensor is located.

Implementation of semantic annotation of M2M data enables opening of the secondary M2M market. While the primary M2M market means vertically connected sensor/applications which target a specific business need, re-using of data collected in primary maket with semantic information which help to understand the original data opens up the secondary M2M market. Data can be in this way shared among different verticals and re-used by different applications.

By providing means to understand M2M data, the available business models can be greatly enhanced. For example, through offering additional semantic information about the data, platform provider can enable (and potentially charge for) the discovery of devices and data by semantic specification. Another possible business that can be provided would be to provide derived information from the provided raw data through intelligent processing, e.g. analysing the data, aggregating data across many different data sources, or to provide interpreted data as an additional service.

Adding semantic information to a system can be done in different ways. The possibilites are yet not standardized, but several options are proposed in early draft [5].

Although the question of adding semantic information in M2M data is under standardization, there are still many questions left unaswered. For example:

How exactly should data be annotated in order to be understandable? Should data be universally understandable, or only inside specific vertical domains? Where is the borderline? What is the best way to create extensible annotations, i.e. from minimal semantic information (simple ontology) towards more elaborate? Where do the semantic descriptions (ontologies) come from and who defines the ontologies? Will there be defined fixed catalogue of ontologies?

All of those questions will have to be answered and standardized in the future work in order to reach the interoperability full scale of M2M applications.

V. CONCLUSION

Recently various studies as well as standardization bodies identified the importance of semantic technologies in M2M communication. Semantics can provide machine-interpretable meta-data and annotations and help describing different attributes of M2M devices and data. Implementing standardized semantic technologies into M2M communication system incerases overall interoperability and opens up new business models.

However, semantics also come with an overhead and standards will need to look for a trade-off between efficiency and expressability.

REFERENCES

[1] http://en.wikipedia.org/wiki/Machine_to_machine

[2] Ericsson White Paper, "More than 50 billion connected devices, 284 23-3149 Uen", February 2011

[3] D. Katušić, M. Weber, I. Bojić, G. Ježić, M. Kušel „Market, Standardization, and Regulation Development in Machine-to-Machine Communications", Software, Telecommunications and Computer Networks (SoftCOM), 2012 20th International Conference

[4] IOT – i The Internet of Things Initiative, „D1.3 Analysis of Existing IoT Strategic Research Directions and Priorities", http://www.iot-i.eu/public/public-deliverables/

[5] ETSI TR 101 584 V0.4.0 (2012-09) „Machine to Machine Communications (M2M); Study on Semantic support for M2M Data"

[6] ETSI TS 102 689 V1.1.1, Machine-to-Machine communications (M2M); M2M service requirements

[7] ETSI TS 102 690 V1.1.1, Machine-to-Machine communications (M2M); Functional architecture

[8] ETSI TS 102 921 V1.1.1, Machine-to-Machine communications (M2M); mIa, dIa and mId interfaces

[9] SENSEI – Integrating the physical with the digital world of the network of the future, FP7 project contract no. 215923, December 2010, http://www.ict-sensei.org

[10] Pachube – data infrastructure for the Internet of Things, http://www.pachube.com

[11] Sensor Web Enablement, Open Geospatial Consortium, http://www.opengeospatial.org/projects/sensorweb

[12] Semantic Sensor Network Incubator Group, W3C, http://www.w3.org/2005/Incubator/ssn

[13] Linked Object Data, http://linkeddata.org and http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[14] Joerg Swetina „ETSI M2M / oneM2M and the need for semantics", http://www.iot-week.eu/iot-week-2012/programme-1/tuesday-1/presentations/semantic-interoperability/K1_ETSI-M2M-oneM2M-and-the-need-for-semantics-IoTWeek2012.pdf

[15] http://www.onem2m.org

[16] David Boswarthick, Omar Elloumi, Olivier Hersent, " M2M Communications: A Systems Approach"

[17] Paul Russel, ETSI M2M Architecture Introduction, "A brief overview for Release 1 and Release 2", ETSI M2M Workshop, 24-25 October 2012

[18] Payam Barnaghi, Wei Wang, Cory Henson and Kerry Taylor, "Semantics for the Internet of Things: early progress and back to the future"

[19] Barnaghi, P., "Sense and sens'ability: Semantic data modelling for sensor networks", In Proceedings of the ICT Mobile Summit

[20] Menghan Chen, Beijun Shen, "A Semantic Unification Approach for M2M Applications based on Ontology", 2011 IEEE 7th International conference on Wireless and Mobile Computing, Networking and Communications (WiMob)

[21] Bachir Chihani, Emmanuel Bertin, Noël Crespi, "Enhancing M2M communication with cloud-based context management"

[22] Selvage M., "Achieve semantic interoperability in a SOA, Patterns and best practices", IBM Technical library, Available at: http://www.ibm.com/developerworks/webservices/library/ws-soaseminterop/ index.html

[23] IBM, Bringing Big Data to the Enterprise, avaiable at: http://www-01.ibm.com/software/data/bigdata/

[24] Evans, D. , "The Internet of Things How the Next Evolution of the Internet Is Changing Everything", CISCO white paper, available at http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

[25] Liam Fallon, Declan O'Sullivan, "Using a Semantic Knowledge Base for Communication Service Quality Management in Home Area Networks"

[26] M.Blockstrand, T.Holm, Lars-Orjan Kling, R.Skorg, B. Wallin, "Operator opportunities in the internet of things", Ericsson Review, No.1, 2011

[27] M. Chen, J. Wan, and F. Li, "Machine-to-machine communications: architectures, standards, and applications.", KSII T Internet Inf 6(2): pp. 471–489, 2012

[28] Paul Barom Jeon, Jangwon Kim, Sukhoon Lee, Chonghyun Lee, and Doo-Kwon Baik, "Semantic Negotiation-based Service Framework in an M2M Environment", 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology

[29] 3GPP TS 22.368 v11.2.0, "Service requirements for machine-type communications," Jun.2011.

[30] 3GPP TR 23.888 v1.3.0, "System improvements for machine-type communications," Jun.2011.

[31] 3GPP TR 22.868 V8.0.0 (2007-03), 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on Facilitating Machine to Machine Communication in 3GPP Systems; (Release 8)

Machine-To-Machine Communication as Key Enabler in Smart Metering Systems

Mićo Dujak[1], Vedran Parać[1], Marko Đurasević[2], Ajdin Herić[3]
[1]Ericsson Nikola Tesla d. d., Krapinska 45, 10000 Zagreb, Croatia
[2]Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 0000 Zagreb, Croatia
[3]Faculty of Electrical Engineering, University of Sarajevo, Zmaja od Bosne, 71 000 Sarajevo, Bosnia and Herzegovina
E-mails: {mico.dujak, vedran.parac}@ericsson.com; marko.durasevic@fer.hr; ajdin.heric@gmail.com

ABSTRACT: Machine-to-machine (M2M) communications is a topic that has recently attracted much attention. Smart utilities will include a lot of M2M connected devices. The industry defines M2M communication as the automated communication between remote entities and central management applications. It provides real-time monitoring and control without the need for human intervention. The most represented example in the M2M segment is smart metering management which is defined as using ICT (Information and Communications Technology) technologies in utilities. It is understood as using one system for measuring and collecting energy consumption and other data of two or more of the above-mentioned energy forms.

Diversity of end-user platforms makes the task of collected metering data dissemination very changeling. In this paper, we present a project, which goal was to develop proof of concept for smart metering based on existing M2M Service Enablement (M2M SE) Node developed in Ericsson. M2M Service Enablement is a platform that enables integration of telecom assets and device information with industry specific services and applications in a common user experience. Furthermore, a comparative analysis with the ETSI smart metering standards has also been made. Presented solution collects data from water, gas and electricity meters and stores the data in M2M SE node. Also, the solution is verified as a web application which provides data presentation of collected data to end user in graph form.

KEY WORDS

1. M2M communication 2. Smart Metering 3.Service Enablement

I. INTRODUCTION

M2M refers to technologies that allow both wireless and wired systems to communicate with other devices that have the same ability. M2M uses a device to capture an event, which is sent through the network to an application that translates captured event into meaningful information. M2M communication is the communication between two or more entities that do not necessarily need any direct human intervention. The goal of M2M services is to automate decision and communication processes Smart metering being the most represented example in M2M segment. The same system is used for measuring and collecting energy consumption and other data of two or more energy forms. Smart Metering represents the cornerstone of future smart grids. It provides a range of applications using an infrastructure comprising networked meters, communication networks, data collection and management system.

According to European Energy efficiency action plan individual meters shall be implemented that accurately reflect the end customer's actual energy consumption and provide information on actual time of use [6]. Billing shall be based on actual energy consumption and shall be performed frequently enough to enable customers to regulate their own energy consumption. Usually, a smart meter is considered for registration of electricity and gas use, however water consumption registration is also a possibility.

In this paper we present an alternative approach in energy consumption metering by using smart meters and the M2M Service Enablement Node. The solution allows for different meter data readings to be stored into the M2M SE Node, and then graphically displayed through a web application. The presented approach is additionally validated through a simple system used for energy consumption metering.

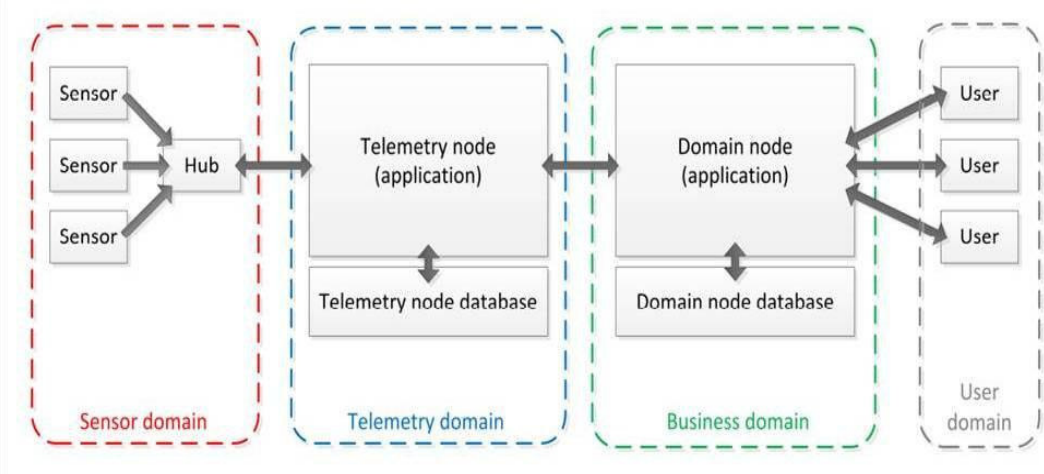


Fig 1 Domains in M2M system

II. M2M SERVICE ENABLEMENT IN SMART METERING SYSTEMS

Machine to Machine (M2M) communications is the communication between two or more entities that do not necessarily need any direct human intervention. M2M services intend to automate decision and communication processes [1]. M2M uses a device (such as sensor or meter) to capture an event (such as temperature, inventory level, etc.), which is relayed through a network (wireless, wired or hybrid) to an application, that translates the captured event into meaningful information [8]. There are four domains in M2M system as shown in Fig. 1.In sensor domain data is sent to a hub over wireless (for example GPRS) or wired network. Telemetry domain is responsible to retrieve data from a hub and store that data in telemetry node database. Also, telemetry domain is responsible to provide web services, so various applications can be developed to present data from sensor domain. Business domain is an application (for example web application) which presents data from telemetry node to end-users.

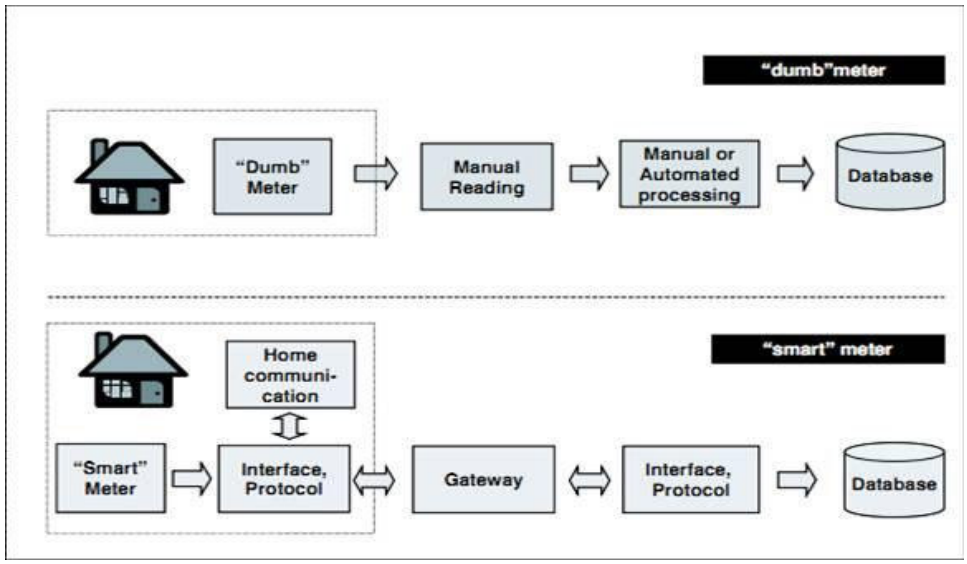


Fig 2 Difference between the conventional and the smart meter data process [2]

Smart metering generally involves the installation of an intelligent meter at residential customers and a regular reading, processing and feedback of consumption data to the customer [2]. Many advantages are attributed to smart metering, including lower metering cost, energy savings for residential customers, more reliability of supply, variable pricing schemes to attract new customers and easier detection of fraud. Additional benefits are foreseen in relation to distributed generation (DG). Smart meter can be used to separately measure electricity delivered by the DG to the grid and the smart metering communication infrastructure can be used to remotely control DG (e.g., in a virtual power plant concept). The differences of smart meters and classical "dumb" meters are shown in Fig. 2. The conclusion which can be drawn on the basis of all the mentioned smart metering characteristics is that in smart metering there is no need for any human intervention, sensor reading and data storing are done automatically [5].
Smart metering is a one of components towards transformation of smart cities.

III. SYSTEM ARCHITECTURE

A simple system has been developed as a proof of concept for the proposed solution, which consists of three main parts: the scheduler, the M2M SE node and the web application. The overall system architecture can be seen on figure 3.
Before the core of the system is described, a few things should be mentioned about the meters first. The meters that were used in this system are normal meters with optical cameras attached to them (meaning they are not real smart meters). Upon the user interaction, by pressing a button, a picture of the current consumption status is taken, and sent to a web server where it is parsed by optical character reader (OCR) software.
The scheduler part of the system periodically pulls the data from the web server with the OCR, parses the received data, and transforms it into a format (an XML document) accepted by the M2M SE Node. Finally, scheduler stores data into the M2M Service Enablement Node. The scheduler also validates the received data, and if data isn't valid (for example if an error occurred during the meter reading) it will not be transferred to the M2M SE Node. It is possible to set the time interval in which the scheduler will pull the data from the web server with the OCR program and transfer it to the M2M SE Node. The scheduler is needed because the M2M SE Node didn't provide the required ability to periodically pull the data from the server with the OCR.
The M2M SE provides a web application interface which in turn provides a possibility to easily store data into the M2M SE Node and to easily fetch that data from the M2M SE Node. The interface for retrieving data from this node allows basic filtering options, e.g. time filtering (for example, filtering for data from a certain time interval), and device filtering, which in turn allows us to request data only from a certain device that was specified. All the data which is retrieved from the M2M SE Node is in the form of an XML document. The M2M SE Node also provides a possibility for user subscription.
The final part of the developed system is a web application which provides a simple user interface from which the user consumption can be checked. This web application offers a variety of different possibilities to visualize the consumption rates, e.g. the possibility to visualize the consumption of water through a single month in a single city [4]. Visualization options are provided for different time intervals (consumption through a year, month or day) and different consumer groups (all consumers in a region or a city or only for a single user).

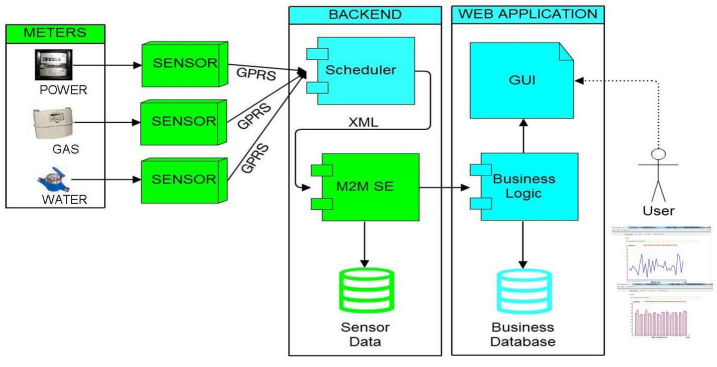


Fig 3 Smart Metering M2M System Architecture

IV. CONCEPT VALIDATION

The proof of concept system is able to collect data from four different meters: gas meter, water meter, single -phase electricity meter and three phase electricity meter. Aside from that, the system supports a two tariff system for the electricity meters. All the meters that were used by this system were classical meters with optical cameras attached to them. Those optical cameras take a photograph of the current meter state, and then send that photograph to a server where the number recognition and parsing takes place. The scheduler then periodically reads the meter data and stores it into the M2M SE Node.

Using those meters, the consumption for multiple users (around 30), during a time period of two years, was simulated. Each user had 6 meters (gas, water, single and three phase electricity meters, both of which had a two tariff system, which were treated as separate meters) and 4 readings a day were simulated for each user. The consumption data for all users was then stored into a single M2M SE Node. During the tests it was shown that the M2M SE Node was able to handle such an amount of data (totaling to more than thirty thousand entries) quite well, and that it fetched and filtered the stored data very efficiently.

As part of the proof of concept, several options for data visualization were developed within the web application. Figure 4 shows an example of a graphical display of electricity consumption through the months of a year (the first tariff is marked blue and the second tariff red). Figure 5, on the other hand shows the water consumption for the days of a month.

The developed system provides a wide variety of advantages when compared to classical meter reading. First of all, meter readings are much more frequent in such a system. The system can easily read the meter state several times in a week, or even day. Aside from that, such meter reading is much more reliable and less error prone. Even if the meter data is incorrectly read, such errors are much easier to find in such a system. All this provides the energy providers with much more control over energy consumption.

Energy providers aren't the only users of the system that benefit from its deployment. Ordinary consumers would have the ability to have a reliable insight into their energy expenditures. This would allow them to adjust their energy consumption on the basis of their past consumption. Later on, even more advanced options could be offered to the consumers. Some of them could be: prepaying for a certain amount of energy for a certain time period, paying the bills through the web application of the system or being warned when a certain amount of energy has been consumed. All things considered, such a system could provide many benefits to both, the providers and the consumers.

V. COMPARATIVE ANALYSIS OF SMART METERING STANDARDS BY ETSI WITH PROOF OF CONCEPT

During the solution development, a short comparative analysis between the developed system and the standards proposed by ETSI in [3] and [8] has been made. This comparison focused mostly on the use case scenarios that the system must satisfy in order for it to adhere to the standards that were proposed in [3] and [8].

The conclusion that was reached is that the developed system does not satisfy the given standards, more specifically, it doesn't support some specifications required by the ETSI standard proposed in [3]. The main reason for this is the fact that the developed system does not use real smart meters for data reading, instead it reads data through optical cameras. One of the requirements for smart meters is to be able to periodically read the meter data and send it to the server, without any user action, which is not the case with optical cameras used in this system, because they needed user interaction for the meter reading to take place. Aside from that, this kind of meter reading, which was used by this system, does not support two-way communication, meaning it doesn't support communication from the supplier to the smart meter located at the users premise. The lack of two-way communication makes it impossible to support the more advanced use cases which require two-way communication between the operator and the meter, e.g. monitoring of supply quality, outages (electricity), network leakage detection (water) and identification of possible meter malfunction, etc.

Although this solution does not comply with the ETSI standards, it does have some advantages. Because the solution is very simple, it is easy to develop applications for such systems. Such a solution, where meters are augmented with optical cameras, is cheaper than solutions with real smart meters (prices for smart meters vary around 100 dollars, even for the cheaper meters).

On the other hand, some disadvantages, apart from the non-compliance with the standards, are that this solution requires user interaction for the readings to take place (this could probably be fixed if the cameras would have an option to automatically read the consumption data). Apart from that, the data which is received has to be interpreted with an OCR program, which could possibly interpret some of the data incorrectly. But if meter readings occur frequently enough, such errors should not pose a big problem, because they could be easily detected and the data marked as invalid or deleted.

Fig 4 Example of electricity consumption through months

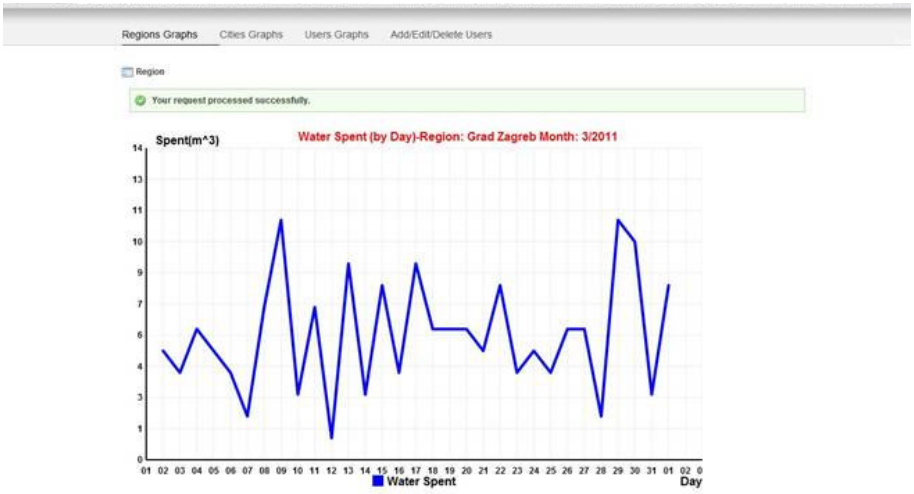


Fig 5 Example of water consumption through a month

VI. CONCLUSION AND FUTURE WORK

The presented system is able to collect data from gas, water and electrical meters with optical cameras using a scheduler which periodically pulls the data from a web service (OCR), stores the data in a single M2M SE node, and then display the data collected in the node. The data is displayed through a simple web application, which is used for retrieving the data from the M2M SE node, and displaying the data in graphical form. The data about consumption can be displayed in a number of different ways, for example, the user can display that data about consumption for all the years in which there is data about consumption or single year, single month, and even a single day. Furthermore, the data can be displayed for a single user, or a group of users located in the same region or city. This system can be further developed by integrating an algorithm, which could detect time intervals in which the energy consumption is high and then offer the users to consume resources in a different time interval, for lower prices, or other benefits. Aside from that, integration with an IPTV would enable users to view their consumption on an IPTV, and thus provides another possibility for the extension and further development of the system.

A comparative analysis between the developed system and the ETSI Use Case standards has also been made. Although many current real-world implementations take advantage of easier deployment of optical cameras attached to existing meters, and provide many benefits to providers and consumers, they do not conform entirely to the full set of smart metering standards.

VII. REFERENCES

[1] Machine to Machine communications (M2M) ; M2M service requirements, European Telecommunications Standards Institute, 2010. ETSI TS 102 689 v1.1.1

[2] Gerven, R, Jaarsma, S, Wilhite, R.. (2006). Smart Metering. Kema, The Netherlands,.

[3] ETSI TR 102 691 v1.1.1 (2010-05) Machine to Machine communications (M2M); Smart Metering Use Cases

[4] Machine-to-Machine communications (M2M); Functional architecture. European Telecommunications Standards Institute. 2011. ETSI TS 102 690 V1.1.1 (2011-10).

[5] Final Guidelines of Good Practice on Regulatory Aspects of Smart Metering for Electricity and Gas. European Regulators Group for Electricity & Gas. Bruxelles, 8 February 2011. Ref: E10-RMF-29-05.

[6] Directive 2006/32/EC of the European parliament and of the Council of 5 April 2006 on energy end-use efficiency and energy services and repealing Council Directive 93/76/EEC.

[7] J.P. Lynch, A. Sundararajan, K.H. Law, A.S. Kiremidjian, T. Kenny, E. Carryer, Embedment of Structural Monitoring Algorithms in a Wireless Sensing Unit. Structural Engineering and Mechanics, Vol. 15, No. 3, p.p. 285-297, 2003.

[8] CEN/CLC/ETSI/TR 50572 Functional reference architecture for communications in smart metering systems

DEVELOPMENT METHODS

Improved Distributed Band-Preference Dynamic Spectrum Management for Upstream VDSL2

Yao Gao, Youming Li, Gang Wang and Ming Jin
College of Information Science and Engineering
Ningbo University, Ningbo 315211, China
Email: liyouming@nbu.edu.cn

***Abstract*—Crosstalk is a key factor that can cause performance degradation in VDSL2. Balancing the transmit power spectra refers to as spectrum management to diminish the impact of crosstalk. The distributed band-preference spectrum management (DBPSM) adopts a bit-moving process to highlight the data rates of weak users. However, it still leaves room for improvement. This paper proposes a new tone and band selection method to improve the performance of DBPSM and proves that the new tone selection method makes the bit-moving process more efficient. Simulation results show that the new band selection method can increase the data rates of weak users.**

***Index Terms*—VDSL2, crosstalk, distributed algorithm, bit-moving, dynamic spectrum management**

I. Introduction

Until the accomplishment of a full Fiber Network, digital subscriber line (DSL) remains the most popular broadband access technology. The skyrocketing demand for higher data rates forces DSL systems to extend bandwidth up to 30MHz for VDSL2. However, using high frequency causes significant far-end crosstalk (FEXT), which is the major obstacle to achieve higher data rates. Thus excellent crosstalk suppression is critical [1] [2] [3].

Dynamic spectrum management (DSM) provides an effective framework to reduce crosstalk. It can be categorized into three levels of coordination, namely DSM 1, DSM 2 and DSM 3 [4]. In DSM1, each modem works autonomously with a spot of information such as the target bit rates, the models of spectrum management. In DSM 2, a spectrum management center (SMC) optimizes the power spectra of all modems centrally to reduce the mutual interference [5] [6]. However, much information for spectrum management has to be computed and distributed by the SMC, and in some cases, the SMC cannot adapt to rapid changing of channel information. In DSM 3, users terminate at least in one end of the loops and crosstalk is canceled by signal processing jointly. However there are some scenarios, like joint central office (CO) / remote terminal (RT) deployments, where joint processing is not possible.

This paper focuses on DSM1. One typical method in DSM 1 is iterative water-filling (IWF) in which each user runs a single-user water-filling process and seeks to maximize its own data rate [7] [8] [9]. Although IWF does not require any central controller and has low complexity, it does not perform well particularly in near-far scenarios such as upstream VDSL2 [10]. To mitigate the interference from strong users and improve the performance of weak users, distributed band-preference spectrum management (DBPSM) is proposed in [11]. Here, the strong users adopt a bit-moving process after the normal bit-loading process [12]. Although DBPSM leads to distinct performance gain, there is still non-negligible room to improve.

In this paper, we propose an improved DBPSM algorithm. According to the characteristics of loading one more bit in a band with exponentially growing incremental power and IWF bit-loading process, we first design a novel band selection scheme, in which all the high frequency bands are considered as a whole, and bits are added to all the tones in the high frequency bands in the bit-moving process. While for the previous method, the high frequency bands are considered separately. Meanwhile, we also introduce a novel tone selection method. It first sorts the tones based on the size of incremental power, and then select the tone according to the sort order which is proved to be faster than tone selection method in DBPSM. Simulation results show that the performance of improved DBPSM achieves significant enhancement than that of DBPSM.

The remainder of this paper is organized as follows: Section II presents the system model of multi-user DSL. Section III formulates the spectrum management problem and proposes improved DBPSM algorithm. Section IV shows and interprets the simulation results. Section V finally summarizes conclusions for improved DBPSM.

II. System Model

VDSL2 systems use discrete multi-tone (DMT) modulation that splits the available bandwidth into many subcarriers also called tones. The transmission for a binder of N modems can be modeled independently on each tone by

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{z}_k, k = 1, \cdots, K \quad . \tag{1}$$

The vector $\mathbf{y}_k = \left[y_k^1, \cdots, y_k^N \right]^T$ contains the received signals on tone k for all N modems. $h_k^{i,j} = \left[\mathbf{H}_k \right]_{i,j}, i \neq j$ represents the crosstalk channel from the transmitter j to the receiver i on tone k. The diagonal elements are the direct channels. $\mathbf{x}_k = \left[x_k^1, \cdots, x_k^N \right]^T$ contains the transmitted symbols on tone k. $\mathbf{z}_k$ is the vector of additive noise on tone k and contains alien crosstalk, thermal noise, RFI etc. K is the number of upstream tones.

In this model, we assume there is no signal coordination between lines and the signals from other users are treated as noise. Under this assumption, the achievable rate per DMT-symbol for modem n on tone k is

$$\begin{aligned} b_k^n &= \log_2 \left(1 + \frac{1}{\Gamma} \frac{\left| h_k^{n,n} \right|^2 s_k^n}{\sum_{m \neq n} \left| h_k^{n,m} \right|^2 s_k^m + \sigma_k^n} \right). \\ &= \log_2 \left(1 + \frac{g_k^n s_k^n}{\Gamma} \right) \end{aligned} \tag{2}$$

Where Γ is the SNR-gap to capacity and is a function of the desired bit error rate (BER), coding gain and noise margin, $s_k^n = \varepsilon \left\{ \left| x_k^n \right|^2 \right\}$, $\sigma_k^n = \varepsilon \left\{ \left| z_k^n \right|^2 \right\}$ are the transmit power spectrum density (PSD) and the noise PSD on tone k for modem n respectively, $g_k^n = \dfrac{\left| h_k^{n,n} \right|^2}{\sum_{m \neq n} \left| h_k^{n,m} \right|^2 s_k^m + \sigma_k^n}$ is the normalized channel gain. From (2), the incremental power to load a bit can be expressed as

$$\begin{aligned} \Delta p\left(b_k^n \right) &= p\left(b_k^n \right) - p\left(b_k^n - 1 \right) \\ &= \Gamma \left(2^{b_k^n} - 1 \right) \Big/ g_k^n - \Gamma \left(2^{b_k^n - 1} - 1 \right) \Big/ g_k^n \\ &= \Gamma 2^{b_k^n - 1} \Big/ g_k^n \end{aligned} \quad . \tag{3}$$

The data rate for modem n is

$$R^n = f_s \sum_k b_k^n \quad . \tag{4}$$

f_s is symbol rate. The total power used by modem n is

$$P^n = \sum_k s_k^n \quad . \tag{5}$$

III. Spectrum Management

A. The spectrum management problem

We restrict attention to the two user case for ease of explanation as depicted in Fig. 1, where user 1 assumed to be far-located and is much weaker than user 2. The spectrum management problem is to maximize the rate of user 1 while user 2 is just subject to achieving the target date-rate. This can be formulated as

$$\begin{aligned} &\max_{\mathbf{s}^1, \mathbf{s}^2} R^1 \\ &s.t. \quad R^2 \geq R^{2,target} \end{aligned} \quad . \tag{6}$$

The optimization (6) is typically subject to a spectral mask constraint on each tone and a total power constraint on each modem as follows

$$0 \leq s_k^n \leq s_k^{n,mask}, \forall k, \forall n \quad , \tag{7}$$

$$\sum_{k \in K} s_k^n \leq P^{n,tot}, \forall n \quad , \tag{8}$$

where $\mathbf{S}^n = \left[s_1^n, \cdots, s_K^n \right]^T, n = 1$ or 2, contains the transmitted power for user n on all K tones. R^1 denotes the rate of user 1, $R^{2,target}$ is the target rate of user 2, $s_k^{n,mask}$ is the spectral mask constraint of user n on tone k and $P^{n,tot}$ is the total power budget for user n.

B. Improved distributed band-preference spectrum management algorithm

In DBPSM, consecutive tones are autonomously grouped into one band after the normal bit-loading process as in IWF [11]. In order to protect weak users, strong users move bits from low frequency bands to high frequency bands. In the bit-moving process, bits are first moved from the first band to the last band. Once a bit can no longer be removed from the first band or a bit can no longer be added to the worst band, the second or the last but one is considered. This process ends when all the bands are considered. Fig. 2 shows the PSDs of DBPSM when the target rate of user 2 is 104 Mbps. From it, we can get the PSD of region A is larger than that of region B. However, direct channel gains of region A are larger than those of region B. Moreover, region A and region B are not occupied by weak user. Thus, the PSD of region A should be larger than that of region B.

To avoid this situation, an improved DBPSM algorithm is proposed in this section. All bands are first divided into two regions. The first region just contains the first band, and the second contains all subsequent bands. Then, bits are moved from the first region to the second region, as long as the limitations of (7) and (8) are met. Once a bit can no longer be moved, the first band is discarded and the rest of bands is continued to be divided into two regions. Now the new first region contains the second band, and the new second region contains all subsequent bands. The process finishes when the second region just contains the last band and cannot be divided.

As in [11], one bit is removed from the tone that needs the largest energy in the first region. In DBPSM, each time there is a bit to move, we have to compare the energy of all tones in the first region in order to find the needed tone. In this paper, we will propose a fast tone selection method. It is presented as follows: For the first region, firstly, we sort the tones from large to small based on the size of incremental power; then we just select the tone in order. If the tone is the last of the sorted tones, we will back to the first tone next time which is called one loop. The following theorem shows that the tone selected by the new method is which used the largest energy.

Theorem 1: Define $k_\max = \arg\max_k \Delta p\left(b_{kE}^n \right)$, $b_{kE}^n > 0$ and let $k_\max = iE$ when the l^{th} bit is moved. As the $(l+1)^{th}$ bit is moved, we can conclude

$$k_\max = (i+1)E, \quad if \quad i+1 \le K, b^n_{(i+1)E} > 0, \tag{9}$$

where kE is the k^{th} tone after sorted.

Proof: See Appendix. ■

Table I shows the complexity of the new tone selection method and the tone selection method in DBPSM. S is the number of moving bits, U is the number of tones in the first region. Normally, S is much larger than U.

To the second region, the tone that needs the least energy is selected to add one bit in the bit-moving process. Similarly, we can adopt the above method except that we sort the tones from small to large based on the incremental power. The new tone selection method can also get the needed tone as there is least one bit in each tone of the second region. For strong users, it is usually able to meet. The entire detailed process is presented in Algorithm 1. The new tone selection method also can be adopted in DBPSM. In the previous section, we have only presented the two-user situation. Nevertheless, the situation of more users can be handled as in [10].

IV. SIMULATION

In this section, the performance of the OSB, IWF, DBPSM and IDBPSM are compared. Based on France Telecom measured data, we simulate a two-user VDSL2 upstream scenario as depicted in Fig. 1, where user 1 is distantly located and user 2 is closely located. The VDSL2 upstream frequency bands are 3.75-5.2MHz, 8.5-12MHz and 18.1-30MHz. The target symbol error probability is 10^{-7}. The SNR-gap is set to 19.5236, the noise PSD is -140 dBm/Hz and the symbol rate f_s is 8 KHz. The PSD constraint is set at -56.5 dBm/Hz. However, it is not applied to IWF and OSB. The total power is 11 dBm.

Fig. 3 shows the rate region. The PSDs are depicted in Fig. 4, Fig. 5 and Fig. 6 when the target rate of the 300 m. line is 104 Mbps. Compared with OSB, the performance of IWF degrades significantly. According to Fig. 4 and Fig. 5, we can discover that the primary reason for this gap is the strong interference in the low frequency induced by the strong user. Fig. 3 also shows that the DBPSM's performance achieves much gain after the bit-moving process and is close to that of the OSB. More significantly, the performance of IDBPSM is better than that of DBPSM. For example, when the target rate of the 300 m user is set as 104 Mbps, the 600 m user achieves 24.272 Mbps in the DBPSM and 27.56 Mbps in the IDBPSM. Therefore, the DBPSM achieves 286% increase over the IWF and achieves 81.24% of the OSB. The IDBPSM achieves 325% increase over the IWF and achieves 92.24% of the OSB. According to Fig. 2

TABLE I. COMPARISON THE COMPLEXITY OF THE NEW TONE SELECTION METHOD AND THE TONE SELECTION METHOD IN DBPSM

Tone selection Method	Complexity
New	$O(U^2)$
In DBPSM	$O(US)$

Algorithm 1 Bit-Moving process

Initialize
Divide the all tones into M bands: $B_1, B_2, \cdots, B_M$
$i = 1$
Iteration:
While $i < M$ **do**
  $k_\max = \max(c), c \in B_i$
  $B_de = \arg sort_c\ \Delta p(b_c^n), c \in B_i$ in descending order
  $B_as = \arg sort_c \Delta p(b_c^n), c \in \{k_\max+1, \cdots, K\}$ in ascending order
  $k = 0, l = 0$
  while $\sum_{c=1}^{K} p(b_c^n) < P^{n,tot}$ **do**
    $k = k+1, l = l+1$
    if $k > k_\max$ **then** $k = 1$
    if $l > K - k_\max$ **then** $l = 1$
    if $b^n_{B_de(k)} = 0$ **then** $k = k+1$
    else if $p(b^n_{B_as(l)} + 1) > s^{n,mask}_{B_as(l)}$ **then** $l = l+1$
    else if
    $\sum_{c=1}^{K} p(b_c^n) + \Delta p(b^n_{B_as(l)} + 1) - \Delta p(b^n_{B_de(k)}) / 2 > P^{n,tot}$
    then $i = i+1$; **break**;
    else
      $b^n_{B_de(k)} = b^n_{B_de(k)} - 1,\ b^n_{B_as(l)} = b^n_{B_as(l)} + 1$
      $p(b^n_{B_de(k)}) = p(b^n_{B_de(k)}) - \Delta p(b^n_{B_de(k)})$
      $p(b^n_{B_as(l)}) = p(b^n_{B_de(l)}) + \Delta p(b^n_{B_de(l)} + 1)$
      updating $\Delta p(b^n_{B_de(k)})$ and $\Delta p(b^n_{B_as(l)})$ using (3)
    end if
  end while
end while

and Fig. 6, we can speculate that the number of bits moved in the improved DBPSM is more than that moved in the DBPSM. Comparing the simulation time of bit-moving process in Fig. 7, although more bits are moved in improved DBPSM, the bit-moving process in improved DBPSM is roughly four times faster than that in DBPSM.

V. CONCLUSION

In this paper we presented an improved DBPSM algorithm. Compared with the DBPSM algorithm, the improved DBPSM algorithm changes the band selection method and the tone selection method in the bit-moving process. It is proved that the new tone selection method makes the bit-moving process more efficient. Numeric simulations show that the improved DBPSM algorithm yields performance gains over the DBPSM algorithm.

ACKNOWLEDGMENT

This work was partially supported by National Science Foundation of China under Grant 61071119, Natural Science Foundation of Ningbo Municipality under Grant (2012A610017)，the Innovation Team of Ningbo (2011B81002).

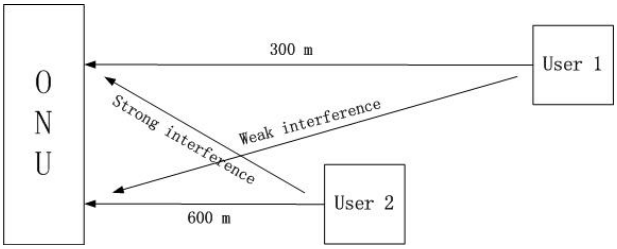


Fig. 1. Upstream VDSL2 scenario for two users

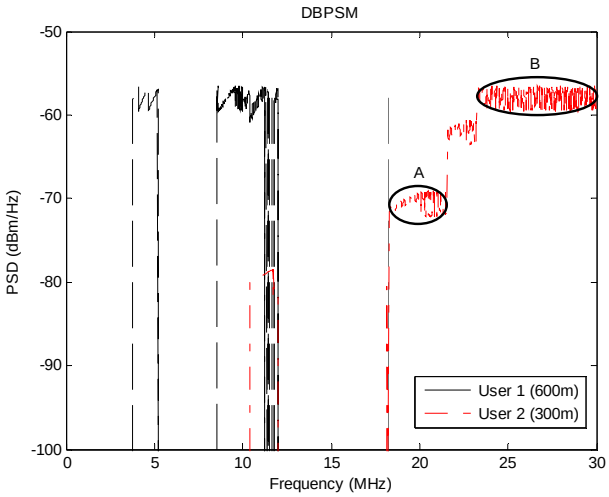


Fig. 2. PSDs of DBPSM with 104Mbps-user 2 target rate, 11dBm-total power constraint, -56.5dBm-spectral mask constraint

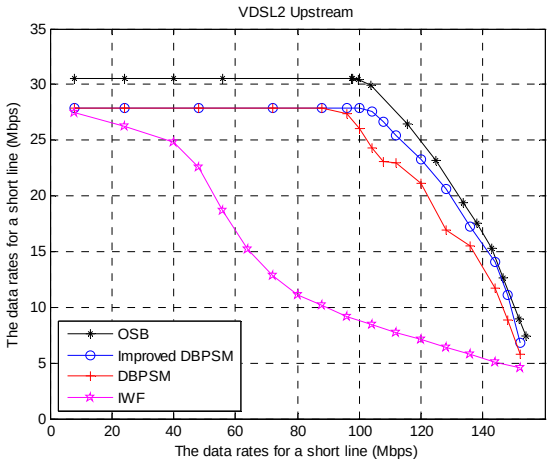


Fig. 3. Upstream rate regions of improved DBPSM, DBPSM, OSB and IWF

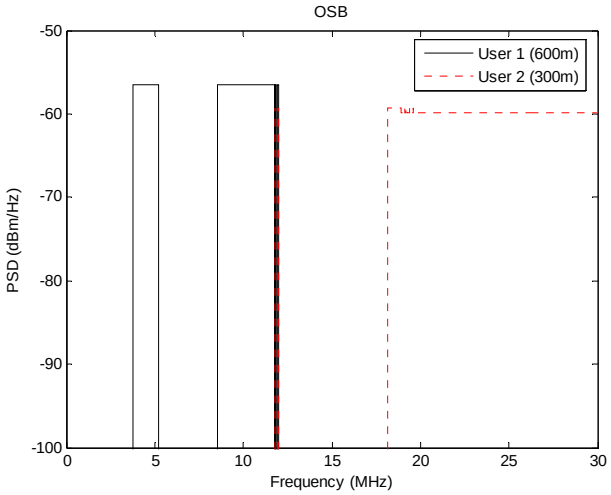


Fig. 4. PSDs of OSB when $R_2^{target} = 104$ Mbps

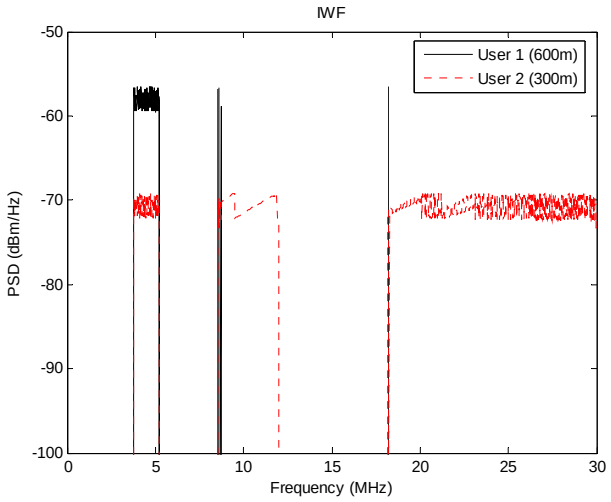


Fig. 5. PSDs of IWF when $R_2^{target} = 104$ Mbps

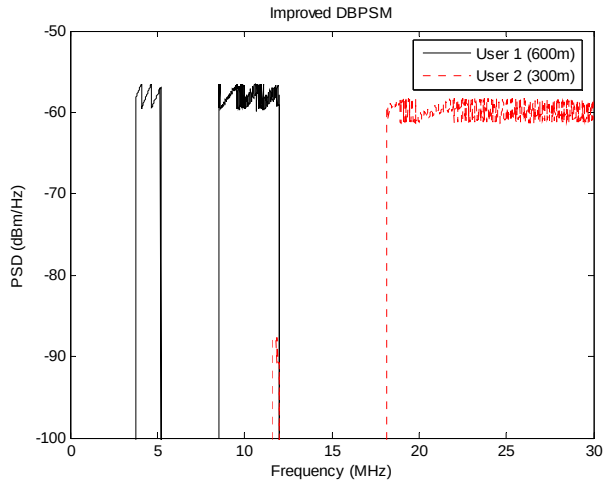


Fig. 6. PSDs of improved DBPSM when $R_2^{target} = 104$ Mbps

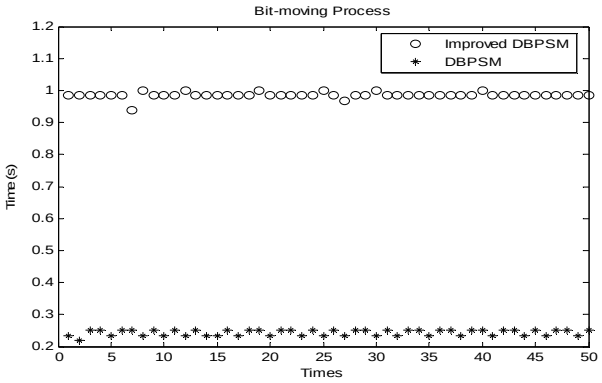


Fig. 7. Simulation time of bit-moving process in improved DBPSM and DBPSM

APPENDIX

Proof of Theorem 1

Here is the open statement, $S(t)$: We moved t bits and will move the $(t+1)^{th}$ bit.

1) $S(1)$: obviously $k_\max = 1$. Hence

$$b_{1E}^n = b_{1E,I}^n - 1, \tag{10}$$

where $b_{k,I}^n$ is the number of bits of user N on tone k after the normal bit-loading process. Next we will prove the following conclusion holds

$$\Delta p\left(b_{1E}^n\right) \le \Delta p\left(b_{2E}^n\right), \tag{11}$$

if $\Delta p\left(b_{1E}^n\right) > \Delta p\left(b_{2E}^n\right)$ holds, then

$$\Delta p\left(b_{1E,I}^n\right) > \Delta p\left(b_{2E,I}^n + 1\right), \tag{12}$$

it is contradictory to the normal bit-loading process and so (11) is true. Moreover, we can easily get

$$\Delta p\left(b_{2E}^n\right) > \Delta p\left(b_{3E}^n\right) > \cdots > \Delta p\left(b_{KE}^n\right), \tag{13}$$

combining (11) with (13) implies

$$k_\max = 2, \tag{14}$$

therefore $S(1)$ is true and we have our basis step and a starting point from which to begin the induction.

2) Assuming Theorem 1 is true for $S(1), S(2), \cdots, S(m-1)$, we want to establish our inductive step by how the truths of $S(1), S(2), \cdots, S(m-1)$ "force" us to accept the truth of $S(m)$. The assumption of the truths of $S(1), S(2), \cdots, S(m-1)$ is our induction hypothesis, and $k_\max = jE$ as the m^{th} bit is moved. To establish the truth of $S(m)$, we need to show that

$$k_\max = (j+1)E, \quad if \quad j+1 \le K, b_{(j+1)E}^n > 0. \tag{15}$$

According to the assumption, we can get

$$\left(b_{(k1)E,I}^n - b_{(k1)E}^n\right) - \left(b_{(k2)E,I}^n - b_{(k2)E}^n\right) = 1$$
$$k1 \in \left\{1, \cdots, j \middle| b_{(k1)E}^n > 0\right\}, k2 \in \left\{j+1, \cdots, U \middle| b_{(k2)E}^n > 0\right\}, \tag{16}$$

in the first step we want to prove

$$\Delta p\left(b_{(j+1)E}^n\right) \ge \Delta p\left(b_{(k1)E}^n\right), \tag{17}$$

assuming there is a tone $f\ (f \in K1)$ such that

$$\Delta p\left(b_{fE}^n\right) > \Delta p\left(b_{(j+1)E}^n\right), \tag{18}$$

combining this with (3) implies

$$\Delta p\left(b_{fE,I}^n\right) > \Delta p\left(b_{(j+1)E,I}^n + 1\right), \tag{19}$$

it is contradictory to normal bit-loading process and so (17) is true.

Now we just need to prove

$$j+1 = \arg\max_k \Delta p\left(b_{kE}^n\right), k \in k2, \tag{20}$$

assuming there is a tone f such that

$$\Delta p\left(b_{fE}^n\right) > \Delta p\left(b_{(j+1)E}^n\right), f \in k2, f \ne j, \tag{21}$$

combining this with (3) implies

$$\Delta p\left(b_{fE,I}^n\right) > \Delta p\left(b_{(j+1)E,I}^n\right), \tag{22}$$

it is contrary to $\Delta p\left(b_{1E,I}^n\right) > \Delta p\left(b_{2E,I}^n\right) > \cdots > \Delta p\left(b_{KE,I}^n\right)$ and so (15) is true. (11) and (15) established the inductive step of Theorem 2.

Consequently, by the Principle of Finite Induction, Theorem 2 is true.

REFERENCES

[1] Paschalis Tsiaflakis, Yung Yi, Mung Chiang and Marc Moonen, "Green DSL: energy-efficient DSM," in IEEE ICC, Dresden, Germany, pp. 1-5, Jun. 2009.

[2] Ming-yang Chen, Wonjong Rhee, Mehdi Mohseni and John M.Cioffi, " Distributed Crosstalk Management for Upstream VDSL Using Dynamic Power Control," IEEE Trans. Commun., vol. 60, no. 4, pp. 940-945, Apr. 2012.

[3] R. Cendrillon, J. Huang , M. Chiang and M. Moonen, "Autonomous Spectrum Balancing for Digital Subscriber Lines," IEEE Trans. Signal Processing, vol. 55, no. 8, pp. 4241-4257, Aug. 2007.

[4] Amir R. Forouzan, Marc Moonen, Jochen Maes and Mamoun Guenach, " Joint Level 2 and 3 Dynamic Spectrum Management for Upstream VDSL," IEEE trans. Commun., vol. 59, no. 10, pp. 2851-2861, Jul. 2011.

[5] Hao Zou, Aakanksha Chowdhery and John M. Cioffi, "A centralized Multi-Level Water-Filling Algorithm for Dynamic Spectrum Management," in ACSSC, Pacific Grove, CA, pp. 1101-1105, Nov. 2009.

[6] Raphael Cendrillon, Wei Yu, Marc Moonen, Jan Verlinden and Tom Bostoen, "Optimal multiuser spectrum management for digital subscriber lines," in IEEE ICC, Paris, France, vol. 1, pp. 1-5, Jul. 2004.

[7] Sean Huberman, Christopher Leung, and Tho Le-Ngoc. "Dynamic Spectrum Management (DSM) Algorithms for Multi-user xDSL," IEEE Commun. Surveys & Tutorials, vol. 14, no. 1, pp. 109-130, First Quarter, 2012.

[8] Wei Yu, "Multiuser water-filling in the presence of crosstalk," in Information Theory and Applications Workshop, La Jolla, CA, pp. 414-420, Jan. 2007.

[9] Sean Huberman, Christopher Leung, and Tho Le-Ngoc, "A Clustering Approach to Autonomous Spectrum Balancing Using Multiple Reference Lines for DSL," in IEEE GLOBECOM, Houston, USA, pp. 1-5, Dec. 2011.

[10] Oshri Naparstek, Kobi Cohen and Amir Leshem, "Parametric Spectrum Shaping for Downstream Spectrum Management of Digital Subscriber Lines," IEEE Commun. Lett, vol. 16, no. 3, pp. 417-419, Mar. 2012.

[11] Wooyul Lee, Youngjae Kim, Mark H Brady and John M

Cioffi, ”Distributed spectrum optimization for multi-carrier interference channels,” Ph. D. dissertation, Stanford University, CA, 2009.

[12] J Campello, ”Optimal discrete bit loading for multicarrier modulation systems,” in IEEE International Symposium on Information Theory, Cambridge, MA, pp. 193, Aug. 1998.

Dictionary Based Optimization for Adaptive Compression Techniques

Ahmed Mokhtar A. Mansour* and Mona A. M. Fouad **
* NI, Giza, Egypt
** NTI, Cairo, Egypt
Ahmedmokhtar_adu@yahoo.com, drmfouad08@gmail.com

Abstract - This paper proposes an enhancement of the two levels dictionary based compression. The enhancement is based on optimizing the mapping tables overhead of the second level compression, usually variable length code, that used over the first level compression (the dictionary based one). The case study will be applied on 'A-M' compression and will introduce the 'A-M' dictionary version 2. This work guarantees the reduction of overhead from the second level compression up to 45% of the original mapping table. Besides, the use of 'A-M' dictionary reduces the processing effort of recreation of dynamic dictionary. The idea is based on reducing the dictionary field of the mapping table by sorting, grouping, and then eliminating common parts (6 Most Significant Bits - MSB) of the dictionary words during compression, which is reconstructed back during decompression. The process is completely transparent with respect to the two levels of compression (dictionary and variable length) and can be applied not only for the 'A-M' compression, but also for any other dictionary based compression technique.

I. INTRODUCTION

Text compression is used long time ago in order to minimize the space of storage of documents. The telecom revolution through internet and mobile communication get use of the existing compression techniques and encouraged the use of compression as a solution to save bandwidth. There are many techniques of compression. In text compression, where changing only one character might change all the meaning, lossless compression is the most important one. Many compression techniques are introduced as lossless compression techniques for text [1] [2]. Arithmetic compression is introduced as an algorithm that transfers group of characters 'word' into a number [3] [4]. The first step is to identify the initial probability of occurrence of each character and build the corresponding probability table. The algorithm creates a variable length number for each group of characters 'word' and replace the word by the number in its bit format. The disadvantage of the code appears directly if the length of words is long and the character set is long or not correctly mapped. In such cases both the processing time and resulting code words will increase dramatically.

Another lossless dictionary based compression is LZW [5] [6]. LZW depends on dynamic creation of mapping table based on an initial mapping table as a header of the coded (compressed) document. LZW generation of the dictionary means a single occurrence of the encoded array appeared on the document. However, it doesn't mean that there will be another appearance of the same string of characters. Using 12 bit to encode all the variable string means that the encoding space will be filled completely very fast with possibility of having non useful data or having new characters appearing after the occupation of the whole mapping space. Such a problem might cause the algorithm to come out with an error. Besides, a portion of the mapping table (dictionary) is not really used efficiently.

One more famous lossless compression technique is Huffman code. Huffman code is not a dictionary based one but it can be used alone or above any dictionary based compression as a second level compression. 'A-M' code and LZW code recommends the use of Huffman as a second level coding in many cases. Huffman is based on creating a weight of each individual symbol (character) depending on the frequency of occurrence of that symbol. The more the occurrence will be the shorter the code should be. Huffman counts the characters in a document and assign to each character a binary code. Huffman uses a recursive operation of arrangement of characters by their weights and building a mapping table for all the characters. The mapping table is sent as an overhead with the same document. If the number of the mapped symbols (characters) increases, the mapping table size will increase as well. Part of the work introduced in this paper is how to reduce the mapping table size of the second level compression. Shannon-Fano algorithm [2] [7] introduces another lossless compression technique that might be used standalone or as a second level compression one. It is similar to Huffman in the concept, however the algorithm is sorting the frequency of occurrence of symbols (characters) in a descending order rather than the ascending one (or balanced on other word) of Huffman [2] [8]. Shannon-Fano algorithm results in almost similar results as Huffman, however, it has been shown in some cases that Huffman is giving better results than Shanon-Fano. In this paper we are using Huffman as example of second level compression algorithm. The same technique can be used with Shannon-Fano as well with the same gain rate. There exist several modifications for the previously discussed compression methods [9] [10] [11]. This paper introduces an enhancement of the dictionary based compression techniques by proposing a new version of 'A-M' dictionary and using this version to optimize the second level compression algorithm by minimizing the coding overhead.

II. 'A-M' DICTIONARY BASED OPTIMIZATION

The 'A-M' Frame work provides high compression ratio on text files. The compression is two level compression that provides a dictionary based compression in its first level with aid of a static dictionary (dictionary version may varies), and the second level is a variable length code compression (usually Huffman code is used). In first version of the 'A-M' dictionary (version 1), variable word length were used with MSB used as indicator of the length, this requires processing in the bit level which results in heavy processing during decompression. The developing of A-M (version 2) compression is important for two reasons, the first one is to increase the space of the encoded records; secondly, to facilitate the optimization mechanism during the second level compression.

In version 1 dictionary, there were three types of records, the compound element records, the special element records, and the simple element records. The total number of the records was 3392 and the available records were 3472. In version 2, the number of available record will increase as well as the number of used records. Besides, the number of compound elements is increased so that the compression ratio will increase as well. This paper will explain in details the specifications of 'A-M' dictionary optimization from version 1 to version 2; moreover, a new mapping table technique will be introduced to minimize the overhead of the mapping table.

A. Specifications of 'A-M' Dictionary (version 1)

The basic idea of 'A-M' Dictionary is to build a pre-defined dictionary for each language. The dictionary will be saved either at the source or the destination (encoding and decoding) or in the cyber space and to be obtained during the encoding and decoding process. In both version 1 and version 2 dictionary, a pre-defined dictionary will be used in the coding and encoding process, the version number identifies which dictionary to be used, a size of one version of the dictionary is less than 16 Kbyte. Dictionary version 1 will be implemented as follows:

For one identified language, the dictionary is built from all the standard ASCII characters from 0 – 127, plus all the combinations of two alphabetical letters. In order to do so, twelve bit length words will be used. Frequently used characters such as full-stop, comma, space, etc. are added and treated in as special case. The following algorithm is used to build the dictionary:

a) The 12 bit space is divided into 2 parts from 000.000000000 to 111.111111111 and the records of 0.x and 7.x classes will be reserved now.

b) The classes from 1.0 up to 6.511 are reserved for 12 length words and avails exactly 3072 different codes. Each record is called the compound record and is formed from two characters; each character is called a compound element. Having 3072 available record allows covering more than the 52 capital and small letters of the English language, as a result, other five special characters are added in dictionary version 1 compound elements. Namely, the comma, the space, the full stop, the new line, and the NULL. The reserved 000.1XXXXXXXX are offering more 256 bits that might be used as extension for and supplementary characters for other languages.

c) The reserved record 0.x and 7.x will be used in different manners. For 0.x; it's divided into two parts, first is 0.0x and 0.1x. 0.1x is reserved as mentioned before. 0.0x or in other words 0000xxxx will be dealt with as only eight bit records, leading 0000 means that this is an eight bit record and will be used for numbers and the special characters (full stop, space, comma, new line). These records are named the special records. The 7.x records will be reserved for the normal 128 ASCII characters, this record will be only 10 bit length rather than 12 bits. The leading 111 bits means that we will read only another 7 bits. In other words, the record length will be only 10 bit if the leading MSB are 111.

'A-M' Version 1 dictionary has its own algorithm for encoding and decoding that takes into consideration the variable length of records. The coding and encoding algorithm shown in [1] tries to identify the group which the character belongs to and search the dictionary to map the character(s) to its right code. The decoding process is doing the same but on the opposite direction and taking into consideration the variable length of each code. The MSB helps to identify the right mapping. Figure 1 summarizes the process used if version 1 dictionary is used in the 'A-M' Framework. Figure 1 summarizes 'A-M' version 1 dictionary based compression algorithm.

The standing alone version 1 based compression process ensures up to 25% compression based on the fact that each 2 characters will use only 12 bit space rather than 16 bit space. Having many single characters on the text such as non alphabetical characters will reduce the compression ratio. Having 256 free records can be used to enhance the compression ratio to exceed the 25%; however, the compression algorithm should be modified to do so.

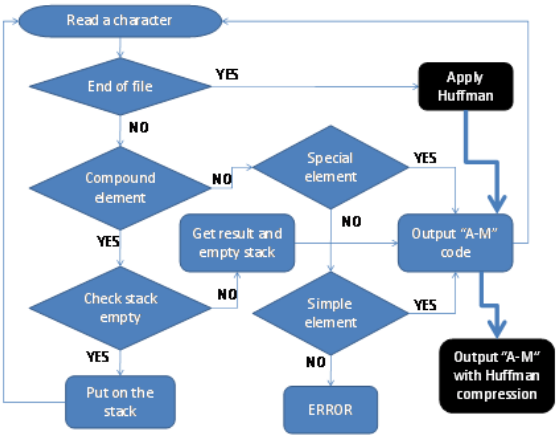


Figure 1: version 1 'A-M' framework [1]

B. 'A-M' dictionary version 2

The 'A-M' dictionary version 2 uses fixed length records of 12 bits. Most of the records will contain a mapping of two ASCII characters as in version 1. This means a direct compression ratio of 25%. The ratio results

from using 12 bit rather than 16 bits for two ASCII characters. The treatment and the code are different from version 1. The records in this version are divided into two parts. Each part of 6 bits to be of the form of x.y while x and y may vary from 0 to 63. The records starting from 0.0 and terminates at 63.63 (or from 000000.000000 to 111111.111111). There are five classes of records, namely; reserved, normal, alphabetical, numerical, and frequent words records. The details of each class and its mapping in the dictionary are as follow:

a) The reserved records are used for flagging in both first and second level compression. The record 0.0 is reserved as the general flag key. All records of the form X.0 are reserved to avoid ambiguity. All reserved records belong to the reserved class. Further explanation is given in the next section.

b) The records 1.y and 2.y are reserved for the 128 normal ASCII characters. Those records are called normal records and they belong to the normal class.

c) The records 3.y up to 59.y are used for alphabetical records. As 'A-M" dictionary version two identifies 60 different character (for the English dictionary, it will be 26 small character, 26 capital character and eight extra characters, namely, space, comma, newline, full-stop, semicolon, open braces, close braces, and question mark. In order to cover any combination of two characters from this pool of character we need 3600 records (60*60). The mentioned pool of record is only 3591 record (57 * 63) as all the x.0 are reserved and not used. The required 9 records will be taken from 60.1 up to 60.10. Covering from aa up to ZZ is the same exactly like version 1, however, some extra characters are added to the alphabetical records. This addition will increase the rate of compression. Each individual character in this class is called basic character.

d) The records from 60.11 up to 61.47 will be used for numerical combinations from 00 to 99. Those records are called the numerical records and belong to the numerical class. This is another addition that has no similarity in version 1.

e) The final class is the frequent words' class; the records of 62.y and 63.y are used for such a purpose. 61.48 to 61.63 can also be used for the same purpose. For facility, only 62.y and 63.y are used. In version 1, 0001xxxxxxxx could be used for the same purpose but the algorithm should be changed to do so. The used algorithm in version 2; shown in the next sections might be adopted with the necessary customization for version 1 enhancement. The use of frequent words class implies flagging the frequent words before starting the compression algorithm. The list of the frequent words contains 126 words of more than 2 characters. Examples of these words are (the, his, her, this, that, which, where, have, had, has, been, however, next, after, before, etc …). Knowing that the most 100 used words in English contains 25% of 2 character or less words, then only 75 word will be added, this will give a space for another 51 word (most of them are lemmas of verbs founded in the 75 words list). The selected words are based on the linguistic researches of each language. The longer the word, the higher the compression rate achieved. For each existing record in the frequent words list, the compression rate will vary from 50% for 3 letters words (the, her, his, etc…) to achieve 80% compression ratio for 5 letters words such as (other, under, above, etc…). The algorithm to flag frequent words before starting the compression will be detailed in the next section.

The 'A-M' dictionary version 2 uses flagging for two different reasons. The first reason is used before the first level compression, which is frequent words flagging or flag sequence alike flagging. The second use of flagging is done after the second level compression. The flagging is used to reduce the mapping table size. The first level flagging is described in the 'A-M' code generation section. The second level flagging is described in the mapping table optimization section.

C. 'A-M' code generation for dictionary version 2

In order to create a code from a given text document, the dictionary should exist either in the local storage or be retrieved from the cyber space; the right dictionary version should be retrieved. Before starting the first level compression, the 'A-M' compression, an optional flagging process should take place, the flagging process is introduced in version 2, and however, it can be implemented in version 1. The objective of the flagging process at this stage is to identify in the frequent words that already stored in the list of frequent words in the dictionary, those words to be treated as whole words rather than character by character. Such words will be given a mapping code from the dictionary (codes between 62.x and 63.x). Applying the flagging process before starting the compression phase will help the algorithm to avoid unneeded parsing effort during compression. The flagging process is done as follows:

a) Search the original document for all words existing in the dictionary as frequent words class (situated between position 62.y and 63.y)

b) For each match, flag the word by adding the flag string. The flag string in version 2 is //. The flag is added before the word and after the word.

c) During the compression process, the added flags will be omitted and the whole word will be replaced by only 12 bit word.

d) If the flag itself is a part of the document, it will be processed also as one of the frequent words and will be surrounded by two flags.

e) To avoid infinite loops, the document is searched only one time without recursion. That is to say, word by word on the document will be compared with the frequent word list. The cursor will move forward either after flagging the word or leaving the word to be dealt with in character by character bases.

The next step after the flagging process will be the coding process. The coding process go through the text document and replace each character, characters, words by

the corresponding code that predefined in the dictionary. The sequence is very simple and will work as follows:

a) Read a character, If EOF, terminate, otherwise, read the next character (there are four options), and then proceed with the next steps.

b) If the two characters are numbers, code is generated from the dictionary as numerical class.

c) If the two characters are alphabetical characters (including the special characters such as space, full stop, etc …), in this case the output will be mapped from the alphabetical class in the dictionary.

d) If the two characters are // and that means to continue reading until finding another // and map the word in between to the frequent word class record (12 bits) and omit the // before and after the mapped frequent word

e) None of the above conditions, which means one of the three conditions:

   i) The first condition is that both characters are neither from the 60 basic characters nor a numerical character. In this case, each character will be handled individually and to be searched in the dictionary under the normal ASCII characters.

   ii) The second possibility is that the first character is from one of both lists but the second is not. In such a case the treatment will be the same as the previous case and both characters will be mapped to the normal ASCII characters.

   iii) The third condition is that the first character is not from the basic or the numerical groups, however, the second character is either from the basic or numerical list; or is a flagging character. In this case the first character will be dealt with as a normal ASCII character. The second character will be moved to be the first character and another character will be read and the process will repeat (go to a)

Arriving to this point means that the first level compression is completed and a dictionary based code is generated. The 'A-M' compression mentioned here based on 'A-M' dictionary version 2 may obtain between 25% and 75% based on the types of words in the document. The output compressed code from the fist level mentioned here can be compressed again via other variable length coding such as Huffman and Shannon-Fano.

The decompression process is much easier than the compression one. It is a direct reverse mapping process from the 12 bit codes in the compressed document into the original character, two characters, or word. In all cases, it will be a direct mapping between 12 bit record and its registered correspondence in the predefined dictionary.

If a second level optimization is used, the 12 bit records output of the first level will be used as the input to the second level coding system such as Huffman or Shannon-Fano. The output of the second level will be a coded file with a mapping table that maps the input symbols to a corresponding code.

D. Second level compression

After completing the first level 'A-M' compression, any variable length code compression can be used. In this paper we are using the example of Huffman as second level compression. Huffman code uses a very simple and effective algorithm. The algorithm is as follows:

a) Initialization: sort all existing symbols in a list according to their frequency of existence, this implies to read the entire file first and then start the sorting process.

b) Repeat the following three steps until the list has only one symbol left.

   i) Choose the lowest frequent two symbols (the symbols that repeated less than all the others). Create the Huffman sub-tree that has these two symbols as child nodes and create a parent node for them.

   ii) Assign the sum of the children's frequency counts to their parent and insert it into the created list in step 1, such that the order is maintained.

   iii) Delete the children from the original list.

c) Assign a codeword for each leaf based on the path from the root.

In the 'A-M' framework, the second level compression is strait forward. Each symbol in the list will be 12 bit symbol. Each symbol will correspond to the Huffman code generated using the above mentioned algorithm. Each record on the mapping table will be of the form <Huffman code><A-M code>

The generated mapping table will be used in the case of decompression. During the reverse process, 'A-M' code will be created from Huffman code. Afterwards, the original text will be created from the 'A-M' code using the pre-defined dictionary. The next section shows a new mechanism to reduce the overhead of the mapping table resulted from the second level compression.

III. MAPPING TABLE OPTIMIZATION

Another contribution of this paper is to introduce an optimization mechanism of the mapping table resulted from the second level compression.

The normal mapping table may contain a maximum of 3969 record. Each record will have a maximum of 24 bit depending on the Huffman code length. The table size may reach 11907 byte + separators size (spaces or new line). Our objective is to decrease this overhead of the second level compression. The suggested algorithm will reduce the overhead table to 9184 byte + separators size. If we neglect the separators size, the compression ratio is 23%.

The optimization technique is based on using flagging groups. All the 'A-M' records with common most significant bits are gathered. Then, the common part from the gathered records is cut. The cut MSB is mentioned just once after the flag. The algorithm is as follows:

a) All records having the 'A-M' code represented by Common-part.x will be grouped together and the

table will be of the form: <Huffman code><63.x>, <Huffman code><62.x>, <Huffman code><61.x>, etc…

b) Each group will be flagged with the unique flag 10000000000000000000000001, 24 zeros headed and followed by 1 that makes it 26 bits.

c) Remove the 6 most significant bits (common bits) for the record of each group and put it only once after the flag.

d) At the end of the table use the flag followed by 000000 as indication of end of the table.

During the reverse operation, the decoder reads the flag and recognizes the common part. The decoder adds the common part to each record of the group while generating the 'A-M' intermediate code. Having the right 'A-M' code, the rest of the decompression operation will be done as explained before.

IV. PUTTING ALL TOGETHER

The whole framework in case of using dictionary version2 will be as follows:

a) The dictionary is generated and saved either on the cloud or locally in the devices.

b) When a document require a compression, the dictionary version will be selected and the correspondent algorithm will be applied

c) In case of dictionary version 2, a flagging process will take place before the compression. During such process, the frequently used words will be flagged by // before and after each word.

d) The first level compression will take place in a straightforward mapping between character(s) or words to 12 bit code word. By producing 12 bit code words for each corresponding element, the 'A-M' normal compression is done and might be used as standalone compression.

e) A second level compression is used using variable length coding such as Shannon-Fano or Huffman coding. Any selected compression will be performed transparently without any modification except that each symbol is 12 bit long (the output of the first level compression)

f) After building the mapping table of the second level compression, another optional enhancement can be done by flagging common codes and apply the mapping table optimization algorithm. The compression process terminates here.

During the decompression phase, the reverse process is done as follows:

a) First regenerate the original mapping table. To do so, search the flag and identify the common MSB that omitted from the 'A-M' Code during the last step of the compression, those bits are the six bits appearing after the flag. Those common bits should be restored as MSB for each 'A-M' code.

b) Repeat the process for each flagged common bits until arriving to the end of the table. The end of the table is recognized by finding 000000 after the flag.

c) Use the mapping table to replace Huffman code (or any second level variable length code) by 'A-M' code in the document. The generated document will be 12bits records document.

d) Use the dictionary version 2 to replace each 12bits by their corresponding character, characters, or words. This will generate the original document.

V. CONCLUSION

This paper demonstrates three contributions to the 'A-M' Framework. The first one is the introduction of Dictionary version 2 that facilitates handling of records and unifies the record size of 'A-M' framework. Such a contribution is simplifying the processing of the dictionary and adding some useful features to it. The second contribution is the use of additional class for frequently used words. Such a contribution increases the compression rate to be between 50% and 90% for the records of this class. The third contribution is the optimization of the mapping table used in the second level optimization. The proposed algorithm reduces up to 23% of the size of the mapping table.

REFERENCES

[1] Ahmed Mokhtar A. Mansour and Mona A. M. Fouad, "A New Compression Framework: 'A-M' Code," Internathional Conference on Modeling, Simulation, and Applied Optimization, ICMSAO'13, Tunisia, April 2013.

[2] Li, Ze-Nian, Drew, Mark S., "Fundamental of multimedia," 2004 Pearson Education, ISBN 0-13-127256-X

[3] Rissanen, J.J.; Langdon, G.G., Jr. (March 1979), "Arithmetic coding," IBM Journal of Research and Development 23 pp. 149–162.

[4] Witten, Ian H.; Neal, Radford M.; Cleary, John G, "Arithmetic coding for data compression," Communications of the ACM 30 (6): pp. 520–540.

[5] Terry Welch, "A technique for high-performance data compression," IEEE Computer, June 1984, pp. 8–19

[6] Jacob Ziv and Abraham Lempel, "Compression of individual sequences via variable-rate coding," IEEE Transactions on Information Theory, September 1978.

[7] Shannon,C.E. (July1948), "A mathematical theory of communication," Bell System Technical Journal 27: pp. 379–423.

[8] Mordecai J. Golin, Claire Kenyon, Neal E. Young, "Huffman coding with unequal letter costs," STOC 2002: 785-791

[9] Nathanael Jacob, Priyanka Somvanshi, and Rupali Tornekar, "Comparative analysis of lossless text compression Techniques"International Journal of Computer Applications pp. 0975 – 8887 Volume 56– No.3, October 2012

[10] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, "Introduction to algorithms" MIT Press and McGraw-Hill, 2001. ISBN 0-262-03293-7. Section 16.3, pp. 385–392

[11] David Salomon, "Data compression," The complete reference, 4th edition 2012.

Adoption of Agile Methodology in Software Development

D. Duka
Ericsson Nikola Tesla, Split, Croatia
e-mail: denis.duka@ericsson.com

Abstract - As adopting Agile software development becomes a trend, there is a need for a more structured definition of what is Agile and what is a high-level of Agile maturity. Traditional development methodologies rely on documents to record and pass on knowledge from one specialist to the next. Feedback cycles are, in many cases, too long or even nonexistent. Agile principles emphasize building working software that people can get hands on quickly, versus spending a lot of time writing specifications up front. Agile development focuses on cross-functional teams empowered to make decisions, versus big hierarchies and splitting by function. It also focuses on rapid iteration, with continuous customer input along the way. This paper deals with Agile methodology and scaling. The special highlight is put on people investigating their contribution in Agile approach success. Some reflections after using Agile in our own organization are also presented.

I. INTRODUCTION

Traditionally the Telecom business has been standardization-driven and regulated on a national level. The development lead times have been long. Consequently the Telecom vendors have developed capabilities to influence standards, develop products and interact with the Telecom operators in a slow-moving industry. The business landscape has changed leaving the major slow-moving vendors to struggle with the pace of the newcomers [1].

The slow-moving market forces us to develop our ability to run major multi-year projects. We became predictable development machinery with extensive mechanisms to ensure predictability and control on the expense of flexibility and customer closeness.

This, in turn, led to organizational setups focusing on the alignment with the project structures and deepening the competencies in narrow areas both in the product and the functional dimensions. The result was organizational silos with multiple related hand-over challenges.

This gradually led to possessing less and less people with a broad systems understanding, which in turn slowed down both the organizational capability to handle broad technical challenges and the individual opportunities to learn and broaden the contribution. The feedback loops within the product development became longer and longer [2].

Only recently the Lean principles and the Agile development philosophy has been recognized as the source for solutions to these challenges.

Paper starts with describing Lean, Agile and Scrum methodology highlighting the difference between traditional *waterfall* and new approach in software development. The third chapter describes some main Agile characteristics. Scaling the Agile architecture and four proposed approaches are also presented here. The final chapter deals with real experiences following the Agile implementation in Ericsson underlining some observed changes and used practices while applying this new way of working. Some concrete measurements following the new concept are also presented.

II. ABOUT LEAN, AGILE AND SCRUM

The classical way to organize software development of large systems is the *waterfall* model, where the development starts from defining and analyzing the requirements and ends to operating (or maintaining) the software. Actual coding is only a minor part of the entire process, whereas there is much emphasis on defining, designing, documentation, testing and operating (maintaining) the software system [3].

Figure 1 [4] illustrates the waterfall software development model.

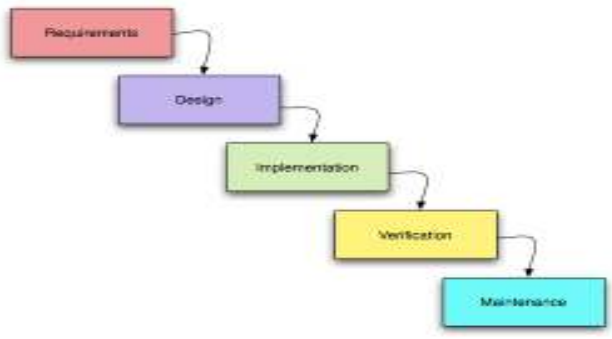


Figure 1. Waterfall Software Development Model

Agile software development model based on short iterations and quick releases challenges the *waterfall* models having emphasis on design and documentation.

The Agile manifest [5] recaps the ideology:

- Individuals and interactions over processes and tools,
- Working software over comprehensive documentation,
- Customer collaboration over contract negotiation,
- Responding to change over following a plan.

Lean manufacturing, initially called Just-In-Time production (JIT) was originally developed by Toyota in the latter part of the 20th century. Lean manufacturing system is built on several principles and philosophies. These include minimization of waste (through pull-production), Kaizen (continuous improvement), getting quality right from the beginning (stop production for fixing), among others. The Lean principles have been applied to Agile software development, and quite often are referred as Agile/Lean. Poppendieck Mary and Tom have published several books on the principles [6].

Agile and Lean are relatively broad concepts. There are several more detailed software development models, the developers of which call them as Agile methods. Some can also be considered Lean. These models include Scrum, Extreme Programming (XP) and Agile Unified Process (AUP), among others.

Stober [7] describes Scrum as a drastic simplification of project management containing three roles, three documents and three meetings. As can be seen from Figure 2, in a Scrum software development project, Product Owner (PO) decides the product backlog, i.e. the features expected from the software (for the next release), signs off all deliverables and represents budget and interests of the stakeholders. At the start of each sprint, the project team decides in a sprint planning meeting, which items from the product backlog are taken to the sprint backlog as use cases for the software. This is based on the prioritization by the Product Owner and teams work estimates and commitments. During each sprint, which duration is typically two weeks, the team completes the sprint backlog. A daily short Scrum meeting is held to follow-up the ongoing work and solve rising issues. Scrum Master (ScM) facilitates issue solving outside the meetings. At the end of each sprint a sprint review meeting is held to review the sprint results. Each sprint results in a working increment of the software product. Sprints are repeated until the product backlog is depleted and the software release is ready. Releases again are repeated for major software updates.

In this paper I refer principally to Scrum model, when discussing Agile, as it constitutes a clear and tangible reference model.

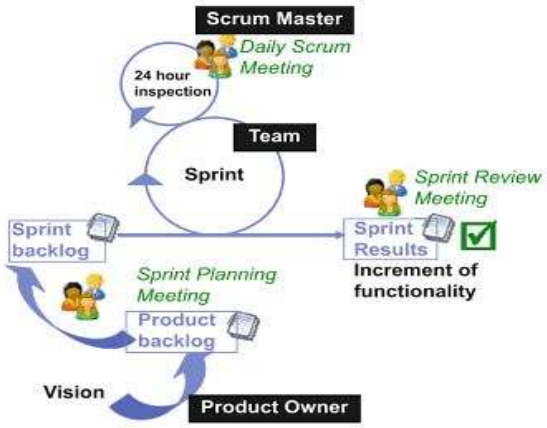


Figure 2. Scrum overview

III. AN AGILE APPROACH

The Agile approach aims to nurture organization assets and to support other groups, such as development teams, within organization. These groups should act in an Agile manner that reflects the expectations of their customers and the ways in which their customers work [8].

First and foremost, the Agile values, principles and practices should help to guide organization architecture modeling and documentation efforts. This is just a good start though. In order to be successful at organization architecture you need to rethink your overall approach and address some fundamental issues. These issues are connected in a synergistic manner; you must address all of them otherwise you will put your effort at risk. Some of these issues are:

- Focus on people, not technology or techniques:
  All of the arguments over "which model is right", "which notation is right", and "which paradigm is right" are meaningless if you don't have a viable strategy for working together effectively. You could create a perfect organization architecture model but it doesn't matter if project teams can't or won't take advantage of it.
- Keep it simple:
  A critical concept is that organization architecture models and documents just need to be good enough, they don't need to be perfect. By keeping organization architecture artifacts simple you increase the chances that your audience will understand them, that project teams will actually read them, and that you will be able to keep them up to date over time. Overly detailed documents might look impressive sitting on a shelf, but a simple model that project teams actually use is what your true goal should be.
- Work iteratively and incrementally:

Agile organization architects work in an iterative and incremental manner. They don't try to create complete models. Instead they create models that are just good enough. When they find that their models are not sufficient they work on them some more. The advantage of this approach is that they evolve their models incrementally over time, effectively taking a just-in-time (JIT) model storming approach that enables them to get the models in the hands of their customers as quickly as possible.

On large Agile teams, geographically distributed Agile teams, or for organization-wide architectural efforts, an architecture owner team will be required. Agile teams at scale are organized into smaller sub-teams. The architecture owner on each sub-team is a member of architecture owner team, which helps to increase the chance that each sub-team understands and follows the overall architecture as well as increases the chance that the overall architecture strategy will address the full needs of the overall solution. There will be an overall architecture owner, this could be a rotating role, who is responsible for facilitating the group [9].

There are four basic strategies for organizing Agile teams at scale:

- Architecture-driven approach:
  With this strategy you organize your sub-teams around the subsystems/components called out in your architecture. This strategy works well when your architecture is of high quality and the interfaces to the subsystems have been identified before the sub-teams really get going. Although the interfaces will evolve over time, they are needed to get a good start at them initially. The challenge with this strategy is that it requires your requirements to be captured in a way which reflects the architecture. For example, if your architecture is based on large-scale business domain components then a requirement should strive to focus on a single business domain if possible. If your architecture is based on technical tiers then requirements should focus on a single tier if possible.
- Feature-driven approach:
  With this strategy each sub-team implements a feature at a time, a feature being a meaningful chunk of functionality to your stakeholders. This strategy should be applied in situation where the architecture exhibits a lot of coupling and where you have sophisticated development practices in place. The challenge with this approach is that the sub-teams often need to access a wide range of the source code to implement the feature and thereby run the risk of collisions with other sub-teams. As a result, these teams require sophisticated change management, continuous integration, and potentially even parallel independent testing strategies in place.
- Open source approach:
  With this strategy one or more subsystems/components are developed in an open source manner, even if it is for a single organization. This strategy is typically used for subsystems/components which are extensively reused by many teams, for example a security framework, and which must evolve quickly to meet the changing needs of the other systems accessing/using them. This strategy requires you to adopt tools and processes which support open source approaches.
- Combinations thereof:
  Most Agile teams at scale will combine the previous three strategies as appropriate.

IV. IMPLEMENTING AGILE

Recently, our R&D management team started discussing how to adapt the way-of-working in our Unit to the increasingly turbulent business environment. Prediction of long-term market development, which had been the basis of our operations and requirement setting, was getting more and more difficult. Looking into the future, it became clear that we needed to do a proactive change in order to more flexibly react to customer wishes. After months of discussions, Agile and Lean principles were chosen as the guiding principles of our way forward.

In the beginning, we became aware of the concept of Agile in large scale telecom industry. It was challenging to realize in advance how we could change ourselves from a component-based technically-thinking organization into a customer-focused cross-functional and value-thinking organization [10].

Even before the beginning of the change journey we could identify multiple obstacles in front of us. However, the expected advantages were greater. Soon we realized that if we want to stay as a serious player in the fierce telecom industry, we need to start the journey towards becoming an Agile software company.

A. Less detailed level planning

We used approximately half a year to internally spread the knowledge of Agile software development methods. The change was complex and thus we were not able to justify the change by utilizing the normal business case study procedure.

We wanted to have a detailed plan of the one year change project. We wanted to understand and solve all the major potential problems beforehand. We wanted to mold this plan into our yearly strategic planning process and have each organizational silo responsible for part of the change. The coaches changed our approach more to "start implementing with a small scale and tackle the problems as they occur".

We started with one cross-functional development team and one Product Owner that had a task to implement a real customer feature into our product. In the beginning, it was a constant struggle to grant room for the experiment and, at the same time, keep the customer promises for the ongoing project. It was clearly two extremely different worlds that met. One world that we knew how it worked, that was predictable and going on with a slow but steady tempo. Then the new world with a greater tempo in which all current weekly and monthly meeting routines felt as road blockers on the way. The previous organization was not ready to react fast enough with the new requirements that two-week Scrum sprints brought to surface. Early on the team realized that our development environment was outdated and not built for making fast and small end-to-end deliveries. The team was encouraged to experiment with new tools and find better means that would support the faster feedback in our R&D organization.

Here are the steps that we have taken during the journey:

- One team, one Product Owner and one Scrum Master,
- Three teams (one off-site), one Product Owner and three Scrum Masters,
- Customer Support Request (CSR) and Fault Handling (FH) Kanban teams with Product Owners and Team Coaches,
- Building the continuous integration machinery,
- More new teams with a Scrum Master and a Product Owner,
- Feature and System integration in close co-operation,
- All development teams are cross-functional, self-organized and stable,
- Feature feasibility study done in co-operation with the lead feature team,
- The Release verification is in close co-operation with the feature development.

B. Intense learning & competence build up

Changing people mindset starts from small and visible actions that are repeated everyday. Management created a guideline to make clear what we want more in our everyday working environment. We believe that fundamental changes that are needed in our minds to succeed with this journey are as follows:

- More people initiative and less top-down control,
- More team players and less individual heroes,
- More courage and less risk avoidance,
- More conversations and less one-way communication,
- More personal growth and less comfort zone.

The teams are expected to learn in two dimensions. One is the functional dimension, i.e. system-design-testing. The other is the product dimension, i.e. different components in the system. They need to learn the tiny end-to-end functionality that goes through multiple components. This new dimension of learning needs different attitude than before.

Combining system, design and test competences into one team gave the possibility to share knowledge between the team members. Visualizing the task both in the Kanban board and the sprint planning board was found to be a valuable tool for the team coach to help the team members to broaden their competences.

C. Innovation

Change from the silo-based R&D center towards the Agile development R&D center included intensive learning, facing new challenges and risk taking which are also typical parameters of innovations. The elements in Lean & Agile and our innovation environment are complementing each others.

We have seen that by bringing people around the same table from the different backgrounds, it generates a spin of positive, innovative energy. Focal point of the current Agile software development is efficient software development in Agile teams. Hence, the majority of the creativity is naturally directed towards the improvements around the ongoing work in the sprints.

D. Involving the whole organization

From the beginning, we chose to be as open as possible during the planning and transformation. We realized that if we want value workers to actively contribute we need to involve them in the decision making and show all information openly. We had all meeting notes and workshop material visible in the intranet.

We introduced the “current best thinking“ to emphasize that we plan only short period ahead and want everybody to participate and look for alternatives.

Once a week we had an open question session called *Fast Forward Friday*. The sessions were found extremely valuable. Everybody had a chance to ask questions and grow their understanding of what is expected in the future. The level of questions told to the management team what the maturity of the rest of the organization was.

We utilized also anonymous web polls to gather individuals’ opinions. The free text answers also gave valuable information regarding where to put effort while driving the change into the proper direction.

E. Making the change visible

Already some years back we had the understanding that working according the strict organizational silo-

responsibility and budget-areas will limit the whole organization to improve. Therefore, we intended to introduce a new organizational setup. First, we started to break down the organizational silos by introducing flow based *pipes* that consist of several organizational resource pools. For instance, *the execution pipe* had resources from system, development and verification department.

The other issue was seating arrangement. We had a long history of everybody having their own 8 m^2 offices where one could close the door while wanting to concentrate in privacy. Therefore, it was understandable that the resistance for the new seating arrangement was substantial. Our management and also the project management team had shared an open space area for a year. Their positive experiences of how much easier it is to communicate in an open space area was supporting the move.

F. Agile practices used

While introducing Agile we have followed the basic Scrum and tried out how it fits into our environment without judging something that we have not tried.

In product maintenance, where the nature of work is a constant flow of customer service requests, we have utilized Kanban combined with practices from the Scrum framework. For instance, since the retrospectives are seen essential while experimenting the new way of working, we added that into the Kanban method.

Behind the decision of using Scrum there was a profound discussion and consideration. As a large R&D center developing a complex product, we needed to have one common approach for the teams. At Ericsson we have a history of tailor-making own tools and methods. This time the fact was that we did not have enough experience or competence to tailor a suitable process. It would have required too much time to develop an Ericsson-specific Agile process.

G. Agile experiences

Following the new methodology an assessment was performed in our organization. Overall 8% improvement is seen across all nodes put together. Results are presented through the following graph:

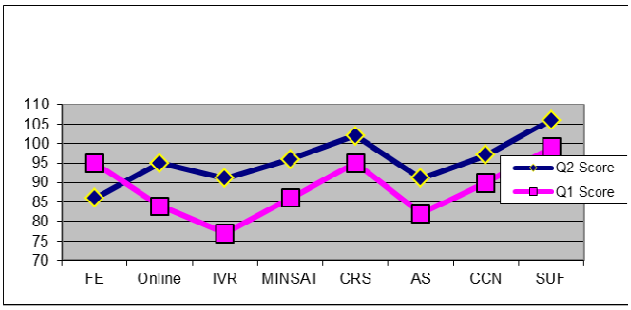


Figure 3. Agile Q1 & Q2 Assessment Scores Trend

V. CONCLUSION

Successfully implementing organizational wide change requires extra time and high energy level from the whole organization.

By using this approach in software design we are able to first recognize and then eliminate activities which don't add value e.g. partially done work without any guarantee if customer will take it in use, unnecessary task swapping, waiting, handoffs, faults, etc.

This way of working also implies improved responsiveness through rapid deliveries allowing customers to delay decisions. This is especially enabled by short feedback loops and continuous integration.

Our motivation behind the change was not only the current product under development but also to increase the level of our R&D center ability. First-hand experience of Scrum framework is the real source of learning. Only through experiencing it is possible to express real meaning of how well the new way of working fits into our situation.

Preliminary evidence we've collected from an assessment, but also from other sources all tell us that Agile software development is making a positive difference in our organization. In addition, interest and respect for Agile practices is constantly growing. Other parts of the company are interested in what we're doing and how we're making it work. Consequently, our future work will target other software teams who are adopting Agile practices and who need our guidance and support. Following such approach we'll be able to improve our pratices used and also to spread Agile culture within and outside our organization.

REFERENCES

[1] J. Astemark, " Agile@ITTE", Internal Ericsson Documentation, Sweden 2011

[2] D. Duka, "Agile approach in Software Development", Proceedings of the 35th Mipro CTI conference, 2012

[3] J. Sääskilahti, J. Röning "Challenges with Software Security on Agile Software development, Internal Ericsson Documentation, Sweden 2011

[4] R. Winston, "Managing The Development of Large Software Systems", Proceedings, IEEE WESCON, August 1970, pages 1-9. http://www.cs.umd.edu/class/spring2003/cmsc838p/Process/waterfall.pdf

[5] K. Beck, "Manifesto for Agile Software Development", 2001 http://Agilemanifesto.org

[6] M. and T. Poppendieck, "Lean Software Development Principles", http://www.poppendieck.com/

[7] T. Stober, U. Hansmann. "Overview of Agile Software Development", Agile Software Development: Best Practices for Large Software Development Projects, Springer. 2010.

[8] S. W. Ambler, "Agile Enterprise Architecture", http://www.Agiledata.org

[9] S. W. Ambler, "Agile Architecture: Strategies for Scaling Agile development", http://www. Agilemodeling.com.org

[10] K. Mikkonen, "How We Learned to Stop Worrying and Live with the Uncertainties", Internal Ericsson Documentation, Sweden 2011

Developments and current trends in Ethernet technology

D. Valenčić*, V. Lebinac* and A. Skendžić**
* University of Applied Sciences Velika Gorica, Velika Gorica, Croatia
** University of Applied Sciences Nikola Tesla, Gospić, Croatia
davorin.valencic@vvg.hr, vladimir.lebinac@vvg.hr, wireless82@gmail.com

Abstract - Ethernet technologies continue to evolve. This paper presents overview of the current major advancements and developments of Ethernet technology and its appropriate standards, such as Higher Speed Ethernet (40 Gbit/s and 100 Gbit/s Ethernet), Carrier Ethernet technologies (implemented by carriers and service providers), Energy Efficient Ethernet (Green Ethernet) technologies, Power over Ethernet Plus, and others. In this paper is also shown some implementations of different new Ethernet standards and features by major vendors of network equipment (Cisco, Juniper, Brocade, Huawei and others).

I. INTRODUCTION

Ethernet is family of networking technologies initially developed for wired Local Area Network (LAN) implementations. Ethernet is networking technology that covers first two layers of OSI reference model. Ethernet dominance is based on low cost, easy to configure and deploy equipment.

Since its introduction almost 40 years ago, Ethernet was several times claimed as an obsolete technology that should be replaced with (at that time) new emerging technologies, for example Token Ring, FDDI (Fiber Distributed Data Interface), Emulated LAN. Anyway, Ethernet has not only survived, but also it was significantly developed and several new implementations were introduced so that today Ethernet is the most widely spread wired networking technology. All the Ethernet improvements had to preserve 802.3 Ethernet frame format utilizing the 802.3 media access controller (MAC), based on CSMA/CD (Carrier Sense Multiple access / Collision Detection).

The latest revision of base standard IEEE 802.3-2012 was approved in September 2012. In this paper we will have a closer look to the following improvements of Ethernet that are included into the IEEE 802.3-2012 Ethernet standard:

- IEEE 802.3ba (40 Gbit/s and 100 Gbit/s Ethernet) approved on June 17, 2010.
- IEEE 802.3at (Power over Ethernet Plus) approved in September 30, 2009.
- IEEE 802.3az (Energy Efficient Ethernet) approved on September 30, 2010.

Also in this paper there will be briefly reviewed Carrier Ethernet as extension of Ethernet from LAN to both Metropolitan Area Networks (MAN) and Wide Area Networks (WAN) and as a new area of Ethernet expansion. Also there has been shown some examples of implementations of analyzed Ethernet improvements by major vendors of network equipment (Cisco, Juniper, Brocade, Huawei and others). In this paper are not analyzed numerous non-standard and proprietary Ethernet schemes and solutions developed by Ethernet equipment vendors.

II. IEEE 802.3 STANDARD AND ETHERNET EVOLUTION

IEEE 802.3 standard was first published in 1985 [1]. Since the initial publication, many projects have added functionality or provided maintenance updates to the specifications and text included in the standard. Each IEEE 802.3 project/amendment is identified with a suffix (e.g., suffix "ba" released in year 2010 in amendment of standard IEEE 802.3ba-2010).

The Media Access Control (MAC) protocol specified in IEEE 802.3 is Carrier Sense Multiple Access with Collision Detection (CSMA/CD). This MAC protocol was included in the experimental Ethernet developed at Xerox Palo Alto Research Center. While the experimental Ethernet had a 2.94 Mbit/s data rate, IEEE 802.3-1985 specified operation at 10 Mbit/s.

Since 1985 new media options, new operation speeds, and new capabilities have been added to IEEE 802.3 standard. For example:

- IEEE 802.3u added 100 Mbit/s operation (also called Fast Ethernet),
- IEEE 802.3x specified full duplex operation and a flow control protocol,
- IEEE 802.3z added 1000 Mbit/s operation (also called Gigabit Ethernet),
- IEEE Std 802.3ae added 10 Gbit/s operation (also called 10 Gigabit Ethernet),
- IEEE 802.3ah specified access network Ethernet (also called Ethernet in the First Mile).

The latest IEEE 802.3 – 2012 revision approved by the IEEE Standards Association (IEEE-SA) incorporates various technical updates and enhancements and consolidates a host of amendments to the base standard that were approved since IEEE 802.3's last full revision, in 2008.

Since publishing of the standard IEEE 802.3-2008, several new standards (as amendments to IEEE 802.3-2008 standard) have been issued and maintained as separate documents. Amendments addressing, energy efficiency, extension to 40 Gbit/s and 100 Gbit/s speeds, 10 Gbit/s Ethernet Passive Optical Networks (EPONs) while maintaining compatibility with previously installed IEEE 802.3 interfaces, enhanced support for loss-sensitive applications and time synchronization are among those that have been incorporated into IEEE 802.3-2012.

III. 40 GIGABIT ETHERNET AND 100 GIGABIT ETHERNET

The growth in bandwidth for network aggregation applications was found to be outpacing the capabilities of networks employing link aggregation with 10 Gigabit Ethernet. In 2006, the IEEE 802.3 working group formed the Higher Speed Study Group (HSSG) and found that the Ethernet ecosystem needed something faster than 10 Gigabit Ethernet. HSSG determined that two new rates were needed: 40 gigabit per second for server and computing applications and 100 gigabit per second for network aggregation applications.

The IEEE 802.3ba 40 Gbit/s and 100 Gbit/s Ethernet amendment to the IEEE 802.3-2008 Ethernet standard was approved on June 17, 2010 by the IEEE Standards Association Standards Board.

The IEEE 802.3ba standard offers a single architecture capable of supporting both 40 and 100 Gigabit Ethernet, while producing physical layer specifications for communication across backplanes, copper cabling, multi-mode fiber, and single-mode fiber. This chapter provides a brief overview of 40 and 100 Gigabit Ethernet underlying technologies.

A. IEEE 802.3ba - 40 and 100 Gigabit Ethernet

The objectives that drove the development of IEEE 802.3ba standard are the following [2]:

- Support the full duplex Ethernet MAC.
- Preserve the IEEE 802.3 Ethernet frame format utilizing the IEEE 802.3 MAC.
- Preserve minimum and maximum frame size of IEEE 802.3 standard.
- Support a BER better than or equal to 10^{-12} at the MAC/physical layer service (PLS) service interface.
- Provide appropriate support for Optical Transport Network (OTN).
- Support a MAC data rate of 40 Gbit/s.

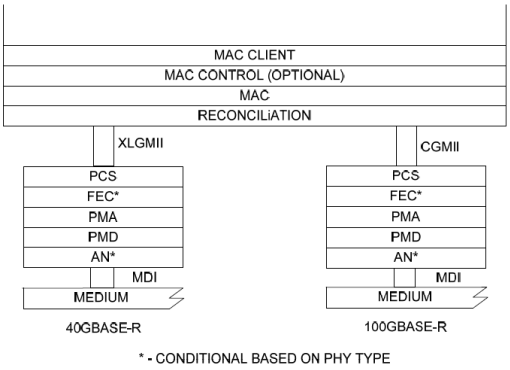


Figure 1. IEEE 802.3ba Architecture [3]

- Provide Physical Layer specifications that support 40 Gbit/s operation over up to the following:
  - At least 10 km on single-mode fiber (SMF)
  - At least 100 m on OM3 multimode fiber (MMF)
  - At least 7 m over a copper cable assembly
  - At least 1 m over a backplane
- Support a MAC data rate of 100 Gbit/s.
- Provide Physical Layer specifications that support 100 Gbit/s operation over up to the following:
  - At least 40 km on single-mode fiber (SMF)
  - At least 10 km on single-mode fiber (SMF)
  - At least 100 m on OM3 multimode fiber (MMF)
  - At least 7 m over a copper cable assembly

The IEEE 802.3ba-2010 amendment specifies a single architecture shown in Figure 1. It accommodates 40 Gigabit Ethernet and 100 Gigabit Ethernet and all of the physical layer specifications under development. The MAC layer, which corresponds to Layer 2 of the OSI model, is connected to the media (optical or copper) by an Ethernet physical layer device (PHY), which corresponds to Layer 1 of the OSI model. The PHY device consists of [3]:

- physical medium dependent (PMD) sublayer,
- physical medium attachment (PMA) sublayer,
- physical coding sublayer (PCS).

The backplane and copper cabling PHYs also include an auto-negotiation (AN) sublayer and a forward error correction (FEC) sublayer.

The PCS sublayer translates between the respective media independent interface (MII) for each rate and the PMA sublayer. The PCS is responsible for the encoding of data bits into code groups for transmission via the PMA

TABLE I. IEEE 802.3BA PHYSICAL LAYER SPECIFICATIONS [3]

	40 GE	*100 GE*
At least 1m backplane	40GBASE-KR4	
At least 7m copper cable	40GBASE-CR4	100GBASE-CR10
At least 100m OM3 MMF	40GBASE-SR4	100GBASE-SR10
At least 150m OM4 MMF	40GBASE-SR4	100GBASE-SR10
At least 10km SMF	40GBASE-LR4	100GBASE-LR10
At least 40km SMF		100GBASE-ER10

and the subsequent decoding of these code groups from the PMA. The PCS leverages the 64B/66B coding scheme that was used in 10 Gigabit Ethernet. It provides a number of useful properties including low overhead and sufficient code space to support necessary code words, consistent with 10 Gigabit Ethernet. The transmit PCS performs the initial 64B/66B encoding and scrambling on the aggregate channel (40 or 100 gigabits per second) before distributing 66-bit block in a round robin basis across the multiple lanes.

The PMA sublayer enables the interconnection between the PCS and any type of PMD sublayer. A PMA sublayer will also reside on either side of a retimed interface, referred to as "XLAUI" (40 Gbit/s attachment unit interface) for 40 Gigabit Ethernet or "CAUI" (100 Gbit/s attachment unit interface) for 100 Gigabit Ethernet. XLAUI and CAUI are low pin count physical interfaces that enables partitioning between the MAC and sublayers associated with the PHY in a similar way to XAUI in 10 Gigabit Ethernet. They are self-clocked, multi-lane, serial links utilizing 64B/66B encoding. Each lane operates at an effective data rate of 10 gigabit per second, which when 64B/66B encoded, results in a signaling rate of 10.3125 gigabaud per second. In the case of XLAUI, there are four transmit and four receive lanes of 10 Gbit/s, resulting in a total of 8 pairs, or 16 signals. In the case of CAUI, there are 10 transmit lanes and 10 receive lanes of 10 Gbit/s, resulting in a total of 20 pairs or 40 signals [3].

Regarding the PMD sublayer, different physical layer specifications for computing and network aggregation applications have been developed. For computing applications, physical layer solutions will cover distances inside the data center for up to 100m for a full range of server form factors, including blade, rack, and pedestal configurations. For network aggregation applications, the physical layer solutions include distances and media appropriate for data center networking, as well as service provider inter-connection for intra-office and inter-office applications. A summary of the physical layer specifications being developed for each MAC rate is shown in Table I.

TABLE II. IEEE 802.3BA VENDOR IMPLEMENTATIONS

Vendor	*Equipment*
Brocade	12-port 40GE linecard for VDX 8770 switch; 2-port 100GE module for MLX series switching routers
Cisco	2-port 100GE board and 6-port 40GE module for Nexus 7000 switch; Nexus 3064-X with 48 1/10GE ports and four 40G links; 40GE module for Catalyst 6500 switch
Extreme	4-port 100G and 12-port 40G modules for BlackDiamond X8 core switch
Huawei	40GE and 100GE linecards for CloudEngine 12800 series of core switches
Juniper	100GE interface card for T1600 Core Router

B. *Implementation*

In the Table II are shown some examples of IEEE 802.3ba standard implementation by networking equipment vendors, mainly interface cards for core and data centre switches and routers [4-8].

IV. POWER OVER ETHERNET PLUS

A. *IEEE 802.3af - Power over Ethernet (PoE)*

The IEEE 802.3af is the standard for Power over Ethernet (PoE) and it was released in 2003. It helped increase the value of an Ethernet port by giving the opportunity of connecting and powering devices such as IP Phones using a common network infrastructure (mainly Ethernet switches).

In IEEE 802.3af standard powered devices (PD) refer to end devices that require power such as IP phones or WiFi Access Points, while power sourcing equipment (PSE) are devices that deliver power such as PoE switches.

The majority of PD number growth comes from IP phones usage by enterprise customers who are deploying converged voice and data networks that operate across a common IP network infrastructure, resulting in cost saving.

The original IEEE 802.3af standard provides up to 15.4 W of DC power (minimum 44 V DC and 350 mA) to each device. Only 12.95 W is assured to be available at the powered device as some power is dissipated in the cable [9].

B. *IEEE 802.3at - Power over Ethernet (PoE) Plus*

Increasing number of powered devices required power greater than the allowable 12.95 W specified in IEEE 802.3af. These powered devices are, for example, IEEE 802.11n WiFi Access Points, pan-tilt-zoom (PTZ) security cameras, and IP Phones with advanced features such as video conferencing. These powered devices benefit from IEEE P802.3at standard.

The IEEE 802.3at (released in September of 2009) is a gradual evolutionary improvement to the IEEE 802.3af standard. The main characteristics of the IEEE 802.3at are the following [10]:

- The IEEE 802.3at standard provides up to 34.2 W of DC power to each device (25.5 W is assured to be available at the powered device).
- IEEE 802.3at use Category 5 (Cat 5) cabling only (IEEE 802.3af supports Category 3 (Cat 3) and Cat 5). PoE Plus requires Cat 5 (8-wire) instead of

TABLE III. STANDARD POE PARAMETERS AND COMPARISON

Value	*802.3af (802.3at Type 1)*	*802.3at Type 2*
Max power delivered by PSE	15.40 W	34.20 W
Power available at PD	12.95 W	25.50 W
Voltage range (at PSE)	44.0–57.0 V	50.0–57.0 V
Voltage range (at PD)	37.0–57.0 V	42.5–57.0 V
Maximum current	350 mA	600 mA
Supported cabling	Cat 3, Cat5	Cat5

Cat 3 (4-wire) because more power can be transmitted over two 4-wire cables.

- IEEE 802.3at abides to the power safety rules and limitations pertinent to IEEE 802.3af.
- IEEE 802.3at power sourcing equipment is compatible with IEEE P802.3af.
- IEEE P802.3at should provide the maximum power to PD's as allowed within practical limits.
- IEEE 802.3at defines a powered device MIB for SNMP management.

IEEE 802.3at provides full backward compatibility and interoperability to existing 802.3af compliant PSE and PD equipment. IEEE 802.3at defines the PD classification to ensure the backwards compatibility [9]:

- Type 1 PD's are IEEE 802.3af powered devices and have a maximum wattage requirement of 12.95W.
- Type 2 PD's which require power from 12.95W and up to 25.5W.

Type 2 PD's that cannot operate with less than 12.95W must give indication to user when connected to Type 1 PSE, while Type 2 PD's that can operate with less than 12.95W must be able to be powered by a Type 1 PSE [9].

Basic parameters and a brief comparison of the IEEE 802.3af and the IEEE 802.3at standards are given in Table III.

C. Implementation

The migration to the IEEE802.3at is slow but steady as more vendors determine how to take advantage of higher power delivered over Ethernet and design and test new PDs that can take advantage of the higher power. The IEEE P802.3at standard helps to combine services such as video used for perimeter security in a single infrastructure. For example, a closed-circuit television (CCTV) camera used for perimeter security requires multi-wiring and analog-based receiver/recorder such as a VCR (Video Cassette Recording). While a pan-tilt-zoom security camera that supports IEEE P802.3at only requires one wire, an RJ-45, to capture the scene, transmit the video, and power the camera. This gives enterprise customers cost savings, a simplified infrastructure, and greater security coverage [10].

Some examples of equipment support for the IEEE 802.3az standard are Brocade ICX 6430 and 6450 Switches [4], Cisco UPOE Line Card for Catalyst 4500E Series switching platform [11], Extreme Networks Summit X460 Series Ethernet Switches [6], Huawei S2700 Series Enterprise Switches [7] and Juniper EX Series Ethernet Switches [8].

V. ENERGY EFFICIENT ETHERNET

There has been a growing focus on the energy usage of IT devices. The programs to reduce IT energy consumption have initially concentrated on the areas of highest energy usage: computers and consumer devices. However, networking equipment has been identified as consuming as much as 10% of all IT energy, so it is logical to consider how networking energy consumption can be reduced without adversely affecting the critical functionality that networking performs [12].

The majority of Ethernet links spend much of the time idle, waiting between packets of data, but consuming power at a near constant level. Energy Efficient Ethernet (EEE) provides a mechanism and a standard for reducing this energy usage without reducing the vital function that these network interfaces perform. Some companies introduced technology to reduce the power required for Ethernet before the standard was ratified, using the name Green Ethernet.

A. IEEE P802.3az – Energy Efficient Ethernet

The IEEE 802.3az standard was approved on September 30, 2010. It is the first standard in the history of Ethernet to address proactive reduction in energy consumption for networked devices and it is designed to provide network managers and networking services consumers with the tools to reduce energy consumption in network-attached devices, network routers and switches, computers, and printers.

The IEEE 802.3az standard deals with the mainstream "BASE-T" interfaces (i.e. 10BASE-T; 100BASE-TX; 1000BASE-T; and 10GBASE-T) that operate over twisted pair wiring [12]. These interface types comprise the vast majority of Ethernet deployments, especially at the edge of networks where the opportunities for energy savings are maximal. The standard also covers Backplane Ethernet interfaces used in blade servers (as well as within proprietary systems) because the amount of change required for those interfaces was considered minor.

The fundamental idea of EEE is that the communication link should need to consume power only when real data is being sent. Most wireline communications protocols developed since the 1990s have used continuous transmission - consuming power whether or not data was being sent. The reasoning behind this was

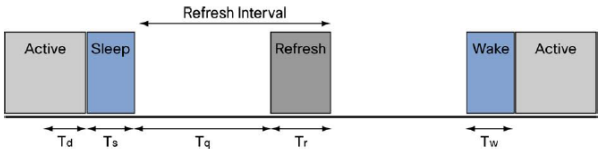


Figure 2. IEEE 802.3ba Architecture [13]

that the link must be maintained with full bandwidth signaling so that it is ready to support data transmission at all times. In order to save energy during times where there is a gap in the data stream, EEE uses a signaling protocol that allows a transmitter to indicate that there is a gap in the data and that the link can go idle. The signaling protocol is also used to indicate that the link needs to resume after a pre-defined delay.

The EEE protocol uses a signal that is a modification of the normal idle that is transmitted between data packets. This signal is termed low power idle (LPI) [13]. Low Power Idle (LPI) mode is an optional mode intended to save power that may be enabled during periods of low link utilization in which either side of a link may disable portions of device or system functionality. The transmitter sends LPI in place of idle to indicate that the link can go to sleep. After sending LPI for a period (Ts = time to sleep), the transmitter can stop signaling altogether so that the link becomes quiescent. Periodically, the transmitter sends some signals so that the link does not remain quiescent for too long without a refresh. Finally, when the transmitter wishes to resume the fully functional link, it sends normal idle signals. After a pre-determined time (Tw = time to wake) the link is active and data can be sent [13].

New Link Layer Discovery Protocol (LLDP) TLVs (type, length, and values) are defined for negotiating system level energy-efficiency parameters [14]. LLDP power negotiation allows the PoE controller to dynamically allocate power to LLDP-enabled powered devices based on their power needs. The PoE controller allocates to an interface only the power currently required by the connected powered device, and it can allocate the power in small increments. (LLDP was defined in the IEEE 802.1ab standard initially released in 2005 and revised in 2009.)

B. Implementation

During the IEEE 802.3az standard development process, a lot of attention was given to backwards compatibility. It is deployable in networks where the majority of equipment uses legacy interfaces and must also seamlessly support the very wide range of applications that already run on these networks. Ethernet interfaces complying with the IEEE 802.3az standard might not save energy when connecting with older devices as long as the existing functions are fully supported. This allows incremental upgrades for networks to increasingly benefit from EEE as the proportion of EEE equipment increases.

The standard also recognizes that some network applications may allow larger amounts of traffic disturbance and includes a negotiation mechanism to take advantage of such environments to increase the depth of energy savings.

EEE represents the beginning of a change of opinion in networking architecture. Previously it had been acceptable that networking devices, like the communications on the links themselves, continue to use energy at the same rate, regardless of the level of usage. The standard for EEE defines the signaling necessary for energy savings during periods where no data is sent on the interface, but does not define how the energy is saved, nor mandate a level of savings. This approach allows for a staged rollout of systems with minimal changes that are compatible with future developments that extend the energy savings.

It should be expected that early implementations of the standard save relatively small amounts of energy (comparing idle energy to full rate usage). However, these systems will be compatible with later products that may save much greater proportions of their energy use. The early systems may use a simple application of static logic design in the physical layer devices (PHYs) to save energy when data is not present. PHYs typically consume between 20 to 40 percent of the system power, and the static design methods allow savings of up to 50 percent of the PHY power. Therefore the expected system-level savings may be in the range of five to 20 percent [13].

Later generations of networking systems will use more aggressive energy savings techniques, such as power islands or voltage scaling. These methods can be applied to all of the system silicon, extending the range of energy savings. However, such aggressive techniques require significant new architecture design and will necessarily follow much later. With these power savings, an individual networking system may be able to save as much as 80 percent of its worst case energy use in certain situations [13].

Some examples of the support for the IEEE 802.3az standard by equipment of networking vendors are Brocade ICX 6430 and 6450 Switches (hardware ready) [4], Cisco UPOE Line Card for Catalyst 4500E Series switching platform [11][13], Huawei S9300 series switches [7] and Juniper EX Series Ethernet Switches [8].

VI. CARRIER ETHERNET

Carrier Ethernet is a marketing term for extensions to Ethernet to MAN and WAN networks in order to enable telecommunications network providers to provide Ethernet services to customers and to utilize Ethernet technology in ISP networks.

Many technologies have been used to deliver metro and wide-area services. Layer 1 TDM technologies used to deliver private line services include E1 and E3 copper circuits, and SDH/SONET-based optical circuits. Layer 2 technologies used to deliver MAN/WAN services include Frame Relay, ATM and PPP. These legacy technologies provide inflexible bandwidth scalability because the bandwidth is dictated by the technology. When a service provider or enterprise needs to add bandwidth, they either bond multiple circuits together or upgrade their network and equipment to support a new technology.

Carrier Ethernet addresses the limitations of legacy WAN technology by providing flexible bandwidth scalability. Once an Ethernet service is deployed, bandwidth can be added simply through remote provisioning up to the Ethernet port speed. With Carrier Ethernet, subscribers can use the same, well understood Ethernet technology for their LAN, MAN and WAN connections. This reduces costs for the equipment to

connect to the service and also simplifies operations and training.

A. *Ethernet as MAN/WAN technology*

Ethernet in his long history has become dominant in enterprise networks and LAN environment but Ethernet has traditionally had a number of limitations in the MAN/WAN application. Ethernet does not scale very well to MAN/WAN networks as it uses layer 2 MAC addressing scheme, transparent bridging, spanning tree protocol and it extends broadcast domain. Moreover, Ethernet has lacked some of the dependability features necessary in service provider application, for example mechanisms to isolate one customer's traffic from another, to measure performance of a customer service instance, and to rapidly detect and repair failures in larger networks.

The industry has made a concerted effort to bring the simplicity and cost model of Ethernet to the wide area network and to resolve the limitations of Ethernet in the WAN. Metro Ethernet Forum (MEF) was formed in 2001 and today it is a global industry alliance comprising more than 200 organizations including telecommunications service providers, network equipment/software manufacturers, semiconductors vendors and testing organizations [15]. The MEF develops Carrier Ethernet technical specifications and implementation agreements to promote interoperability and deployment of Carrier Ethernet worldwide. The MEF accelerates the worldwide adoption of Carrier-class Ethernet networks and services.

The five Carrier Ethernet attributes distinguish Carrier Ethernet from LAN based Ethernet for ubiquitous, standardized, carrier-class services and networks [15]:

- Standardized services
- Service management
- Scalability
- Reliability
- Quality of Service

To create a market in Ethernet services, it is necessary to clarify and standardize the services to be provided. MEF has played a key role in defining:

- E-Line: a service connecting two customer Ethernet ports over a WAN.
- E-LAN: a multipoint service connecting a set of customer endpoints, giving the appearance to the customer of a bridged Ethernet network connecting the sites.
- E-Tree: a multipoint service connecting one or more roots and a set of leaves, but preventing inter-leaf communication.

All these services provide standard definitions of such characteristics as bandwidth, resilience and service multiplexing, allowing customers to compare service offerings and facilitating service level agreements (SLAs). The Metro Ethernet Forum does not specify how Ethernet services are to be provided in a carrier network.

The key roles in developing Carrier Ethernet technologies and standards have been played also by the Institute of IEEE 802.1 and 802.3 standards committees. For example, IEEE 802.1 has addressed the scalability and management issues in the standards for Provider Bridges (802.1ad) and Provider Backbone Bridges (802.1ah). IEEE 802.3ah-2004 (Ethernet in the first mile (EFM)) refers to using of the Ethernet protocol between a telecommunications company and a customer's premise.

The Optical Internetworking Forum (OIF) and the Ethernet Alliance have also been working cooperatively with their members to enable future enhancements to Ethernet for the WAN while looking to the future speed of Ethernet technologies and services.

Some examples of vendor's Carrier Ethernet solutions, architectures and product information could be found on [16] [17][18].

VII. CONLUSION

Today Ethernet is unifying technology implemented in various environments, from enterprise networks, data centers to carrier networks.

Ethernet technologies continue to evolve. The latest revision of base standard IEEE 802.3-2012 was approved in September 2012 and it incorporated the IEEE 802.3ba (40 Gbit/s and 100 Gbit/s Ethernet), IEEE 802.3at (Power over Ethernet Plus) IEEE 802.3az (Energy Efficient Ethernet) and other amendments and improvements.

Carrier Ethernet uses many of the Ethernet LAN technologies but required further augmentation in order to function as a service delivery technology for MANs and WANs.

REFERENCES

[1] IEEE Std 802.3™-2008, approved 26 September 2008.

[2] IEEE Std 802.3ba™-2010 (Amendment to IEEE Std 802.3™-2008) approved 17 June 2010.

[3] Ethernet Alliance, "40/100GbE Technology Overview", October 2010

[4] www.brocade.com

[5] www.cisco.com

[6] www.extremenetworks.com/

[7] www.huawei.com/en/

[8] www.juniper.net

[9] Morty Eisen: "Introduction to PoE and the IEEE802.3af and 802.3at Standards", presentation slideware, 2010

[10] Ethernet Alliance, "Power Over Ethernet Plus", v2.0, 2008

[11] Cisco: "Cisco Universal Power Over Ethernet", white paper, 2012

[12] IEEE 802.3 Working Group: "Energy Efficient Ethernet CFI", presentation slideware, 2006

[13] Intel/Cisco: "IEEE 802.3az Energy Efficient Ethernet", white paper, 2011

[14] IEEE Std 802.3az™-2010 (Amendment to IEEE Std 802.3™-2008) approved 30 September 2010.

[15] www.metroethernetforum.org

[16] www.brocade.com/solutions-technology/service-provider/carrier-ethernet/index.page

[17] www.cisco.com/en/US/netsol/ns577/networking_solutions_solution_category.html

[18] www.juniper.net/us/en/solutions/service-provider/network-infrastructure/carrier-ethernet/

Pregled metoda ublažavanja utjecaja endogenih faktora na neuspješnost IT projekata

Nikša Marušić, Odjel za informatiku Sveučilišta u Rijeci, Rijeka, Hrvatska
niksa.marusic@uniri.hr

Sažetak - Rad opisuje najčešće endogene faktore koje utječu na neuspješnost IT projekata u području razvoja programske potpore. Navode se definicije uspješnosti, odnosno neuspješnost projekta u okvirima upravljanja projektima. Analizirana su moguća rješenja i metode za ublažavanje utjecaja endogenih faktora na konačni ishod projekta i na unaprijed definirane ciljeve. Naglašena je problematika velike fluktuacije ljudskih resursa u projektima programske potpore i projektima općenito.

I. UVOD

Neuspješnost projekata kao činjenično stanje najbolje opisuje statistike (CHAOS Chronicles – 2003.) koje navode postotak potpuno neuspješnih projekata na čak 66%, odnosno čak 82% projekata s kasnom isporukom, 52% projekata s krivim dosegom, dok je 43% projekata prekoračenog budžeta. Informacijska tehnologija (IT) i vezani projekti nisu izuzeti iz ove statistike. U ovakvom okruženju od velike je važnosti proučiti glavne razloge neuspješnosti projekata, te proučiti dostupne metode uspješnog upravljanja projektima u skladu sa zahtjevima i okolnostima projekta. Potrebno je analizirati rizike pojedinog projekta i usvojiti adekvatne metode upravljanja rizicima kako bi ostvarili glavni cilj upravljanja projektom, uspješnost projekta.

Uspješnost projekta opisuje se kao stupanj zadovoljstva svih dionika projekta samim projektom, odnosno projektnim produktom. Samim time faktori koji utječu na neuspješnost projekta su mnogobrojni i raspodijeljeni u svim fazama projekta. Činjenica je da glavni faktori neuspješnosti projekta proizlaze upravo iz lošeg upravljanja projektom [8]. To nam daje smjernice za daljnje istraživanje metoda ublažavanja utjecaja faktora na neuspješnost projekta kroz učinkovitije metode upravljanja projektima.

II. POJMOVI I DEFINICIJE PODRUČJA

Uspješnost projekta se prema mnogima definira u okvirima upravljanja projektima (eng. *Project Management*). Definicija uspješnog projekta nadovezuje se na definicije pojmova područja upravljanja projektima: projekt, proces upravljanja projektima, rizik u projektu, upravljanje rizicima projekta, uspješnost projekta.

A. Projekt i rezultati projekta

Projekt se definira kao vremenski određeno nastojanje da se proizvede jedinstven proizvod, usluga ili neki drugi rezultat [1]. Vremenska određenost, odnosno privremenost, navodi da svaki projekt, neovisno o svom trajanju, mora imati jasno određen početak i kraj. Projekt završava u trenutku kada su ciljevi projekta dostignuti, odnosno kada se zaključi da ciljevi projekta ne mogu ili neće biti dostignuti. Jedinstvenost se u projektu odnosi na rad na nečemu što prije nije postojalo i što se razlikuje od rezultata nastalih sličnim projektima [3].

Progresivna razrada projekta (eng. *Progressive elaboration*) opisuje postepenost razvoja projekta što podrazumijeva neprekidno poboljšavanje i detaljiziranje plana kroz niz ponavljanja u kojima prikupljene informacije postaju sve detaljnije, a procjene sve preciznije.

Rezultat projekta je najčešće proizvod ili artefakt koji se može kvantitativno odrediti, a koji može biti krajnji proizvod ili sastavna komponenta. Uobičajeno je to materijal ili roba. Rezultati projekta čine i sposobnosti obavljanja usluga, ishodi, dokumenti ili znanje.

Interesni sudionici projekta, dionici (eng. *stakeholders*) su [3]:

- Sponzor projekta (eng. *project sponsor*), osoba ili grupa koja osigurava (financijske) resurse za projekt;
- Projektni tim, ekipa (eng. *project team*);
- Potporno, prateće osoblje (eng. *support staff*);
- Naručitelj, klijent (eng. *customer*);
- Krajnji korisnici (eng. *users*);
- Dobavljači (eng. *suppliers*);
- Oponenti.

B. Upravljanje projektom

Upravljanje, rukovođenje projektima primjena je znanja, vještina, alata i tehnika u projektnim aktivnostima da bi se ispunili projektni zahtjevi [1]. Projekt je definiran i kao redoslijed aktivnosti, kao dio posla koji ima određen ulaz i izlaz.

Osnovni parametri projekta, po navođenju većeg dijela literature su [1] [3] [2]:

- Doseg, opseg projekta (eng. *scope*) koji definira granice projekta, što će biti napravljeno i što neće biti napravljeno u projektu;
- Kvaliteta projekta (eng. *quality*);
- Trošak projekta (eng. *cost*);

- Vrijeme (eng. *time*), raspored (eng. *schedule*) ;
- Resursi, sredstva (eng. *resources*)

Odnos između parametara projekta opisan je trostrukim ograničenjem (eng. *triple constraint*), odnosno ravnotežom projekta sačinjenu balansom zahtjeva za opsegom i kvalitetom, vremenom, troškovima te raspoloživih resursa.

Skup međuzavisnih akcija i aktivnosti koje se izvode kako bi se postignuo predodređeni skup proizvoda, rezultata ili usluga nazivamo procesom [2]. Razlikujemo dvije vrste procesa koje su međusobno povezane i zavisne [3]:

- Procesi orijentirani prema proizvodu (eng. *project-oriented processes*), specifični za određenu domenu i životni ciklus proizvoda;
- Procesi upravljanja projektima, zajednički svim projektima grupirani u šest grupa: pokretanje (eng. *Initiating*), planiranje (eng. *Planning*), izvršavanje (eng. *Executing*), praćenje i nadzor (eng. *Monitoring and Controlling*), završavanje (eng. *Closing*).

Svi procesi pripadaju određenim područjima upravljanja projektom, ili područjima znanja. PMI (*Project Management Institute*) Standard prepoznaje devet tematskih područja za upravljanje projektima [3]:

- Koordinacija projekta (eng. *Project Integration Management*)
- Upravljanje dosegom projekta (eng. *Project Scope Management*)
- Upravljanje vremenskim rasporedom projekta (eng. *Project Time Management*)
- Upravljanje troškovima projekta (eng. *Project Cost Management*)
- Upravljanje kvalitetom projekta (eng. *Project Quality Management*)
- Upravljanje ljudskim resursima projekta (eng. *Project Human Resource Management*)
- Upravljanje razmjenom informacija u projektu (eng. *Project Communications Management*)
- Upravljanje rizicima projekta (eng. *Project Risk Management*)
- Upravljanje nabavom za potrebe projekta (eng. *Project Procurement Management*)

C. *Upravljanje rizicima projekta*

Prema UK Standardu rizik se definira kao kombinacija vjerojatnosti (eng. *Probability*) pojave nekog događaja i njegove posljedice (eng. *Consequence*). Australijski/Novozelandski Standard definira rizik kao vjerojatnost da se nešto dogodi, što ima utjecaj na ciljeve projekta. US COSO okvir za upravljanje rizikom u poduzećima opisuje rizik kao unutarnje (endogene) i vanjske (egzogene) događaje koji utječu na postignuće ciljeva promatranog entiteta. Događaji mogu imati negativan upliv (prijetnje), pozitivan upliv (prilike) ili oboje istovremeno [6][7].

Upravljanje rizikom obuhvaća koordinirane aktivnosti kojima se upravlja organizacijom ili obavlja kontrola unutar neke organizacije vezano uz rizik [6]. Upravljanje rizikom znači suočavanje s brigom prije nego što ona preraste u problem ili krizu. Sastoji se od [3]:

- identifikacije rizika;
- odluke o postupanju prema pojedinom riziku;
- uklanjanja rizika.

Aktivnosti vezane za upravljanje rizikom potrebno je svoditi na razinu koja odgovara projektu. To u praksi znači da se mali projekti mogu uspješno provesti koristeći jednostavne liste rizika, dok se u velikim projektima provodi formalno upravljanje rizikom [3]. Prema Pressmanu (1993.) upravljanje rizikom dijelimo na sljedeće razine [3][6][7]:

- Upravljanje krizom (eng. *crisis management*), nakon što rizik postane problem;
- Prihvaćanje rizika (eng. *risk acceptance*), ispravka u trenutku pogreške (eng. *fixon failure*) - prepoznavanje i brza reakcija na rizik, neposredno po ostvarenju rizika, prije nego što postane problem.
- Ublažavanje rizika (eng. *risk mitigation*) - planiranje akcija koje treba poduzeti u slučaju da nastupi rizična situacija, ali ne učiniti ništa da bi se nastupanje rizične situacije spriječilo.
- Sprječavanje (eng. *prevention*) - izrada i provedba plana prepoznavanja rizika i sprječavanja problema koji su posljedica rizičnih situacija.
- Uklanjanje (eng. *avoidance*) uzroka problema - prepoznavanje i uklanjanje čimbenika koji omogućuju nastajanje rizika.

Projekt se prvenstveno mora uključiti u upravljanje na razinama sprječavanja i uklanjanja jer upravljanje rizikom u ostalim razinama vodi prema nemogućnosti poštivanja vremenskih rokova [3].

Prepoznavanje, identifikacija rizika sastoji se od utvrđivanja liste rizika, naročito onih koji bi mogli utjecati na vremenska kašnjenja projekta. Analiza rizika sastoji se od procjene vjerojatnosti i utjecaja svakog pojedinog rizika, te procjena rizika za različite alternative. Utvrđivanje prioriteta rizika sastoji se u utvrđivanju liste prioriteta rizika koja predstavlja temelj kasnije kontrole rizika. Izbjegavanje rizika, rizičnih aktivnosti ili odluka, predstavlja jedan od načina rješavanja rizika. Rizik se izbjegava izostavljanjem i nepokretanjem određenih projektnih faza ili funkcija ili korištenjem samo dokazanih tehnologija i postupaka [3][6].

D. *Uspješnost i neuspješnost projekta*

Uspješnost projekta definirana je kao stupanj zadovoljstva kod svih dionika projekta. Zbog kompleksnosti i definicije samog projekta, uspješnost projekta se ne može mjeriti u određenoj točki u vremenu, već se ona promatra kroz cijelo trajanje projekta [8]. U drugim izvorima uspješan je projekt samo onaj koji završi u planiranom roku, ne prekorači planirani proračun i kod kojeg je rezultat projekta upravo ono što je naručitelj tražio, s očekivanom, unaprijed utvrđenom razinom kvalitete. Pritom se kao mjerila uspješnosti projekta uzimaju rok, troškovi i kvaliteta [11].

Neuspješnost projekta definiramo kao suprotni događaj uspješnom projektu. Rizici u projektu direktno utječu na neuspješnost projekta. Faktore koji utječu na rizik dijelimo na [6][7]:

- Endogeni faktori rizika, odnose se na one faktore koje se pojavljuju unutar projekta ili projektne ekipe;
- Egzogeni faktori rizika, odnose se na one faktore koje su izvan projekta, te utječu na sve projekte u istim područjima djelovanja.

Upravljanje projektima podrazumijeva umješno balansiranje resursima (kako bi se željeni cilj ostvario u zadanom vremenu i bez prekoračenja dozvoljenih troškova) i prilagođavanje promjenama u stvarnom okruženju. Upravljanje projektima povećava vjerojatnost uspješnog završetka projekta [11].

TABLICA II. FAKTORI S LOŠIM UTJECAJEM NA ODVIJANJE PROJEKTA (THE STANDISH GROUP) [3]

Rbr.	Naziv faktora	Udio (%)
1.	Nedostatak informacija od strane korisnika	12.8
2.	Nepotpuni zahtjevi i specifikacije	2.3
3.	Promjena zahtjeva i specifikacija	11.8
4.	Nedostatak podrške izvršnog menadžmenta	7.5
5.	Pogrešna tehnologija	7.0
6.	Nedostatak resursa	6.4
7.	Nerealna očekivanja	5.9
8.	Nejasni ciljevi	5.3
9.	Nerealni vremenski rokovi	4.3
10.	Nova tehnologija	3.7
11.	Ostalo	23.0

III. METODE UBLAŽAVANJA UTJECAJA RIZIKA NA ISHOD PROJEKTA

Upravljanje rizicima projekta jezgrena je funkcija, ili područja znanja, upravljanja projektima [3]. Metode upravljanja projektima koje vode računa o uspješnosti projekta uključuju i metode za upravljanje rizicima projekta kao njihov sastavni dio. Metode upravljanja projektima u najširem smislu dijelimo na tradicionalne, dinamične i ekstremne [23].

TABLICA I. GLAVNI FAKTORI USPJEHA IT PROJEKATA (THE STANDISH GROUP) [3]

Rbr.	Naziv faktora	Udio (%)
1.	Sudjelovanje korisnika u projektu	15.9
2.	Podrška izvršnog menadžmenta	13.9
3.	Jasni zahtjevi	13.0
4.	Ispravno planiranje	9.6
5.	Realna očekivanja	8.2
6.	Postojanje malih projektnih miljokaza	7.7
7.	Kompetentni sudionici projekta	7.2
8.	Vlasništvo	5.3
9.	Jasne vizije i ciljevi	2.9
10.	Fokusirani tim koji naporno radi	2.4
11.	Ostalo	13.9

A. *Definicija metode*

Metoda je praksa, tehnika, procedura i pravilo koje se koristi na području određene discipline [2][4]. Metodologijom se mogu smatrati uloge, ekipe, vještine, procesi, tehnike, alati i standardi koje rabi projektna ekipa [14].

Dobra metodologija sadržava sve važne procese upravljanja projektom. Primjenom dobre metodologije upravljanja projektom omogućuje se bolja kontrola nad dosegom projekta što dovodi do smanjenja rizika projekta, boljeg procesa donošenja odluka, većeg zadovoljstva sponzora te više vremena za dodane vrijednosti. Trebamo imati na umu da ne postoji jedinstvena metoda, pomagalo niti proces primjenjiv na sve projekte. Odabir metoda ovisi o tipu, veličini i ostalim značajkama projekta i njegove okoline [3][4].

B. *Tradicionalni pristup*

Tradicionalni pristup upravljanju projektima (eng. *Traditional project management*) temelji se na upravljanju ljudima. Temeljni principi uključuju niz tehnika za planiranje, predviđanje i kontrolu aktivnosti radi postizanja željenog rezultata prema zadanim specifikacijama u određenom vremenu i s određenim troškovima [3][4][14].

Glavne faze životnog ciklusa tradicionalnog pristupa upravljanja projektima su [3][15]:

- Definiranje zadataka i izvršitelja – faza u kojoj se određuje doseg projekta i popisuju svi rizici koji mogu utjecati na uspješnost projekta;
- Planiranje – faza u kojoj se određuju svi zadaci i dodjeljuju se članovima tima;
- Izvršenje – faza u kojoj se određuju potrebni resursi, dodjeljuju zadaci prema planu, raspodjeljuju članovi tima i pokreću planirane aktivnosti;

- Kontrola – faza nadzora planiranih aktivnosti i reagiranja na promjene;
- Završetak – faza formalnog zatvaranja, evaluacije, oslobađanja resursa i predaje projekta.

U tradicionalnom modelu naglasak je postavljen na upravljanje kvalitetom proizvoda, usluga i procesa projekta, ali i na upravljanje projektom općenito [4].

Faza kontrole najčešće služi za upravljanje rizikom unutar projekta. Promatraju se sve aktivnosti unutar projekta kako bi se pravodobno reagiralo na promjene. Nažalost, proces reagiranja na promjene nije učinkovit zbog krutosti i slabe prilagodljivosti tradicionalnog pristupa upravljanja projektima u odnosu na novonastale promjene. To znači da se za svaku novonastalu promjenu mora ponovno proći kroz sve faze životnog ciklusa kako bi se povećala uspješnost samog projekta.

U tradicionalnom pristupu postoje jasni korisnički zahtjevi i mala očekivana količina zahtjeva za izmjenama. Naglasak je na formalnoj dokumentaciji, a članove tima se marginalizira i ne uzima u obzir kod rizika povećane fluktuacije radne snage koja dovodi do disperzije znanja izvan projektnog tima, smanjene kvalitete rezultata i na kraju neuspješnosti projekta.

Tradicionalni model zbog svoje robusnosti i jasno definiranih koraka djelovanja koristi se u najkompleksnijim projektima dužeg trajanja i u organizacijama sa jasno definiranom, rigidnom strukturom s jasno definiranim razvojnim procesima i funkcijama [24].

C. *Dinamični i ekstremni pristup*

Dinamični pristup upravljanju projektima (eng. *Dynamic project management*) javlja se kao nadopuna tradicionalnog pristupa u pogledu kontrole troškova tijekom projekta i ubrzanim razvojem, odnosno isporukom rezultata pojedinih aktivnosti projekta. U literaturi dinamični pristup navodi se još kao agilni pristup ili adaptivni pristup [15][23][24], dok se kod drugih agilni pristup navodi kao krovni i idejni pristup ekstremnog i adaptivnog upravljanja projektima [20].

Ekstremni pristup nadopunjuje dinamični pristup upravljanju projektima u ograničenjima vezana za povećanje koristi proizašle iz projekta, u bržem razvoju i učinkovitijeg upravljanja velikim promjenama unutar projektnog plana.

Dinamični, ekstremni, agilni ili adaptivni pristup u konačnici su varijante inkrementalnog upravljanja i razvoja kod kojih se isporuka konačnih produkata projekta vrši u ciklusima. Iako se u različitim literaturama spominju i druge metodologije razvoja (adaptivna metodologija, objektno orijentirana metodologija, metodologija temeljena na rješenju, evolucijska metodologija i druge) agilna metodologija (eng. *Agile project management*) jedna je od najprihvaćenijih. Razlog tomu je i izdani Manifest za agilni razvoj (eng. *The agile manifesto*) u kojemu su postavljene četiri temeljne vrijednosti svakog projekta [22]:

- Odnosi i osobe važniji su od procesa i alata;
- Gotov i funkcionalan proizvod važniji je od opsežne dokumentacije;
- Suradnja s korisnicima važnija je od pregovaranja oko ugovora;
- Reagiranje na promjene važnije je od striktnog slijeđenja plana.

Takve temeljne vrijednosti metodologije donose brzi razvoj i isporuku projektnih produkata. Poticanjem komunikacije između članova tima i korisnika ubrzava se proces usuglašavanja i ažuriranja specifikacija. To u konačnici rezultira zadovoljstvom korisnika što je jedna od karakteristika uspješnog projekta. Kratki i učestali sastanci članova tima poboljšavaju komunikaciju, prijenos znanja i informacija u projektu. Time se učinkovitije reagira na promjene u bilo kojem obliku, od velike fluktuacije radnika do tehnološkog rizika i rizika promjene početnih korisničkih zahtjeva.

Najprikladnija primjena agilnih metodologija je u projektima s velikom nesigurnošću i visokim razinama promjena, iako se i u ostalim projektima mogu primijeniti, ali uz nadopune sa tradicionalnim dijelovima [4].

IV. ZAKLJUČAK

Projekti vezani uz informacijsku tehnologiju podložni su velikom broju faktora koji direktno utječu na uspješnost, odnosno neuspješnost projekta. Pokazano je da kvalitetnim upravljanjem projektima, što uključuje i upravljanje rizikom, možemo efikasno upravljati krizom, prihvatiti, ublažiti, spriječiti i ukloniti rizik.

Moderne metode upravljanja projektima, kao i agilne metodologije, potenciraju komunikaciju i prijenos znanja unutar samog projekta i zanemarivanje nepotrebnog posla kako bi se endogeni rizici prihvatili i riješili u što kraćem vremenu. Upravljanje rizicima u takvim okolnostima postaje učinkovitije. Brzim prijenosom znanja sprječavamo i rizik fluktuacije ljudskih resursa kao jednog od čestih uzroka neuspješnosti projekta i smanjene kvalitete projektnog produkta.

U ovom kontekstu važnost upravljanja projektima je neupitna. Metodologije upravljanja projektima prilagođavaju se ciljanoj organizaciji i projektu s obzirom na vrstu i trajanje projekta, organizacijsku strukturu, potrebne resurse i iskustvo organizacije u upravljanju projektima. Na taj se način ciljano vodi upravljanje projektima što u konačnici povećava vjerojatnost uspješnosti projekata.

LITERATURA

[1] APM Body of Knowledge, fifth edition , Association for Project Management

[2] Project Management Institute, A Guide to the Project Management Body of Knowledge (PMBOK Guide), 2000.

[3] K. Fertalj, Upravljanje informacijskim sustavima. Zagreb: Fakultet elektrotehnike i računarstva, 2009.

[4] M. Špundak, „Upravljanje projektima – definicije i metodologije“, kvalifikacijski članak, VIPnet d.o.o., Zagreb.

[5] M. Pavlić, Informacijski sustavi, Školska knjiga, Zagreb, 2011.

[6] N. Debrecin, Upravljanje rizikom, Zagreb: Fakultet Elektrotehnike i računarstva

[7] IRM, A Risk Management Standard. IRM, 2002

[8] Cortex, Project Failure Statistics – Failure Causes, http://www.it-cortex.com/Stat_Failure_Cause.htm [18.02.2013]

[9] Al Neimat, Taimour. "Why IT projects fail." The Project Perfect White Paper Collection. Retrieved November 1 (2005): 2011.

[10] J. Hiner, TechRepublic15 glavnih razloga za neuspjeh projekata

[11] D. Jureković, Uvod u upravljanje projektima; Udruga za projekt menadžment, PMI ogranak Zagreb, 2011

[12] B. Whittaker, What went wrong? Unsuccessful information technology projects;

[13] M. Latković, Upravljanje rizicima: identifikacija, mjerenje i kontrola, HAGENA, Zagreb, 2001.

[14] A. Cockburn, People and Methodologies in Software Development, Doktorska disertacija, Sveučilište u Oslo, 2003.

[15] Kerzner, Harold. Project management: a systems approach to planning, scheduling, and controlling. Wiley, 2009.

[16] D. S. Kezsbom, D. L. Schilling, K. A. Edward, *Dynamic project management: a practical guide for managers and engineers*. Wiley, 1989.

[17] J. Knudson, W. Lawrence, M. S. Crego. "Dynamic project management system." U.S. Patent No. 5,765,140. 9 Jun. 1998.

[18] L. S. Hyun, Feniosky Peña-Mora, and Moonseo Park. "Dynamic planning and control methodology for strategic and operational construction project management." *Automation in construction* 15.1 (2006): 84-97.

[19] Highsmith, Jim, and James A. Highsmith. *Agile project management: creating innovative products*. Addison-Wesley Professional, 2009.

[20] A. Cockburn, J. Highsmith. "Agile software development, the people factor." *Computer* 34.11 (2001): 131-133.

[21] J. Highsmith, "Agile project management: Principles and Tools." *Agile Project Management Advisory Service Executive Report* (2003).

[22] M. Fowler, J. Highsmith. "The agile manifesto." *Software Development* 9.8 (2001).

[23] Wysocki, R. K. Effective Project Management: Traditional, Adaptive, Extreme. Wiley, 2006.

[24] Wysocki, Robert K. Effective project management: traditional, agile, extreme. Wiley, 2011.

[25] Goodpasture, J. C. Quantitative Methods in Project Management. J. RossPublishing, Inc. 2003.

[26] Charvat, J. Project Management Methodologies: Selecting, Implementing, and Supporting Methodologies and Processes for Projects. Wiley, 2003.

[27] "Design Patterns and Refactoring", University of Pennsylvania, 2003,

[28] B. Jarvis, S.P. Gristock, „Extreme Programming (XP) Six Sigma CMMI“

[29] M. Miller, The Analysis of IT Project Success: IT Project Success & Failure, 2003.

INFRASTRUCTURE DEVELOPMENT

Terastream
Implementation of All IP New Architecture

B. Zaluški, B. Rajtar*, H. Habjanić, M. Baranek, N. Šlibar, R. Petračić, T. Sukser
Croatian Telecom, Zagreb, Croatia
* Deutsche Telekom, Bonn, Germany

Abstract: In this paper we describe Terastream all IP architecture including market and economic drivers, IPv6 addressing scheme, routing and tunneling protocols, interfaces and optical integration, front end data centers, home networking and performance aspects initial findings. We point out standardization efforts where we are contributing and present initial findings after actual implementation of pilot network.

I. INTRODUCTION

Work on new IP architecture started in 2011 within Deutsche Telekom. In 2012 pilot Terastream network was implemented by Croatian Telecom. In 2013 we will continue research activities and provide additional contributions to standardization work performed by IETF and other standardization organizations.

This paper is describing architecture, implementation and initial findings during implementation. New research activities are only mentioned where appropriate.

II. MARKET AND ECONOMIC DRIVERS

Over the top providers, cable operators, mobile operators and fixed operators are in fierce competition. They are facing massive IP traffic growth driven by improved broadband access technologies, new Internet services and new business models. Mobile and Fixed operators are moving from legacy multi-layer networks with too much complexity.

In order to reduce complexity and enable service innovation, more features, faster provisioning resulting in revenue increase — operators need simplification and one truly converged de-layered network to enable cloud era economics.

With Internet traffic growth forecasts increasing from conservative 10 times to more than 50 times in next ten years operators are facing major changes. Ongoing migration from IPv4 to IPv6 is adding to complexity. In order to get revenue and cost balance right, new service delivery model based on simplified IPv6 network architecture is needed.

III. ARCHITECTURE OVERVIEW

Traditional models produce IP connectivity using separate networks for telecom services and other providers in a way that is too costly to provide for traffic growth. In order to enable more economical IP connectivity Terastream relies on following design principles:

- Reduce the amount of technologies used
- Use IPv6[1] only for all internal functions and services
- Size the network to handle IP traffic without packet losses
- Integrate optical and IP network as much as possible
- Avoid internal interfaces
- Use one network for all services

Architecture is using two types of routers. One we call R2 - they deal with peering, data center connectivity and are meshed among them self to produce telecom network. They have dual stack capability IPv4/IPv6, huge number of 100GE ports to connect so called R1 routers in a horseshoe and number of 10GE ports for directly connecting front end data center computers.

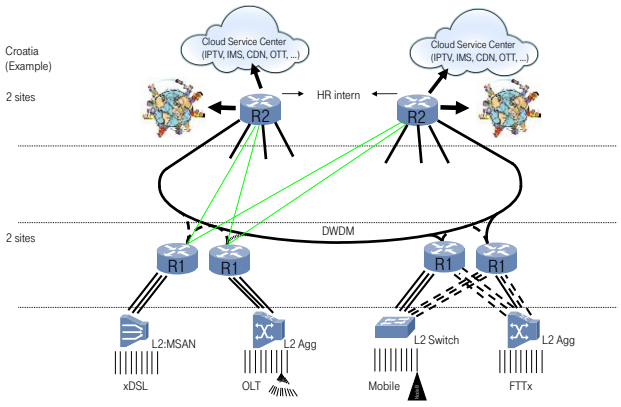


Figure 1. Terastream Architecture Overview

In order to benefit from x86 architecture, network services are implemented using off the shelf servers connected directly to R2 routers. Services like L2 and L3 virtual private networks or IPv4 through DS-Lite[2] tunnel termination require throughput and low latency offered by front end data center infrastructure. Store and forward services like mail or messaging are implemented in regular data centers.

IV. IPv6 ASPECTS, ROUTING AND TUNNELING

IPv6 address uses 128 bits which enables huge address space with near 3.4×10^{38} available addresses. This solves shortage of IPv4 address and its address translation mechanism (except in some special cases).

There are different address scopes in IPv6: local, global addresses and Unique Local Addresses - ULA. ULA addresses are equivalent of IPv4 private addresses.

Addresses can be generated automatically by the client without preserving state using *SLAAC – Stateless Autoconfiguration* which relies on RA (*Router Advertisement*) messages from the router. DHCPv6 protocol can also be used for better control.

Service providers used to build hybrid networks using Ethernet in aggregation part and IP/MPLS in the core part.

TeraStream is pure IP network which lowers the number of used protocols and simplifies network management.

More specifically TeraStream is pure IPv6 network because all other protocols are tunneled (if needed). That includes IPv4 protocol which becomes only one of the network services. Tunneling is done between the Home Gateway and the network AFTR (Address Family Translation) service that sits in the Data Center (using DS-Lite or Lightweight 4over6) as shown on Figure 2.

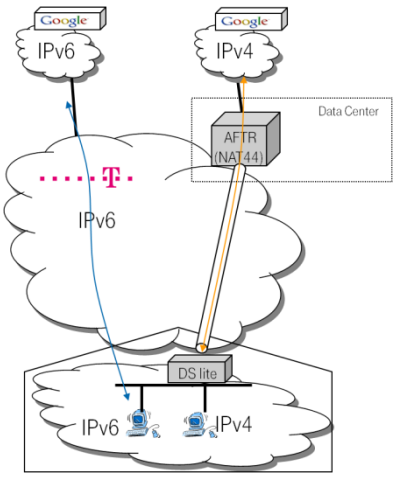


Figure 2. IPv4 as a service over IPv6 using DS-Lite

Routing in Terastream is done using familiar IS-IS and BGP (Border Gateway Protocol) protocols. IS-IS protocol is used only for exchanging loopbacks and interface addresses. Whole network is one IS-IS Level 2 domain. All customer prefixes are exchanged using BGP with R1 routers being RR clients of R2 routers. R2 routers can have external BGP peerings.

In Terastream concept IPv6 address bits (Figure 3.) have special meaning attached to them:

- Defining if address is public, belongs to infrastructure or service (so called PIE bits),
- Defining service type (so called SSS bits, they have values for Internet, IPTV or other service)
- Defining user location (router, interface,VLAN)

```
Examples:                          Source              Destination
                                     PIESSS               PIESSS
  ------------------------------------------------------------------
   User -> IMS                       000110               011110
   IMS  -> User                      011110               000110
   User -> User (best effort)        X00001               X00001
   User -> Internet (best effort)    100001               XXXXXX
   Internet -> User (best effort)    XXXXXX               100001
   Lan-Lan service                   010101               010101


R Registry 32 bit, Registry/IANA assigned
P Public 1 bit, 0=traffic internal to Telekom, 1=traffic external to Telekom
I Infrastructure 1 bit, 0=end user traffic, 1=infrastructure packet
E Endpoint/Service 1 bit, 0=network endpoint, 1=service
S Service type 3 bits, 0=res, 1=internet, 2=res, 3=res, 4=video, 5=L2 service, 6=voice, 7=mgmt
a R1 Area 5 bits, Indicates what R1 that the address is delegated from, max 32 R1
p R1 User 13 bits, User identifier
u User subnet , Delegated to user
```

Figure 3. Terastream IPv6 addressing PIESSS examples

Each customer gets /56 prefix for each of the used services (Internet, IPTV, etc.). Concept enables not only connection of the HGW but also end devices to the Terastream network.

Idea is to also enable more complex topologies so the home network can have connections to multiple service providers or just be multihomed.

Multi-functional end devices can get multiple address prefixes, one for each service. Current IPv6 standards use longest prefix match rules to choose the right source prefix for the destination. As such approach has many limits suggestion is to use so called “coloring” mechanism which will “color” the prefix given to the customer. Each device or application will then be able to choose the right “color” to reach the desired service prefix (that will also be colored). Policies could be preconfigured or downloaded from network management server.

V. IP AND OPTICAL INTEGRATION

Key idea is integration of 100G transponders in routers without classical Dense Wavelength Division Multiplexing – DWDM – components: (ROADM, MUX, DEMUX) using only splitters and amplifiers. This is now possible using coherent tunable transponders which can tune its transmitters and receivers to any of 80 optical channels.

With this approach we can avoid double optical/electrical conversion and reduce number of interfaces as shown on Figure 4 with separate transponders and on Figure 5 with integrated transponders.

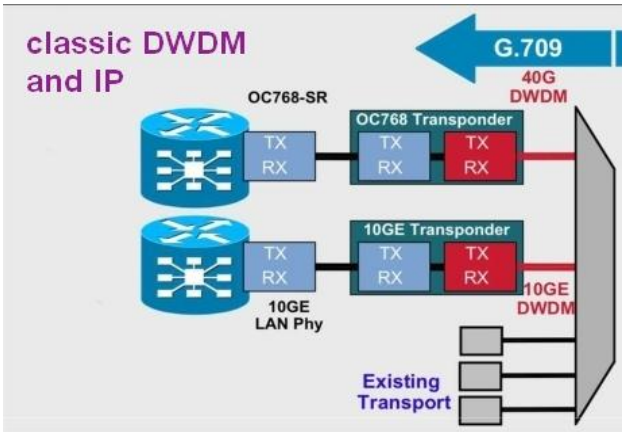


Figure 4. Classic concept with DWDM and IP separated ports

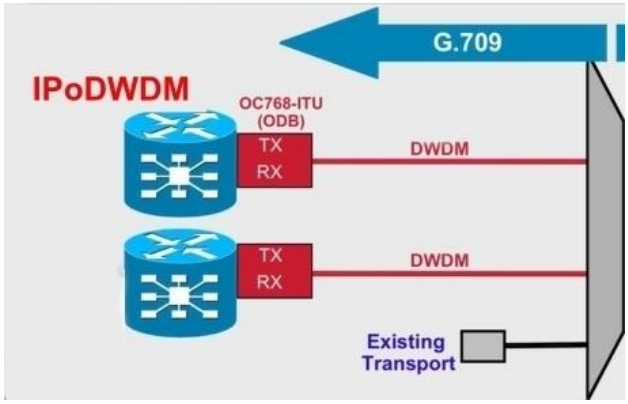


Figure 5. IPoDWDM with integrated transponders

Client ports in router and transponder are not needed since transponder is integrated in router card. This concept reduces CAPEX costs and provide improved resilience of the network (less OPEX cost).

Each R1 (max 8) site has a splitter combination to add / drop channels and is connected to two different R2 routers to enable services in case of fiber interruption. Optical connectivity is shown on Figure 6.

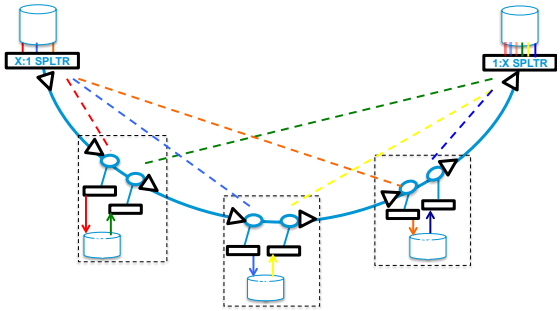


Figure 6. Terastream optical connectivity

All circuits are 100Gb/s[3] Ethernet point-to-point utilizing 50GHz spacing. Typical overall distance of the horseshoe is 600Km.

The solution must also be adapted to the Core network of meshed R2 sites (not depicted).

Each R1 site is equipped with 8 optical channels. We start with utilized two 100Gb/s channels per R1 site.

The DWDM network leverage a "Drop and Waste" architecture where every wavelength that is dropped at an R1 site will be wasted or not available to upstream nodes as shown on Figure 7.

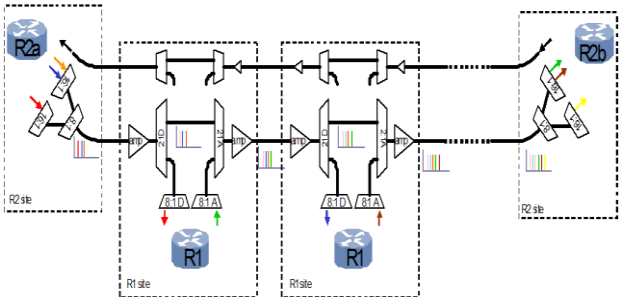


Figure 7. Terastream "Drop & Waste" architecture with splitters

Important fact in this architecture is that every optical channel is available at every transponder due to the usage of splitters and not filters as in classical DWDM system. With previous transponder technology this architecture would be impossible to implement.

Coherent technology enables us to simplify optical network using only splitters and amplifiers from classical DWDM systems eliminating need for expensive ROADM, MUX/DEMUX and other components.

This concept is applicable in presented "Horseshoe" topology with 600km span limit.

VI. Front End Data Center

Data center directly attached to the TeraStream network is called front end data center. The name was derived from its function. Such data center primarily runs front end (customer facing) services. Data centers traditionally offer compute, storage and network resources, while here it is a platform for hosting:

- basic network service
- voice (VOIP), television (IPTV)
- content delivery network

The front end data center is designed in form of cloud, which enables high flexibility, agility and scalability for the services hosted in it. By using standardized cloud platform[4] with well-defined API (Application Programming Interface) it is possible to decouple hardware from software, so that servers from different vendors can be used in the same data center more easily. This approach would enable easier replacement of hardware in future.

Services which are running in the data center are delivered in form of virtual machines, where each virtual machine would contain the software for the service. Each service can run its own operating system which is the best for it, reducing the compatibility issues.

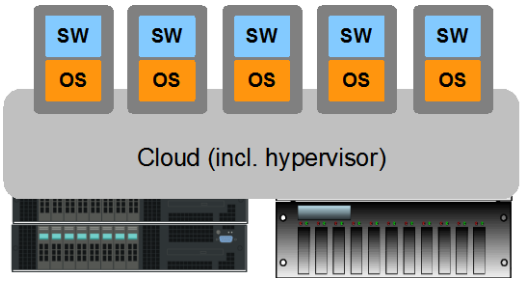


Figure 8. decoupling hardware from software using hypervisor

Having a single data center might be beneficial for easier management, but in order to bring services closer to customers, it is necessary to have the data center at each R2 site. Management of multiple data centers, each running lots of services, is a nightmare and requires a lot of people involved.

By using orchestration tools, it is possible to reduce human errors, improve service delivery times and reduce service outages. The most important functions of the orchestrator:

- configuration of the whole data center from the scratch,
- start and configuration basic data center and network services (DNS, NTP, orchestrator, logging facilities),
- deployment and configuration of a service (as a virtual machine or set of virtual machines),
- tearing down a service (removal from the data center),
- detection and correction of service, virtual machine or physical server failures,
- communication with other network management systems in order to have entire end-to-end overview,
- reporting of data center usage and trend prediction.

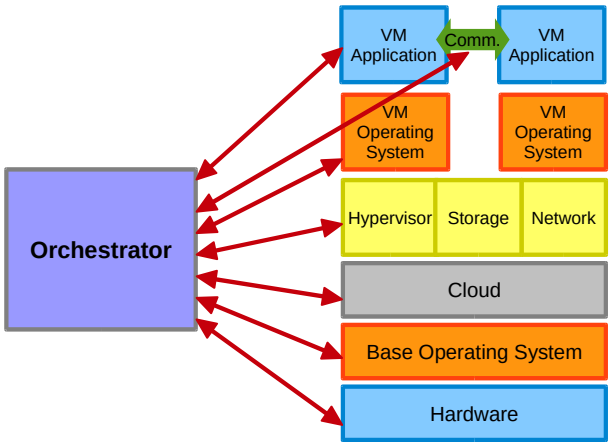


Figure 9. Orchestrator interfacse

Orchestrator acts on different layers as shown on Figure 9: hardware, base operating system on physical servers, cloud software (including hypervisor, storage and network resources), virtual machines as virtual appliances, operating system inside virtual machines, applications inside virtual machines, and helper for establishing communication between various virtual appliances.

VII. Home Networking

With the introduction of IPv6 in the home, customers now receive 2^72 IP addresses they can use for networking in the home. Comparing to the IPv4 addressing space, this means the customers can connect more devices at their home than they are existing devices on the Internet today reachable via IPv4. This allows for all sorts of opportunities such as home automation, connecting all home appliances into one inter-connected network, etc.

The addressing space for IPv6 is a public addressing space which means the customer is directly available from the Internet and is "on" the Internet rather than "connected" to the Internet. Unlike IPv4, where the lack of addresses implies that each customer has only one public IPv4 address, the whole customer IPv6 addressing space is public. NAT – Network Address Translation from the single public IPv4 address to the private addressing space (typically 192.168.1.0/24) used for IPv4 home networking brings a lot of limitations regarding home accessibility from the Internet and is non-existent with IPv6.

The typical scenario of Internet connectivity uses a Home Gateway as a central point of home administration and a point of interconnection. However, the TeraStream architecture allows that the customer connects himself directly to the R1 router and uses only low-end switches to build their own home network. However, this is possible only with devices which are IPv6 capable.

The Home Gateway is used for three primary reasons:

- Firewall used for the protection of the home against malicious attacks from the Internet;
- Access point for wireless connectivity.
- IPv4 connectivity.

Since nowadays a lot of existing equipment are IPv4-only, the Home Gateway is used as a endpoint for tunneling customer IPv4 traffic. This is accomplished using DS-Lite, a protocol that encapsulates customer IPv4 traffic (from the private address space) and sends it to the data center where NAT to the public IPv4 address is performed. The traffic is then forwarded to its final destination on the IPv4 Internet.

VIII. Conclusion

Terastream pilot in Croatian telecom was implemented in less than three months. It proved number of new IPv6 technologies including DS-Lite — by using front end data center it was shown that it is possible to implement network functions like DS-Lite on standard PC hardware in cloud environment — as well as that it is possible to build 100GE coherent drop & waste optical system with integrated optics managed by routers on IPv6 only network.

Acknowledgment

We would like to thank colleagues from Deutsche Telekom that made simplification and improvements of traditional IP architecture possible in Terastream architecture drafted number of early design documents that started pilot in Croatia: Peter Lothberg, Axel Clauberg, Rainer Schatzmayr, Guenter Honisch and many others.

References

[1] Hinden, RM. "RFC 4291 - IP Version 6 Addressing Architecture." 2006.

[2] Durand, Alain et al. "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion." draft-ietf-softwire-dual-stack-lite-04 (work in progress) (2010).

[3] Winzer, Peter. "Beyond 100G ethernet." Communications Magazine, IEEE 48.7 (2010): 26-30.

[4] 26-30.Computing, Rackspace Cloud. "OpenStack Open Source Cloud Computing Software." 2010-12-10. http://openstaek, org (2012).

Merging Dynamic Address Autoconfiguration and Security Key Protocols in Mobile Ad Hoc Networks

Mohammad M. Shurman[1], Mamoun F. Al-Mistarihi[2], Khalid A. Darabkh[3]
[1]Jordan University of Science and Technology/ Network Engineering and Security Department, Irbid, Jordan
[2]Jordan University of Science and Technology/ Electrical Engineering Department, Irbid, Jordan
[3]The University of Jordan / Computer Engineering Department, Amman, Jordan
alshurman@just.edu.jo, mistarihi@just.edu.jo, k.darabkeh@ju.edu.jo

Abstract - **Mobile ad hoc network (MANET) is formed by collaboration of nodes characterized by having the same capabilities whereas neither central infrastructure nor central coordination is required. Notwithstanding the communication between nodes in MANET requires each node to have a unique identity (IP-address), many critical networks require some level of security in which all security principles require the node to have a security key. A tree-based dynamic address autoconfiguration protocol (T-DAAP) is found to be one of the best protocols for address assignment based on network throughput and delay. MANET security ideally takes place during the design phase of the routing module. In this paper, we propose a novel scheme that merges the process of assigning IP addresses and security keys to the nodes in MANET using one protocol such that as soon as a node enters the network, it will be assigned an IP address and security key. To the best of our knowledge, no single protocol provides both IP addresses and security keys for MANET nodes. The results of the proposed method, which is based on T-DAAP, show significant reductions in the control packets and packet delay. Additionally, it provides not only efficient security to the nodes from the first moment they join the network, but also secure delivery of IP addresses and security keys to the network nodes.**

Keywords: Mobile Adhoc networks, control packets, packet delay.

I. INTRODUCTION

In adhoc networks, nodes at the lead of a network joining must be assigned an IP address or identification to communicate with other nodes without confusion. In addition to identification, a shared security key must be established before communicating with other nodes. Symmetric cryptography schemes are so convenient for ad hoc networks while the public key cryptography schemes are not feasible for many reasons [1]. Node address assignment is an active research topic since it can be performed using several techniques [2]. Additionally, the security key assignment in MANETs is also of interest, but unfortunately there are no widely-accepted standards among researchers heretofore.

Nodes' addresses in MANET can be assigned using centralized, decentralized, or neighbor-based approaches. Centralized approaches use a central server for address assignment in a newly joining node. The disadvantage of these methods is the possibility of a single point of failure in the central server which can disrupt the entire network [3]. Decentralized approaches allow the node to pick an IP address and then check for uniqueness [4]. The main disadvantage of these methods is that the verification required for address uniqueness/conflicts with other nodes is accomplished by network-wide flooding. In neighbor-based approaches, the network does not suffer from wide flooding. Achieving a unique address requires a flooding for each node to its neighbors only. These approaches have been found to be the best for MANET in terms of delay, flooding, and throughput [1, 2, 5].

For node address assignment in MANET, many neighbor-based protocols have been proposed [6-11]. It is worth mentioning that all of these protocols were designed for address assignments only without incorporating any secure protocol. Tree-based dynamic address auto-configuration protocol (T-DAAP) [6] divides the network into zones and therein each zone, a leader node responsible for address assignment. Distributed-based dynamic address assignment protocol (D-DAAP) [7] is based on an even distribution of address pool between the new node and its allocator. MANET configurations [8, 9] allow any new node to pick a random IP address and accordingly check with the rest of network nodes if this IP address is unique or not.

When a MANET requires security, it will be required for each node to have a security key. Key assignment can be performed using either central or distributed approaches [12]. Nevertheless, due to the nature of MANET, central approaches suffer from many obstacles [1], out of which the saved keys may be revealed, the central server may be stolen, hacked, or intercepted which in turn lead to a very serious security breach. Additionally, it may go offline for a number of reasons (i.e., power failure) and stop distributing the keys to network nodes. Another concern may be the range of accessibility because of network partitioning [1].

Threshold cryptography is a cryptography scheme suited for mobile networks [13-15] since it provides robustness and defense against a single point of failure in the central server schemes [1]. In threshold cryptography, the central server function is distributed among a group of servers. When one of these servers is compromised or fails, the rest of the group can take the place of this server. This distribution provides not only a cooperative security model, but also robustness against a single point of failure.

The main drawback of threshold cryptography is the difficulty of applying the distributive function among wireless nodes [16].

Most research ideas in MANET security and address management assume two protocols (i.e., one protocol for address assignment and another for security key assignment). Few papers have been published concerning the autoconfiguration and key distribution in adhoc networks [6]. The idea of these approaches is to correlate IP address with a security key which incurs many problems include NATing of local address, fixed mapping of IP, and security key generation. In this paper, we merge using IP address assignment and key distribution into one protocol. The proposed protocol assigns IP addresses to the nodes in a way complies with the neighbor-based approaches and delivers security keys to these nodes in a distributed manner resulting in preserving the network power and reducing the flooding as well. The novelty of the proposed protocol emerges by enforcing security from the beginning of node joining the MANET. Furthermore, the performance improvements of applying the proposed protocol are extremely impressive. In other words, the simulation results demonstrate that the proposed protocol delivers security keys and IP addresses to the nodes with about 26% lower packet delay and 30% lower number of assignment packets compared to what are obtained using separate protocols.

The paper organization is as follows. Section II details the proposed protocol for key and IP address assignments in MANETs. Section III summarizes the simulation environment and results in addition to the significance of the proposed protocol. Conclusion and future research ideas are presented in Section IV.

II. PROTOCOL DESCRIPTION

This section presents a secured and distributed IP address assignment scheme. The proposed protocol has similarities with T-DAAP for address assignment. The proposed protocol overlaps address assignment with security key assignment into one protocol. A newly joining node during the joining process receives an address and a security key to associate itself with the network. The network is divided into zones where each zone has one leader, taking into consideration that the whole network has only one root. According to their functions, network nodes can be classified into three categories include root, leader, or normal node as follows:

A. *The* Root

Only one node in the network should be in this state. The root controls many protocol functions. Its role is as follows:

- Maintains all zones' leaders IP addresses and their corresponding number of free IP addresses, leaders' indices needed for candidate root selection, and shared security key of each one in its database.
- Periodically sends unicast *rootAlive* messages to every leader to announce its presence provided that the *rootAlive* message contains the root IP address, leaders set, and number of free IP addresses for each leader. When the leader receives the *rootAlive* message, it will update its database with the new information of the other leaders. Moreover, the leader replies by *leaderAlive* message that contains its IP address and number of free IP addresses.
- Responsible for security key distribution to the leaders.
- Once a security breach is detected, it is responsible for key invalidation and rekeying the nodes.
- Responsible for generating the security matrices G and D.
- Has common security key with every leader under its supervision.

B. *The Leader Nodes*

Leader nodes exist in the MANET tree structure under the supervision of the root. The number of leaders in MANETs varies and depends on the area of the network. Leaders have knowledge of available free IP addresses where each leader has a disjoint set of free IP addresses in its free IP address pool. In addition, each leader retains information of the other leaders which includes their IP addresses and corresponding number of free IP addresses. This does not accrue a heightened load on the network since it can be accomplished easily by exchanging the triggered updates advertisement packets between the leaders. Basically, these leaders are responsible for IP address and key assignments to the new nodes joining the network. Matrix D_{zl} is known only by the zone leader. Leaders also preserve a table of zone node indices along with the shared security key between the leader and every node. The leader broadcasts its presence periodically and implicitly by including the node type in the *hello* messages. After initialization, each normal node in the MANET has to be in the neighborhood of at least one leader.

C. *The Normal Nodes*

The normal nodes can relay IP addresses and security keys assignments packets to the new nodes if the leaders are outside their transmission range. Furthermore, normal nodes have the generator matrix G distributed among them in a (n, t) threshold manner, where n is the number of nodes inside the zone and t is the threshold value.

D. *Network Initialization*

At the initialization stage of the network, there will be a lower number of nodes than the threshold value. However, the following steps occur during initialization:

- The network follows T-DAAP initialization phase (assigns IP addresses only) until there is enough number of nodes in which these nodes will be the zone leaders (*no. of nodes* ≥ *threshold*).

- Once the number of nodes reaches the threshold, the root generates G_{R_temp} and D_{R_temp} matrices.
- The root sends to each node (where these nodes will be the zone leaders) a row of G_{R_temp} to satisfy the threshold behavior.
- The root sends to each node a row of A_{R_temp} ($A_{R_temp}=(D_{R_temp}\times G_{R_temp})^T$), such that a leader i receives i^{th} *row of* A_{R_temp}.
- Each leader collaborates with other leaders to calculate its initial key chain using k_{i_temp}= (i^{th} *row of* A_{R_temp})$\times G_{R_temp}$.
- The root now generates a new G_R and distributes it to the leaders using their initial keys.
- The root generates new A_R and delivers to each leader its corresponding row.
- Each leader collaborates with other leaders to calculate its key chain using k_i=(i^{th} *row of* A_R)$\times G_R$.
- The root backs up its information (leaders' IP addresses and security keys control information) on the leader that has the lowest index to be the candidate root.
- Now, leaders can move to cover all the network area. Communication between any two leaders can take place securely even if they are out of range to each other since mutual shared keys are established between them (these nodes are the zone leaders).

The same steps are carried out by any leader located within its zone. The leader has to generate another G_{zi} and D_{zi} and distribute them to the normal nodes once there is enough number of nodes (*no. of nodes* ≥ *threshold*) inside its zone. The leader then backs up its control information on one of its zone nodes that has the highest IP address value. It must be mentioned that the threshold value of the leader does not have to be the same as that of the normal nodes, that is to say, the number of leaders do not have to be the same as the number of zones' normal nodes.

E. Node Joining the Network

After the network initialization, any node that joins the network, which is called a requestor, will receive replies for its *addressKeyRequest* broadcast message. The reply message contains the assigned IP address for this node and a row from A_{zi} corresponding to the node index at the zone leader. After the *addressKeyReqTimer* timer expires, the new node checks the responses and sets the *allocatorChosenTimer* timer. According to the received responses, one of the following possibilities may occur:

- At least one leader that has free IP addresses is found in the responses. The node chooses the leader that has the largest free IP addresses to be the allocator and sends *allocatorChosen* message to that node. Upon receiving this message, the chosen leader calculates $A_{zi} = (D_{zi}\times G_{zi})^T$ matrix, removes the highest available IP address from its free IP address pool, as well as assigns it to the new node along with the netId, leader IP address, and i^{th} row of A_{zi} via *addressKeyAssign* message.
- None of the leaders in the responses has IP addresses. Node chooses one of the leaders randomly and sends a*llocatorChosen* message to the selected one. When the leader receives this message, it will send a *waitPeriod* message to the requestor and extend the *allocatorChosenTimer* timer. Next to that, the allocator starts the IP address search process in order to find a free IP address and then assigns it to the new node. It will ask t nodes to collaborate and send G_{zi} matrix rows. After receiving a threshold number of rows for G_{zi} from any threshold number of neighbors, the chosen leader calculates $A_{zi} = (D_{zi}\times G_{zi})^T$ matrix, removes the highest available IP address from its free IP address pool, and assigns it to the new node along with the netId, leader IP address, and i^{th} row of A_{zi} via *addressKeyAssign* message.
- If none of the replies is originated from a leader, it will pick a random neighbor node from the replies and select it as its allocator. This allocator will try to search for an IP address to this new node by extending the *allocatorChosenTimer* timer. It will then start searching for a leader that has free IP addresses by broadcasting the *findLeaderAllocator* message. Any leader receives the *findLeaderAllocator* message replies with the *findLeaderAllocatorReply* message that contains its IP address and number of free IP addresses. The relay selects the leader that has the biggest number of free addresses to be the remote allocator by sending the *leaderAllocatorChosen* message. When the leader receives this message, it will remove the highest available IP address from its free IP address pool and send it to the relay via *leaderAddressAssign* message. In this case, the normal nodes act as a relay between the requestor and remote allocator that eventually sends the IP address to the new node via *forwardIpAssign* message. When the *allocatorChosenTimer* timer expires, the node will check its status. If it finds that it is not configured, it will broadcast *addressKeyRequest* message and start over. If it is configured, it will ask any t neighbor nodes to collaborate to generate G_{zi} matrix. After receiving a threshold number of rows for G_{zi} from any t neighbors, the requestor calculates K_i = (i^{th} row of A_{zi})$\times G_{zi}$.

III. SIMULATION RESULTS

To evaluate the performance of the proposed tree-based assignment scheme, several simulation experiments were performed. Network Simulator ns-2 (version 2.33) was used for this task. Performance was evaluated by measuring the number of control messages per node during the simulation time and IP address and key assignments latency.

A. Simulation Setup

The protocol was tested under various conditions and distribution models. In these tests, the following parameters were used:

- Random waypoint mobility model.
- Network area is 1000 m×1000 m.
- Nodes move with a maximum speed of 5 meters/second.
- The routing protocol used was the ad hoc on demand distance vector (AODV).
- Transmission range of the node is 100 m.
- Data link layer was IEEE 802.11 for all the nodes.
- The number of nodes in the network is 1000 node including 20 leaders
- The network has 20 zones at most where no more than 50 nodes available in each zone.

The proposed protocol was tested and compared with the other well-known protocols (T-DAAP for address assignment, and threshold cryptography protocol for key assignment) in terms of not only the number of packets needed for address and key assignments, but also the address and key assignments delay.

B. Address and Key Assignments Packets

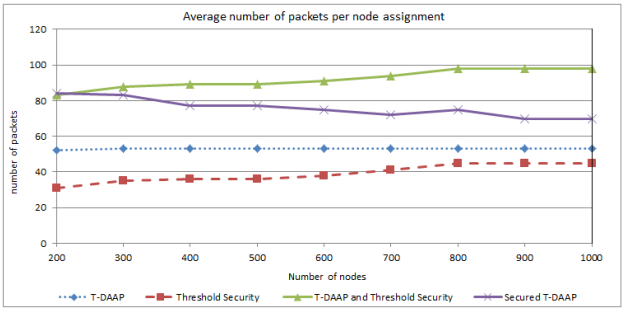


Fig. 1 Average number of control packets/node assignments

Fig. 1 depicts the average number of packets per node for address and key assignments. We can see that the T-DAAP needs almost a fixed number of packets regardless of the growth of the network nodes while the threshold cryptography protocol requires an increasing number of packets as the network size grows larger due to the slight increase in the threshold value to recover *G*. Our proposed protocol clearly shows a reduction in the number of packets needed per node for IP address and security key assignments. As the network grows in size and when the number of nodes reaches the threshold value, our proposed protocol outperforms the other separate protocols as far as the number of required packets is concerned. In fact, our proposed protocol shows an enhancement of about 30% when the network nodes are 1000. On the other hand, when having 200 nodes (network threshold value), the proposed protocol requires almost the same number of packets as required in the separate protocols (T-DAAP and threshold cryptography protocol).

C. Address and Key Assignments Delay

Fig. 2 shows the delays required for address and security key assignments to the node using reference protocols and our proposed protocol. As far as the threshold cryptography protocol delay in concerned, we can see that it tends to settle down when the number of network nodes is dramatically increased. The T-DAAP delay increases as long as the number of nodes increases due to a larger broadcast domain. Compared with the two protocols needed to perform the IP address and security key assignments, it is clearly noticed that the secured T-DAAP shows a great reduction in the assignments delay. When the number of network nodes gets larger and larger, this delay can be up to 26% less than that when employing two separate protocols.

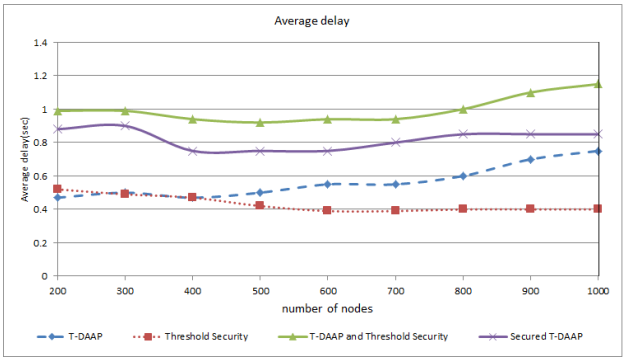


Fig. 2 Address and key assignments average delay

IV. Conclusion And Future Work

Secured T-DAAP overlaps address and key assignments for MANET nodes. Simulation results are very promising and show impressive reductions in the packet delay and average number of packets per node needed to perform the address and key assignments over the other known separate protocols (T-DAAP and threshold cryptography). As a summary, we obtain the improvements of about 26% and 30% for the packet delay and required packets for address and security key assignments, respectively. However, there are a number of potential future research directions for this work. It will be extremely interesting to modify the proposed secured T-DAAP protocol to accommodate two different networks along with two different roots. On the other hand, studying the effect of threshold value variation on the protocol behavior will be a valuable addition.

References

[1] B. Wu, J. Wu , and M. Cardei, "A Survey of Key Management in Mobile Ad Hoc Networks," *Handbook of Research on Wireless Security*, Y. Zhang, J. Zheng, and M. Ma (eds.), Idea Group Inc., ISBN: 978-1-59904-899-4, 2008.

[2] J. Sheu, S. Tu, and L. Chan, "A Distributed IP Address Assignment Scheme for Ad Hoc Networks," *Proceedings of the 11th International Conference on Parallel and*

Distributed Systems 2005 Proceedings (ICPADS'05), Fukuoka, Japan, pp. 439- 445, July 2005.

[3] X. Li, Y. Deng, V. Narasimhan, A. Nayak, and I. Stojmenovic , "Localized address autoconfiguration in wireless ad hoc networks," *Proceedings of 2010 International Conference on Wireless Communications and Signal Processing (WCSP'10),* Suzhou, China, pp.1-6, October 2010.

[4] N.I.C. Wangi, R.V. Prasad, M. Jacobsson, and I. Niemegeers, "Address autoconfiguration in wireless ad hoc networks: protocols and techniques," *Proceedings of IEEE Wireless Communication*, vol.15, no.1, pp.70-80, February 2008.

[5] L. Zhou, Z. Haas, "Securing adhoc networks, " *IEEE Network Magazine*, vol.13, no.6, pp. 24-30, November 1999.

[6] M. F. Al-Mistarihi M. Shurman, and A. Qdaimat, "Tree based dynamic address autoconfiguration in mobile ad hoc networks," *Computer Networks,* vol. 55, no. 8, pp. 1894-1908, February 2011.

[7] M. R. Thoppian and R. Prakash, "A Distributed Protocol for Dynamic Address Assignment in Mobile Ad-Hoc Networks," *IEEE Transactions on Mobile Computing*, vol. 5, no.1, pp. 4-16, January 2006.

[8] K. Weniger, M. Zitterbart, "Address Autoconfiguration in Mobile Ad-Hoc Networks: Current Approaches and Future Directions," *IEEE Network Magazine (Special issue on Ad-Hoc Networking)*, vol. 18, no. 4, pp. 6-11, July 2004.

[9] S. Nesargi and R. Prakash, "MANETconf: Configuration of Hosts in a Mobile Ad Hoc Network," *Proceedings of the 21st Annual Joint Conference of IEEE Conference on Computer Communication (INFOCOM'02)*, New York, NY, USA, pp. 206-216, June 2002.

[10] H. Zhou, L. Ni, and M. Mutka, "Prophet Address Allocation for Large Scale MANETs," *Proceedings of the 22nd Annual Joint Conference of IEEE Conference on Computer Communication (INFOCOM'03)*, San Francisco, CA, pp. 1304-1311, April 2003.

[11] Y. Hsu and C. Tseng, "Prime DHCP: A Prime Numbering Address Allocation Mechanism for MANETs," *IEEE Communications Letters*, vol. 9, no. 8, pp. 712-714, August 2005.

[12] M. Taghiloo, J. Taghiloo, M. Dehghan, "A Survey of Secure Address Autoconfiguration in MANET", *Proceedings of the 10th IEEE Singapore International Conference on Communication Systems (ICCS2006)*, Singapore, pp. 1-5, March 2006

[13] T. Matsumoto, H. Imai, "On the Key Predistribution System: A Practical Solution to the Key Distribution Problem," *Proceedings of the Conference on the Theory and Applications of Cryptographic Techniques on Advances in Cryptology (CRYPTO '87)*, Santa Barbara, California, USA, Lecture Notes in Computer Science 293 Springer 1988, pp.185-193, 1988.

[14] R. Blom, "An Optimal class of symmetric key generation systems," *Proceedings of the EUROCRYPT84 workshop on Advances in cryptology: theory and application of cryptographic techniques*, Paris, France, pp. 335-338, April 1984.

[15] I. S. Reed, G. Solomon, "Polynomial codes over certain finite fields," *Journal of the Society for Industrial and Applied Mathematics (SIAM),* vol.8, no.2, pp. 300-304, June 1960.

[16] K.-H. Lee, S.-B. Han, H.-S. Suh, C.-S. Hwang, and S. Lee, "Authentication Protocol Using Threshold Certification in Hierarchical-cluster-based Ad Hoc Networks," *Journal of Information Science and Engineering*, vol. 23, no. 2, pp. 539-567, March 2007.

MAC Layer Back-off Algorithm for Ad hoc Networks

Mohammad M. Shurman[1], Mamoun F. Al-Mistarihi[2], Zaid A. Alomari[3]

[1] Jordan University of Science and Technology/ Network Engineering and Security Department, Irbid, Jordan

[2] Jordan University of Science and Technology/ Electrical Engineering Department, Irbid, Jordan

[3] Jordan University of Science and Technology/ Computer Engineering Department, Irbid, Jordan

alshurman@just.edu.jo, mistarihi@just.edu.jo, zaid.alomari@gmail.com

Abstract - **Channel efficiency and channel throughput in any network are affected by packet collisions. Reduction of packet collisions plays a greater role in Mobile Ad hoc Networks (MANETs), due to the open medium topology where carrier sensing is not practical or feasible. Consequently, these collision issues introduce a key problem in MANET networks where researches proposed back-off mechanism to overcome packet collisions. Back-off algorithm is proposed to reduce collisions that happen when more than one node tries sending data on the channel simultaneously. The Binary Exponential Back-off (BEB), the first Back-off algorithm that is deployed in Medium Access Control layer, is used by IEEE 802.11 to avoid collision in MANETs. BEB algorithm uses a uniform random distribution Back-off values to solve the collision problem. In this paper, we propose *Fibonacci Liner Increment Back-Off (FLB)* algorithm to achieve an effective value for the Back-off timer of the nodes before accessing the channel to avoid collision. We carried on simulations using ns-2 to compare the performance of the proposed protocol with existing protocols like Fibonacci Increment Back-off (FIB), Binary Exponential Back-off (BEB), Pessimistic Linear-Exponential Back-off (PLEB) in terms of packet delivery ratio (PDR), average End-End delay and normalized routing load. Simulation results demonstrated that our FLB algorithm outperforms existing protocols and achieves the best results in routing packets, normalized routing load and End-End delay.**

Keywords: MANETs, Packet collisions, back-off algorithm, BEB algorithm, FIB algorithm, PLEB algorithm, FLB algorithm.

I. INTRODUCTION

Wireless networks spread rapidly these days; they become a core network for many applications. In infrastructured networks, the nodes communicate with a central point BS (base station) to send and receive the data. Infrastructureless network or mobile ad hoc networks (MANET) consists of mobile nodes, these mobile nodes move dramatically and communicate with each other without the hassle of configuring the nodes and setup the network [1]. Nodes function in MANET differs from infrastructured network nodes due to the lack of central point to control the network operations; also the mobility of nodes forces the network topology to change frequently. Consequently, nodes in MANET operate as a client, a server, and a router at the same time, MANET nodes communicate with each other with the help of multi-hop routing, where intermediate nodes relay data until it reaches the destination.

These features allow incorporating MANET in many applications, such as in military battlefield to collect information about the soldiers, vehicles, and enemy's movement in the war time for strategic planning. Also it is used in disaster recovery operations where the existence of current infrastructure or setting up new infrastructure is impossible as in earthquakes, floods, and fires [2].

The open medium of wireless networks allows any node to send data without coordination, because of that, collision will occur which degrade the network throughput. The Media Access Control (MAC) [3] is a data communication protocol sub-layer of the Data Link Layer specified in the OSI model (layer 2). It provides addressing and channel access control mechanisms allowing several terminals or network nodes to share the medium. Authors of [4] proposed medium access collision optimization protocol using ready to send/clear to send (RTS/CTS) to reduce packet collisions in wireless transmissions. When a node wants to transmit its data, it sends RTS to the destination, which in turn will reply with CTS using the broadcast technique. At this time, the nodes that receive CTS will wait until the time of the CTS finished. Only nodes that sense the RTS packet can transmit, while the nodes that sense the CTS will wait until the medium becomes free [5].

Many researchers studied the waiting time that the nodes should wait before starting to resend when the collision happen. Researchers proposed back-off algorithms to avoid the collision between nodes. Binary Exponential Back-off (BEB) algorithm was first proposed. Until now, many researches focused on how to optimize the Back-off algorithm to utilize the medium efficiently by giving nodes the optimal waiting time value.

The rest of this paper is organized as follows. Section II describes back-off time related works. Section III presents problem statement. Section IV explains the proposed scheme, while section V presents experiments and simulation results. Finally, section VI concludes the paper.

II. RELATED WORK

In this section we present an overview of previous work about Back-off time the nodes should wait to overcome collisions. Logarithmic back-off algorithm that uses logarithmic equation to find the contention window (CW) is introduced in [6]. Authors of [7] use Fibonacci Increment Back-off Algorithm (FIB). The Pessimistic Back-off algorithm (PLEB) that combines two functions is proposed in [9].

Many researchers studied the Back-off algorithm and tried to improve its functionality to reach the optimal Back-off waiting time. When the node wants to transmit data, firstly it senses the medium status to start transmission or wait, i.e., if the medium is not busy with any transmission process, the node will start sending the data. Otherwise, the node will wait a period of time (Back-off time) before sending. This time was first proposed by BEB algorithm, which increases exponentially. The Back-off time is computed as:

$$BO = [Rand()Mod(CW)] \times slot\ time$$

where BO is the Back-off time that the node should wait before next sending, Rand () is a random number generated between 0 and the length of the Contention Window (CW), CW is the contention window, and Slot time: the activity of the medium is sensed by the node at every basic unit of timing called slot time.

Each time slot, the node should sense the medium. If the medium is idle, the BO will be decreased based on following equation:

$$BO = BO_{old} - slot\ time$$

where BO is the new Back-off time value the node will wait before next sending, and BO_{old} is the old back-off time value the node was waiting before next sending.

In logarithmic Back-off algorithm [6], improvement on BEB algorithm was done by using a logarithmic function instead of exponential. The Back-off time the node will wait is calculated as:

$$BO = \log(BO_{old}) \times BO_{old} \times slot\ time$$

This model depends on the network nodes number and speed. The results showed that Log function has more efficiency and higher throughput in large networks than BEB, because when the network is large, the collision will increase, thus, we need a rapid and large CW at the beginning to allow enough waiting time for the node before sending its data to prevent collision. The Log function has the ability to increase its values rapidly.

In FIB, researches focuses on reducing the time increase of CW. When the CW time increases in the Back-off algorithm, the performance of the network will be decreased and wastes the channel resources. In [7], Fibonacci series was used to manage the incremental time after each collision. Fibonacci series equations are as follow:

$$F(0) = 0$$
$$F(1) = 1$$
$$F(n) = F(n-1) + F(n-2), \text{ for } n \geq 2$$
$$BO_{n+1} = F(n) \times \text{slot time}$$

This function will increase the CW size by small factor instead of using exponential function. Mathematical researches that studied this function, found that if we sum two values sequentially in this series and divide the result by the larger of the two numbers, the average will be 1.6 [8]. The series looks like this:

0, 1, 2, 3, 5, 8, 13 …
(2+3)/3 = 1.6
(5+8)/8 = 1.6, and so on

If the node finds the channel busy, the waiting time will increase by Fibonacci value. Simulation results demonstrated that the FIB algorithm has the largest throughput for different node speeds rates.

In [9], Pessimistic liner exponential Back-off (PLEB) algorithm is introduced, it is claimed to improves the back-off time by combining two arithmetic functions (liner and exponential). In this algorithm, authors added liner functionality to BEB algorithm to prevent the CW size from reaching large value, at the same time short period for CW is not needed.

If the network traffic is heavy, the collision may happen more frequently. In this situation, PLEB will give the node exponential waiting time. After the exponential reaches predefined number of Back-offs, the liner increment will start, preventing the CW from reaching large values. This algorithm is compared with LMILD (Linear/Multiplicative Increase and Linear Decrease) algorithm which is an improvement of MILD (Multiplicative Increase Linear Decrease) algorithm that takes into account the impact of the non-sending node channel state. In LMILD scheme [10], colliding nodes increase their contention windows multiplicatively, while other nodes overhearing the collisions increase their contention windows linearly. After successful transmissions, all nodes decrease their contention windows linearly.

A technique to adjust the value of contention window is proposed [11]. Many algorithms use fixed value of CW, while algorithm in [11] uses an adaptive algorithm to calculate the CW. Logarithm increase is used here to increase the waiting time for the nodes when the medium is busy. In case of idle channel and idle time reaches DIFS, the node can transmit and the Back-off window is reduced linearly. In [12], a new Back-off algorithm based on logarithmic nodes number to determine the CW is introduced. Simulation results showed that logarithmic function can outperform the BEB algorithm in throughput and fairness.

Algorithm concerns network traffic was proposed in [13]. If the transmission in the network is low, the CW will be short, while if there are a numerous transmissions from the nodes, the CW window will be large because more collisions will occur. In this algorithm if there is a collision, CW will increase exponentially. After the CW reach threshold value, the Back-off will increase linearly.

Exponential Increase Exponential Decrease (EIED) Back-off algorithm uses the technique to increase and decrease the Back-off time exponentially [14]. This algorithm assumes that when the collision happens during

transmission, the increase of Back-off time will be exponentially. In idle case, when no collision occur (success transmit), the node decreases its waiting time exponentially. The exponential increase and decrease has different factors affecting CW calculation.

Tahiry and Isabelle use their SBA algorithm to increase the fairness and efficiency of the network using two parameters (CWmax and CWmin) [15]. Results of simulation showed that SBA has more efficiency and fairness relative to other Back-off algorithms. Unified analytical model is established to analyze the performance of BEB and Quadratic Back-off (QB). Quadratic Backoff (QB) is proposed in [16], where the Back-off window size is quadratic increased upon collisions. Results showed that QB has better performance and throughput over BEB.

Quality of services (QoS) using Backoff algorithm is proposed in [17]. Researchers built their new Back-off algorithm to control the value of the Back-off time threshold and CW. E-BEB is the modified algorithm to the BEB, which focused on improving the fairness of the network. The algorithm increases the CW when the medium has a failed transmission; while decreases CW when a success transmission takes place. This algorithm provides a technique to fast decrease the value of CW when the medium is free, resulting in an increase of the network throughput. Simulation results showed that E-BEB overcomes the BEB and other algorithms in the fairness.

In EFB Back-off algorithm [18], using two functions, the behavior of a medium size network choosing incremental value is investigated. In [18], the authors did not identify the liner function and the value of the CW limit. Choosing any value for the CW or any liner function that increases the CW rapidly will not give any benefits. Choosing these values (liner function and threshold value) depends mainly on the network behavior, since network behavior may not be stable at the beginning, middle or at the end of network life time.

Adaptive algorithm to adjust the CW and Back-off time is introduced in [19]. This New Adaptively Varying Back-off (NAVB) algorithm is proposed to make the network more dynamic by having three levels of thresholds for CW size. These levels are low level, medium level and high level. The network will move between these levels based on the network collision, if there is a collision, the network will go from low to medium level or from medium to high level. For a successful transmission, the network will move from high level to medium or from medium to low level. Each level has a function to adjust the value of CW. The change from one level to another will not occur in one step, it depends on the decrement values in each level. Simulation results shows that NAVB outperform BEB, MILD and other algorithms in throughput and fairness.

III. Problem Statement

Network nodes experiencing collisions on the shared channel need to Back-off for a random period of time, which is uniformly selected from the Contention Window (CW). This contention window is dynamically controlled by the Binary Exponential Back-off (BEB) algorithm. Increasing contention window size and decreasing the possibility of granting access to channel are the most of Back-off algorithm drawbacks. The large window size leads to waste of network resources by increasing the channel idle times. Thus, BEB scheme suffers from a fairness problem and low throughput under high traffic load. In this paper, to achieve the best Back-off time, we use the Fibonacci function to increase the value of the contention window (CW) with a liner function to limit the increment value of the Fibonacci.

IV. Methodology

In this paper, a new Back-off algorithm named Fibonacci Liner Increment Back-Off (FLB) is proposed, in which a Fibonacci selection of the Back-off value based on the old Back-off value will be used. Our proposed algorithm as shown in Fig. 1, selects the Fibonacci Back-off values to start node waiting time before granting access to the channel. We limit the increment of Fibonacci values by using the upper limit "N" value to prevent the Fibonacci from reaching a large value. Threshold value "N" is investigated for the optimal value that will enhance the proposed algorithm results. The value of N used in this paper is 63. This value is chosen based on different experiments to select a best value. Different researches focused on how to get the range of the best contention window. As mentioned in [20], the optimal minimum contention window (CW) is 31 and the maximum value is 1023. So when the Fibonacci reach the value of N, another function will start to work (liner function).

Fibonacci/Liner Increment Back-Off Algorithm (FLB)

```
Step 0:  Set BO to initial value
Step 1:  While BO ≠0 do
For each time slot
     If channel is idle then BO=BO -1
Step 2: Wait for a period of IDFS then Send
If (send Failure) then
     if ( number of Back-offs <= N)
          then
       BO=Fib(n)
   else
BO = BO + T
          Go to step 1
Stop
```

Fig. 1 FLB Algorithm

The FLB algorithm is starting by using BO values depends on the FIB series. After the FIB reaches the value of N which is 63, the liner function "T" will start to prevent the FIB from reaching large values.

The value of "T" is a small constant integer {1, 2, 3….10}, in our experiment a value of 6 is used and a small size, high traffic network environment is chosen at the beginning. A liner function is used to add a constant value to the BO to get an increment of constant and small value.

V. Simulation Results and Discussion

NS2 simulator is used to test the proposed FLB algorithm, where we run the BEB code based on our simulation environment to find the initial results of the BEB

code to use it as a reference for our proposed approach. The simulation is carried on using different scenarios for the same number of nodes. For example, we use different scenarios to simulate the network having 20, 40, 60, 80 and 100 nodes.

A. Simulation Parameters

The parameters that we have used in our experiment are shown in Table 1.

Four important performance metrics we are looking to analyze for the proposed FLB algorithm:

- *Average End-to-End delay* of data packets; the time it takes to transfer a packet from the sender to the receiver. This includes all the delays in between.
- *Packet Delivery Ratio*; the ratio of the number of packets correctly received to the number of packets sent by the application layer
- *Normalized routing load average*; the ratio between routing packets and received ones.
- Average *dropped* packets

Table 1. Experiment Parameters

Parameter	Value
Area	500m×500m
Nodes Number	20, 40, 60, 80, and 100
Simulation Time	500
Packet Generation Rate	4
Speed	2 m/s , 5 m/s , and10 m/s
Pause Time	0

B. Simulation Results

1. End-End Delay

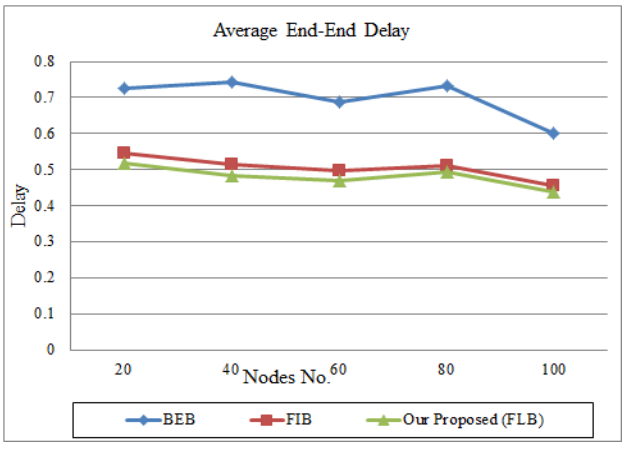


Fig. 2 End to End delay average for all algorithms at speed 10m/s

In Fig. 2, we see that FIB and FLB algorithms have more efficient behavior in end-end delay compared with that of BEB. Comparing FIB and FLB algorithms, we can also see that the proposed FLB algorithm has better results in terms of the average end-end delay than FIB.

Fig. 3 shows the end-end delay verses speed. From the figure, we can see that the proposed algorithm out performs all other algorithms at different speeds. The delay is decreasing when the nodes speed is increasing. We can conclude that end-end delay can be optimized using our proposed FLB algorithm.

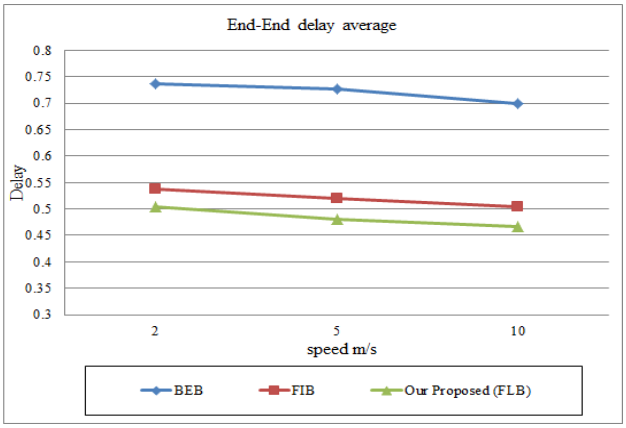


Fig. 3 End-End delay for all algorithms at different speeds

2. Packet Delivery Ratio

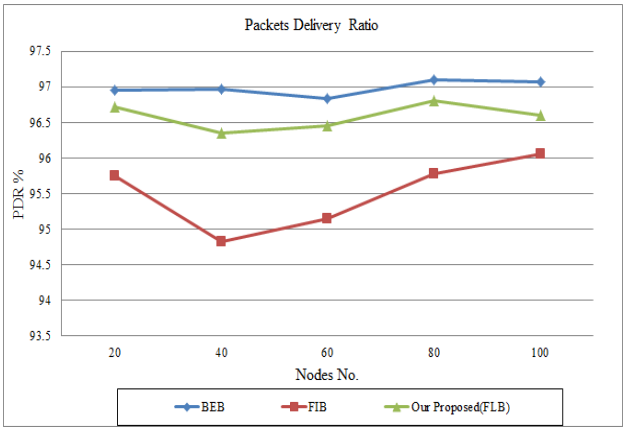


Fig. 4 Packets Delivery Ratio for all algorithms at speed 10m/s

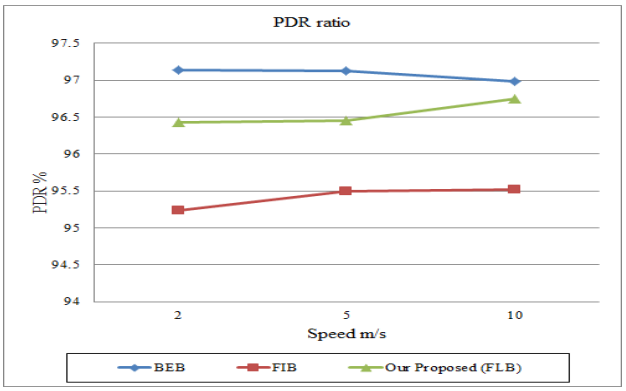


Fig. 5 PDR ratio for all algorithms at different speeds

Fig. 4 shows the packet delivery ratio verses number of nodes at speed 10m/s. The figure shows that BEB al-

gorithm has the best results, while our proposed protocol incurs little delay in the network due to higher channel efficiency over BIB. If we compare the FLB proposed algorithm to FIB algorithm, we can see that FLB perform better than FIB in terms of the packet delivery ratio.

Fig. 5 shows the packets delivery ratio verses speed for a network having 100 nodes. From the figure, BEB performs the best among the other protocols; FLB outperforms FIB but incurs little more delay over BEB.

3. Normalized Routing Packets

One of the most important factors that should be investigated is the routing packets factor. This factor always treated of low importance, but in fact, it is a main factor which controls the packet overflow through the network. When we reduce the number of routing packets, we will increase the chance for the packet to be delivered since probability of the channel to be idle will be higher. Fig. 6 shows the number of routing packets verses the number of nodes at speed 10m/s during the simulation. Note that; FLB algorithm generates packets little higher than the FIB algorithm, but BEB algorithm has the smallest values among them. This can be misleading since it might mean the routing protocol cannot find paths to the destination and fails to deliver packets to their designated destination. Normalized routing packets give a better overview about correctness of routing protocol to find the path from source to destination and packet delivery, the higher the normalized routing packets means more packets are delivered to their destinations.

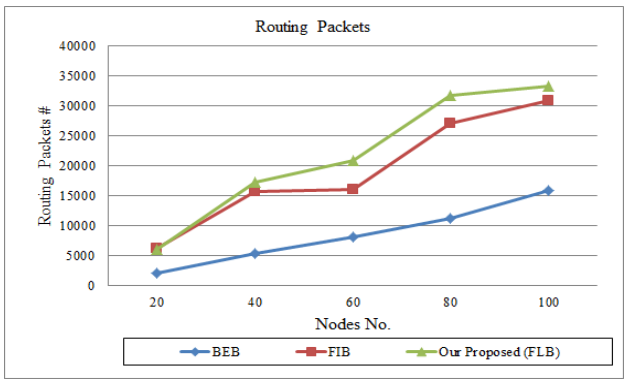


Fig. 6 Routing packets for all algorithms at speed 10m/s

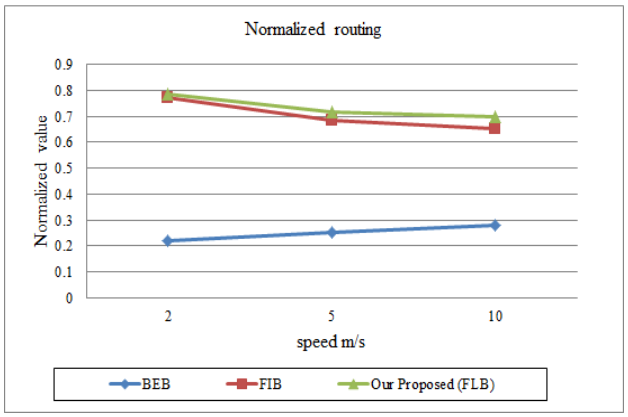


Fig. 7 Normalized routing packets at different speeds

Fig. 7 shows the ratio between the routing packets and the received packets (normalized routing) for all algorithms at different speeds. The proposed FLB algorithm generates the highest values of routing packets which means the protocols found a route for the packet to reach the destination at different speeds.

4. Average Dropped Packets

Another metric to be tested is the average number of packets dropped at different speed, where we measure the average drop packets at speed values 2, 5 and 10 m/s. Figure 8 illustrate how the three algorithms behave. From the figure, we can see that the BEB algorithm has the best results, since it has the minimum number of dropped packets. FIB and the proposed FLB algorithms have very close values in average. The reason that BEB has the minimum average drop packets is that the number of routing packets is less than FIB and FLB.

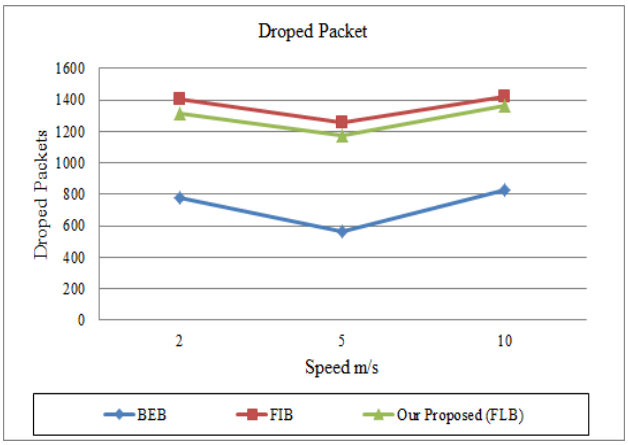


Fig. 8 Dropped packets at different speed for all algorithms

VI. Conclusion

Back-off time is the main approach nowadays to avoid collisions in MANET. In this paper, the Fibonacci Liner Increment (FLB), which uses Fibonacci increment to tolerate the gap between different values of Back-off time, is proposed. Simulation results reveal better results for our proposed algorithm (FLB) compared to BEB and FIB algorithms in normalized routing load and End-End delay. The FLB proposed algorithm has the best End-End delay over BEB and PLEB, it also show better normalized routing packets but add a little delay on the packet route to the destination.

References

[1] Kalepu, S. Mehra and C. Yu, "Experiment and Evaluation of a Mobile Ad Hoc Network with AODV Routing protocol," Mobile Computing Research Lab., ECE Dept., Cleveland State University, October 2002.

[2] Jun-Zhao Sun, "Mobile ad hoc networking: an essential technology for pervasive computing," In Proceedings of ICII International Conference, pp. 316-321, 2001.

[3] http://en.wikipedia.org/wiki/Media_access_control .

[4] N. Sze-Yao, T. Yu-Chee, C. Yuh-Shyan and S. Jang-Ping, "The broadcast storm problem in a mobile ad hoc network," In Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking, pp.151-162, August 1999.

[5] V. Bharghavan, A. Demers, S. Shenker and L. Zhang, "MACAW:

A Media Access Protocol for Wireless LAN's," In Proceedings of the ACM SIGCOMM Conference, pp. 212–225, August 1994.
[6] S. Manaseer and M. Ould-kauoa,"A New Backoff Algorithm for MAC Protocol in MANETs," 21st Annual UK Performance Engineering Workshop, pp. 159-164, 2005.
[7] S. Manaseer, M. Ould-Kha and L. M Mackenzie, "Fibonacci Backoff Algorithm for Mobile Ad Hoc Networks," DCS Technical Report series, Department of Computing Science, University of Glasgow, pp. 1-6, 2006.
[8] Falbo and Clement, "The Golden Ratio—A Contrary Viewpoint," College Mathematics Journal, v36, n2 pp. 123-134, March 2005.
[9] S. Manaseer and M. Masadeh. "Pessimistic Backoff for Mobile Ad hoc Networks," Al-Zaytoonah University, the International Conference on Information Technology (ICIT'09), Jordan, 2009.
[10] J. Deng, P-K. Varshney, and Z-J. Hass "A new backoff algorithm for the IEEE 802.11 distributed coordination function," CNDS '04, In Proceedings of Communication Networks and Distributed Systems Modeling and Simulation, San Diego, CA, USA, January 18-21, 2004.
[11] Y. Zhi, Y. Min and P. Yali, "Logarithm increase-linear decrease Backoff algorithm based on self-adaptive optimal contention window," In Proceedings of ICCASM 2010, pp. 657-661, October 2010.
[12] F. Hong-yu, X. Lei and L. Xiao-hui, "Logarithmic Backoff Algorithm of MAC Protocol in AdHoc Networks," In Proceedings of E-Business and E-Government (ICEE) International Conference, pp. 1695-1698, 2010.
[13] C. Han-Chieh and C. Hsiao-Hwa, "Slow Start Backoff Algorithm for Ad-Hoc Wireless Networks," In Proceedings of GLOBECOM 2010, pp.1-5, December 2010.
[14] S. Nai-Oak, K. Byung-Jae, S. Jabin and M.E. Miller," Enhancement of IEEE 802.11 distributed coordination function with exponential increase exponential decrease Backoff algorithm," In Proceedings of the 57th IEEE VTC, pp. 2775-2778, April 2003.
[15] L. Fratta et al."SBA: A Simple Backoff Algorithm for Wireless Ad Hoc Networks," In Proceedings of IFIP International Federation for Information, LNCS 5550, pp. 416–428, 2009.
[16] X. Sun and L. Dai, "A Comparative Study of Quadratic Backoff and Binary Exponential Backoff in IEEE 802.11 DCF Networks," In Proceedings of the 45th CISS Annual Conference, pp. 1–6, April 2011.
[17] M. Al-Hubaishi, T. Abdullah, R. Alsaqour and A. Berqia, "E-BEB Algorithm to Improve Quality of Service on Wireless Ad-Hoc Networks," Research Journal of Applied Sciences, Engineering and Technology, pp. 807-812, April 2012.
[18] M. Bani Yassein, O. Al Oqaily, G. Min, W. Mardini, Y. Khamayseh and M. Manaseer, "Enhanced Fibonacci Backoff Algorithm for Mobile Ad-Hoc Network", IEEE International Conference on Computer and Information Technology (CIT 2010), pp. 749-754, 2010.
[19] M.Abdul Gafur et al, "Enhancing the Performance of Adhoc Networks Using New Adaptively Varying Backoff Technique," International Journal of Engineering Science and Technology, vol. 12, pp. 7078-7086, 2010.
[20] A. Ksentini, A. Nafaa, A. Gueroui and M. Naimi, "Determinist contention window algorithm for IEEE802.11," In Proceedings IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), pp. 2712-2716, Vol. 4, 2005.

Heterogeneous Networks in theory and practice

Tomislav Ivanković
Ericsson Nikola Tesla d.d
Krapinska 45, Zagreb, Croatia
Phone:+(385) (0)1 3653489; email:tomislav.t.ivankovic@ericsson.com

Abstract - Expansion of using SmartPhones last few years changed all data usage forecasts and it seems that only macro-cellular network layer will not be sufficient in meeting users request. Definition for HeterogeneousNetwork(HetNet) according to Wikipedia is : 'Mobile Experts defines the Heterogeneous Network as a network with complex interoperation between macrocell, small cell, and in some cases WiFi network elements used together to provide a mosaic of coverage, with handoff capability between network elements.' HetNet is an attractive means of expanding mobile network capacity. A HetNet is typically composed of multiple radio access technologies, architectures, transmission solutions, and radio base stations(RBS) of varying transmission power. The aim of this paper is to explain main aspects of increasing mobile networks capacity using HetNet. It will be explained design aspects of HetNet deployment also expected capacity gains and main concerns and problems which might occure while introducing HetNet in existing mobile networks such as signal imbalance between macro-cellular layer and HetNet.

Key words: HetNet, SmallCell, imbalance.

I. INTRODUCTION

Enhancing, densifying HSPA/LTE macro-cellular network and HetNet deployment are methods of meeting increasing traffic demands and fulfilling users high expectations for mobile broadband. This article analyzes expansion strategies, covers important design choices for heterogeneous networks also presenting some field trials which are showing potential of HetNet in meeting high traffic demands for mobile broadband networks. Mobile broadband traffic has surpassed voice and is continuing to grow rapidly. This trend will continue in the future, with global traffic figures expected to double annually over the next five years. This traffic growth, driven by new services and terminal capabilities, is paralleled by user expectations for data rates similar to those of fixed broadband. Actual figures per subscriber can be different depending on geographical market, subscription type, terminal type and some users with mobile devices are already creating traffic in the order of gigabytes and predictions are estimated to be several GB per month for some devices and certain user behavior. The mobile industry is preparing for data rates in the order of tens of Mbps for indoor use as well as outside and gigabyte traffic volumes. Figure 1 shows method of meeting request for higher capacity:

Improve macro layer means to implement new software features, new antennas, more antennas etc.

Densify macro layer means to reduce site to site distance(add more sites) for example reduce site to site distance from 400m to 200m by adding new sites.

Add small cells to existing macro layer which means introduction of HetNet.

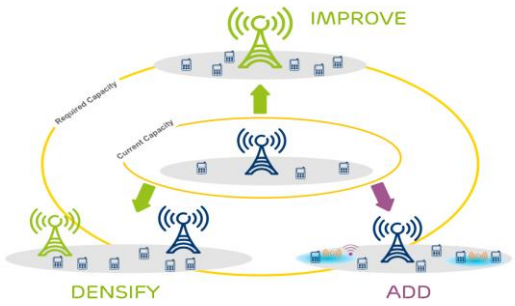


Figure 1 Different solutions to improve coverage and capacity through heterogeneous networks

II. IMPROVING, DENSIFYING AND COMPLEMENTING THE MACRO NETWORK

There are several methods that can be used to meet traffic and data rate demands. On a high level, the key options to expand network capacity include:

- Improving the macro layer- this can be done by adding a carrier, applying different optimization techniques or by introducing new features
- Densifying the macro layer- reducing the inter-site distance between macros by adding additional macro cells or by performing higher-order sectorization (HOS) in which a three-sector site becomes a six-sector site.
- Complementing the macro layer with low power nodes- creating a heterogeneous network.

These approaches are discussed here in more detail and use the example of a dense urban environment to show achievable performance. Upgrading the radio access (HSPA or LTE) of existing sites can give very high user data rates and improved system capacity, which can be further enhanced through the addition of more spectrum(10Mhz-20Mhz), more antennas, and advanced processing within and between nodes. Increasing capacity and data rates with this approach is attractive as it reduces need for new sites.

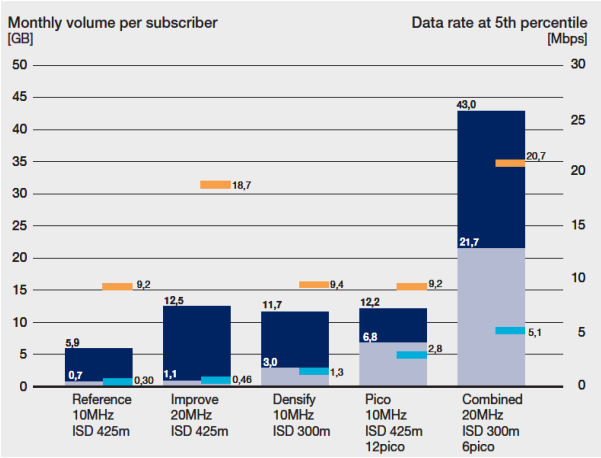


Figure 2 DL and UL monthly data volumes and data rates at 5th percentile(95 percent coverage) supported by the different expansion strategies

The reference system is a 10MHz HSPA system with an inter-site distance of 425m(Figure 2 ,Blue bars-DL monthly data volumes, Light blue bars-UL monthly data volumes, Blue markers-UL 5th percentile data rates, Orange markers-DL 5th percentile data rates), achieving monthly data volumes per subscriber of 5.9GB in DL and 0.7GB in UL. Increasing the spectrum to 20MHz, data volumes for the DL approximately double. Figure 2 shows data rates achievable at low load with 95 percent coverage probability. In the DL, data rates of tens of Mbps are achieved. In the UL, however, this data rate is significantly lower (a few 100kbps) and increasing spectrum does not improve the situation since there is no carrier aggregation in UL with current SW releases. Also this condition is because power limitation: data rates are limited by relatively low received power, which is due to attenuation between terminal and base station caused by a combination of distance and bad radio propagation (such as in indoor locations at the cell edge). At some point, achievable the capacity and/or data rates in the existing network with enhanced radio access will no longer be acceptable. If possible, densifying the macro network is a good solution in these cases. In dense urban areas, networks exist with inter-site distances down to 100-200m. Figure 2 shows that by doubling the number of macro sites, DL capacity is doubled. Despite of more sites DL capacity per site is more or less the same. UL capacity is more than doubled and users become less power limited – better capacity per site. A big increase in UL data rates is achieved. Complementing the macro networks with low power nodes, such as micro and pico base stations,. This approach offers very high capacity and data rates in areas covered by the low power nodes. Performance for users in the macro network should improve if low power nodes can serve a big number of hotspots and coverage holes.

Deploying low power nodes can be challenging it brings complexity in current network deployment, as performance depends to where traffic is generated. In addition, due to the reduced range of low power nodes, more of them are required. To manage all this new complexity requires proper and very detail design and integration of the low power nodes. Figure 2 shows results for the deployment of 12 pico base stations per macro site in traffic hotspots. This returns the same downlink(DL) capacity increase as the previous two approaches (more spectrum and densification). However, the biggest gain is achieved in the uplink(UL), which is a result of mitigating the power limitation. As result of this approach UL data rate improvement is greater than for the other two approaches. The way to meet future capacity demand is by combining all three approaches: improving the macro layer; densifying the macro layer; and adding pico nodes, as indicated by the last example in Figure 2. This three approaches should be combined on the existing network,. Such a heterogeneous network configuration, exploiting macro and low power nodes, can in principle support arbitrary data volumes and very high data rates.

III. DESIGN OPTIONS FOR HETEREGENIOUS NETWORKS

TABLE I RULES OF THUMB FOR LOW POWER NODE DEPLOYMENT

	Design choice Decision criteria	***Decision criteria***
1	Access Open access Closed subscriber group	Deployment conditions Operator deployed User deployed
2	Deployment Indoor deployment Outdoor deployment	Hotspot spread and position Large indoor hotspot Outdoor hotspot or many smaller indoor hotspots
3	Type of low power node RRU Conventional pico Relay	Backhaul availability Fiber (P2P or WDM PON) Copper / fiber / microwave No backhaul
4	Frequency reuse Reuse macro spectrum Separate spectrum	Capacity need and access Capacity is driver Closed subscriber group
5	Power and cell selection Power Biased cell selection	Hotspot area Cover the hotspot Cover the hotspot value varies significantly

Several technical inputs can bring us to effective design of heterogeneous networks. Demand perspective, traffic volumes, traffic geolocation and target data rates are the most important. For supply perspective the important aspects include radio environment, macro-cellular coverage, site availability, backhaul transmission, spectrum and integration with the existing macro network.

Commercial aspects as technology competition, business models, and marketing and pricing strategies must also be considered. To summarize, Table I includes rule of thumbs for some of the key design choices operators encounter.

Deployment aspects and choice of radio access technology how to best complement the macro network depends on the network scenario. HSPA or LTE operating within the licensed spectrum should be used if the base station is deployed in a public area, or if coverage is important. If the base station is well isolated from interference and range is not crucial – if it is used in a private home, for example – then WiFi exploiting unlicensed or license-exempt bands is an attractive solution. For authentication, simple sign-on, and access to mobile operator services the WiFi access point should be connected to the mobile core network. A 3GPP-based HeNB provides little gain over WiFi in such scenarios. On the contrary, HeNBs may create coverage holes or use spectrum that would otherwise be available for the macro layer. In the network-design process, it is important to consider the business model. While a single operator often manages outdoor macro-cellular networks in urban areas, indoor systems are often shared between operators (cf Distributed Antenna Systems). WiFi access points and similar smaller scale solutions are often user-deployed (by individuals, enterprises or a third party), where access can be open for all subscribers or available for certain users only (Closed Subscriber Group)[CSG]).

For public systems, particularly outdoors and in difficult radio environments, open access for all subscribers is important so that users connect to the best base station. This explains the first rule of thumb in Table I. Local traffic hotspots can cover a wide area, such as an entire block and include several buildings. In such cases, deploying an outdoor low power node that also covers indoor locations would be suitable. If the existing macro-cellular grid is too sparse to meet the traffic demand and provide adequate indoor service, deploying outdoor low power nodes is a useful technique to achieve general coverage improvement. When traffic is concentrated to one specific indoor location, such as a shopping mall, indoor deployment is preferable.

Backhaul transmission becomes more important as the number of nodes increase, in part because it will constitute a larger share of the total cost of ownership, but also as the availability of fixed backhaul affects the feasible placements, installation costs, and time needed for site acquisition and installation. A low power node can be connected just to the core network or to the core network and other base stations. Each connection, also called backhaul link, may have different bandwidth and latency characteristics. The capacity of the backhaul link not only affects user throughput, but also the overall radio access network performance as a high-capacity backhaul allows for tighter coordination between nodes. There are several types of low power nodes that put different requirements on backhaul support. For networks where the backhaul has low-latency and high-capacity characteristics, deploying remote radio units (RRUs) is the preferred approach; otherwise stand-alone pico base stations is an option.

An RRU has the potential to improve overall network performance through tight coordination between nodes and usually comprises antennas to send/ receive the radio transmission as well as a radio frequency processing unit. A central control unit that can collect baseband signals from several RRUs performs baseband signal processing and higher layer processing. The control unit and its distributed antennas/ RRUs must be directly connected via a low-latency and high-capacity interface. An optical fiber-based backhaul is suitable for RRU deployment and such solutions are increasingly being used in HSPA high-capacity networks. Where RRUs are not applicable, a stand-alone base station can instead be connected to the radio network controller (RNC) for HSPA and the core network for LTE. In contrast to RRUs, stand-alone pico base stations have loose backhaul requirements and may, therefore, fit with networks that have a high-latency and low-capacity interface. Two examples of cooperation schemes that can be applied to stand-alone picos are: soft handover in WCDMA Rel-99; and Inter-Cell Interference Coordination (ICIC) available in LTE release 8, which enabled simple interference management between base stations (picos and macros). Additionally, a relay or repeater may be employed to improve coverage. The relay needs to communicate with the macro-cellular donor base stations, either inband or outband. If spectrum is available out-of-band relaying – using one band for the access link between terminal and relay and a separate band for the backhaul link between relay and donor base station – is the preferred approach.

IV. IMBALANCE

Building the HetNet with „small cells“-cells that are deployed with less Common Pilot Channel(CPICH) power than the surrounding macro cells could have an imbalance between downlink and uplink in relation to the macro cell. Figure 3 interprets imbalance between macrocell and smallcell. Imbalance can lead to degradation in troughput and reduced capacity.

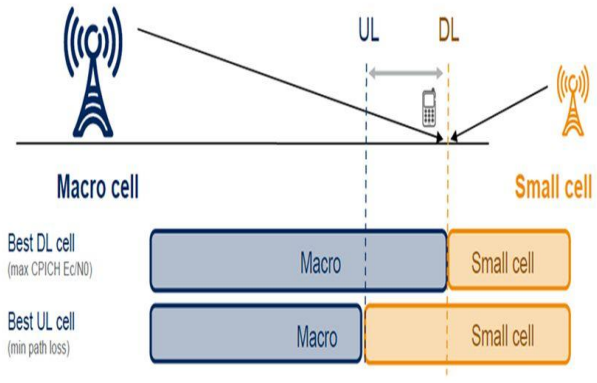


Figure 3 Link imbalance illustration

To calculate the imbalance between two cells, equation (1) is used. This equation consists of two parts. The first part is the difference in CPICH power between the two cells and is used to calculate how much the downlink cell border has moved towards the small cell. The second part is the difference in uplink noise figure and this is used to calculate where the uplink cell border is located. The imbalance is defined as the difference between downlink and uplink.

$$I=(P_{CPICH_macro}-P_{CPICH_smallcell})-(N_{f_small}-N_{f_macro}) \quad (1)$$

Where:

- P_{CPICH_macro} is the CPICH signal strength [dBm] at RBS antenna port (Explanation of RBS antenna port can be found in notes below. Special care needs to be taken if TMA is used).
- $P_{CPICH_smallcell}$ is the CPICH signal strength [dBm] at the RBS antenna port.
- N_{f_small} is the noise figure of the small cell RBS
- N_{f_macro} is the noise figure of the macro cell RBS

To reduce the imbalance it is necessary to change the CPICH setting in the affected cells. This will influence the first part of the imbalance equation (1). If the CPICH power in the small cell is increased, then the imbalance will be less. However, care must be taken when changing the CPICH power so that the proportion of CPICH power out of total cell power remains the same.

Increasing the CPICH power in the small cell is a way to mitigate the imbalance to make sure it is less than the recommended 6dB. However, when doing so it is important to ensure the CPICH to total RBS power ratio remains within 8-10%. A too low CPICH ratio will affect the performance in the cell. However, it should be noted that changing the CPICH power will also affect coverage of the cell as CPICH power is the definition of coverage.

V. CONCLUSION

Mobile-broadband traffic is increasing. In parallel, new applications are raising expectations for higher data rates in UL and DL. In this paper has been analyzed that with creating a heterogeneous networks by introducing low power nodes is an attractive approach to meeting traffic demands and performance expectations, particularly in situations where traffic is concentrated – in hotspots, or areas that cannot be suitably covered by the macro layer. Introducing HetNet by combining low power nodes with an improved and densified macro layer, very high traffic volumes and data rates can be supported. The nature of the existing network, as well as technical and economic considerations, will dictate which approach – improving the macro layer; densifying the macro layer; or adding pico nodes – or combination of approaches best meets volume and data-rate targets. Low power nodes give high data rates locally and also offer benefits to macro users by offloading and cooperating with the macro layer. It shows that tight integration of low power nodes with the macro network provides gains over the uncoordinated case through favorable combining of received signals and avoiding interference.

REFERENCES

[1] Next generation LTE, LTE-Advanced, S. Parkvall, A. Furuskär and E. Dahlman, Ericsson Review 2/2010.

[2] Mobile data traffic surpasses voice, http://www.ericsson.1.com/news/1396928.

[3] M. Almgren, L. Bergström, M. Frodigh, K. Wallstedt, Channel Allocation and Power Settings in a Cellular System with Macro and Micro Cells Using the Same Frequency Spectrum, Vehicular Technology Conference, 1996.

[4] Heterogeneous networks – increasing cellular capacity, F.Kronestedt,L.Falconetti,A.Furuskär,K.Johansson,S.Landström, Ericsson Review 1/2011.

Information systems infrastructure planning based on collected data from monitoring tools

Ivana Golub, Branko Radojević
Croatian Academic and Research Network (CARNet), Zagreb, Croatia
Ivana.Golub@CARNet.hr, Branko.Radojevic@CARNet.hr

Abstract - **Management of large computer systems and networking environments usually demand huge investments and lot of human resources. It is almost impossible to calculate minimum investments needed to maintain high level of availability and functionality. The gap between needed and acquired investments in equipment and human resources may lead to degraded performance and functionalities. The possibility to use collected data from monitoring tools and to combine them with inventory data of the same equipment can become an essential analysis tool for the system owners and/or maintainers. A model is presented in this paper, with which current status of an system can be determined based on the age and utilization of the system, as needed and sufficient factors. Presented approach can be used for complex systems, as well as for simple - single product systems.**

I. INTRODUCTION

Managing large computer system and/or large computer network is very complex, time consuming, and human resources intensive task. It was observed that the complexity is rising higher than linear with respect of the number of nodes added to the system.

While there are several types of management applications that can simplify above mentioned tasks, most of them that were analyzed focus on a set of tasks (performance, resource, system and network monitoring tools, risk analysis tools, inventory management tools, etc.) and lack the ability to easily predict infrastructure elements that will eventually fail or become unresponsive.

In this paper, a new model is proposed according to which there are two dimensions - age and load of individual elements within the system that could be used to determine current status of an information system infrastructure, or its selected part.

The ability to visualize and produce ready-to-publish diagrams that are easy to read and understand by management and decision makers is a clear win situation.

If continuous measurements and assessments were recorded through longer periods of time, there is additional analytic possibility to draw velocity vector and to analyze how fast a single device is moving from good state (quadrant) to a bad state (quadrant).

Those single devices can be grouped by model and/or manufacturer to get a general overview how they are doing in the production environment. This type of analytic data can be used to decide upon actions that should be taken in order to improve the performance of selected system or system parts, as well as to make future purchasing decisions.

II. EXPLORATION DETERMINATION MODEL

Current status of individual elements that form selected part or the whole information system infrastructure can be determined based on the information about the age of the element, as well as the information about the usage (current load) of an element. While age is easily determined based on the date of purchase of the element, usage characteristics depend on the type of element that is observed.

The usage characteristics of a server are determined by the processor power and memory used by the server, extendable to the amount of disk (e.g. in storage systems), number of concurrent users (e.g. web servers), etc. For networking equipment, usage is defined by the number (percentage) of used ports, as well as memory and CPU usage. It is up to the maintainer and/or owner of the system to define the exact parameters of a type of an element in order to describe its system on the best and relevant way. All such performance indicators are usually monitored through respective monitoring systems, so such information does exist in each monitored system. The information about the age of devices should exist in the inventory databases of the corporation that has bought equipment. This model, therefore, combines two usually separated database systems such as asset management (or inventory management systems) with device-usage monitoring system (such as network monitoring systems or server monitoring systems) and provides added-value information about the system.

The model that is then used to determine current status of the system infrastructure is presented at Figure 1. First it is necessary to define the scope of the analysis (1). Once defined, the scope should not be changed until the end of the analysis. Analysis could be repeated as many times as needed, on the same or different systems.

For the selected system it is necessary to perform categorization of devices that for the system (2). It is important because for different component different metrics (parameters) should be used in order to exactly define usage of the component. It is very important to use

all the relevant parameters, and define respective measures for such parameters.

According to the monitoring systems, data should be gathered for each element, and for each recognized parameter. Those values should be expressed in percentage on the scale of the minimum (0%) to the maximum availability (100%). With all respective parameters presented on the same scale it will be easy to perform mathematical operations on those values.

management system about the current age of a system element, gathered data are graphically represented in a two-dimensions graph. For each element only one dot will exist in the graph, with the sum of respective percentages on the x-axis and the age of the element on y-axis.

Positioning of all elements in the graph will show critical point of the system, either in the case of high load of an element, or in the case of old equipment.

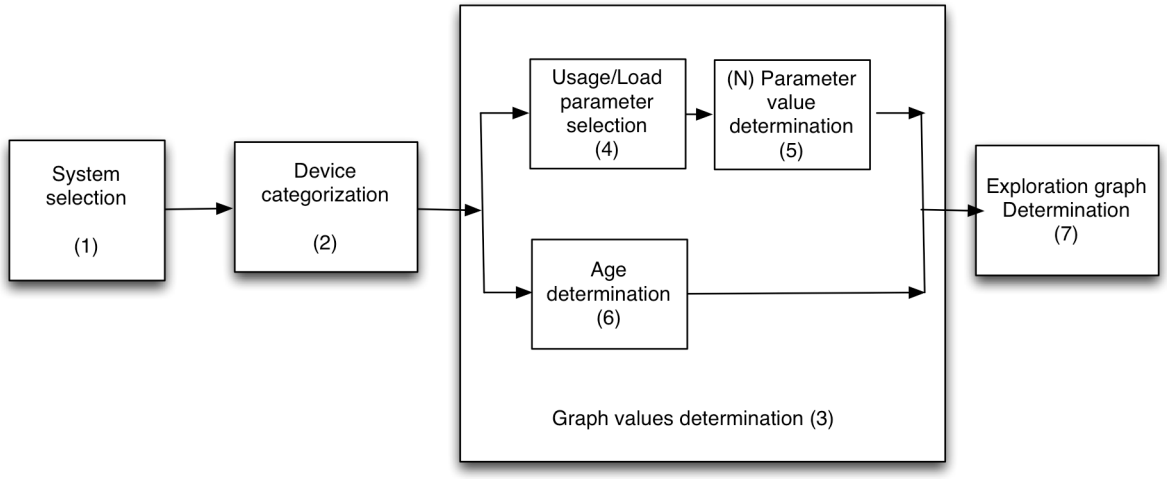


Figure 1: Exploration Determination Model

It is important to notice that different parameters might be selected for different types of the equipment, i.e. for different elements, because of the different nature of those elements, and different role that they have in the system. However, sine all those elements form examined system, it is important to look at and take into consideration all relevant components.

Together with parameter and respective values determination, it is necessary to determine the age (6) of each selected component. Age of an element is determined as present year minus the year of the purchase of that element. It will be presented at y-axis of the exploration graph. Although those data exist in inventory managements systems of an organization, it is usually not easy to integrate with the information from a monitoring system, neither are data from a network monitoring system from an organization integrated with the data from an computer monitoring system.

However, combination of those data is source of the valuable information for the owner(s) and maintainers of such systems, as presented in this document.

Based on the measured data from the monitoring system as well as on the data from the inventory

III. TEST CASE RESULTS

As the proof of the concept, a part of a data center is examined. Selected system consists of 27 components: 20 servers, 4 switches and 2 routers. Group of servers contains 4 categories of servers, bought in four different phases in the period from year 2007 until 2012. Switches have two categories, two of them bought in 2008, and two of them in 2010. Routers are also of different kinds and time periods.

Observed parameters for the servers are average CPU load, and average memory usage for a period of last six months. Server usage can be defined in several different ways - maximum load, instead of average, shorter or longer duration. However, for examined servers selected parameters are considered based on the system baseline.

For networking equipment - switches and routers - relevant parameters are considered to be usage of networking ports (used number of ports from the overall number of networking ports), CPU load (on the scale from 0 - 100%), and used RAM memory (also in the percentage ration from the maximum amount of memory).

For both - servers and networking equipment, age is determined as a difference between current year (2013), and the year of the purchase. Having in mind generally accepted limit of five years for the equipment amortization, it is easy to determine which elements could be considered "young", and of which elements are above the 5-year limit.

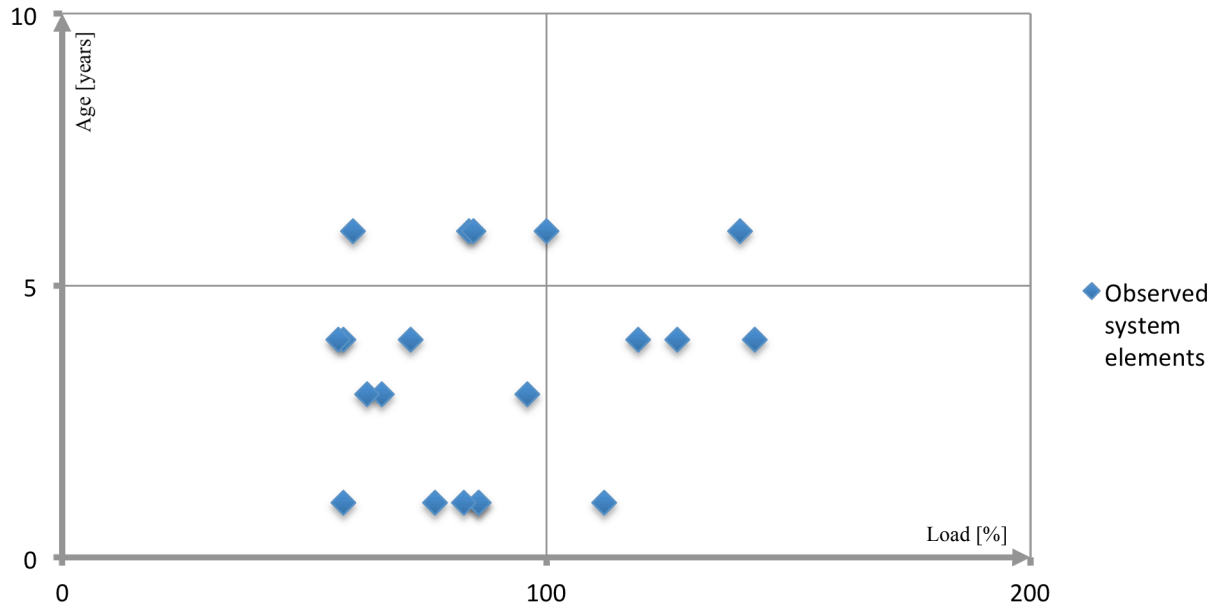


Figure 2: Exploration graph for the servers from the test case System A

Exploration graph on the Figure 2 represent results for the servers from the test case system A. two thresholds are also presented at the figure as two lines, one parallel with the x-axis at the level of 5-years and the other parallel with the y-axis at the level of 100%, which is exactly half of the whole possible interval.

Figure 3 provides similar results for the networking equipment, with two threshold-lines at 5 years and at 50% of the total value for axis-x.

The level of the threshold for y-axis is taken based on the general acceptance of the amortization level of 5 years. If an organization uses different model for amortization period, the threshold should be adjusted

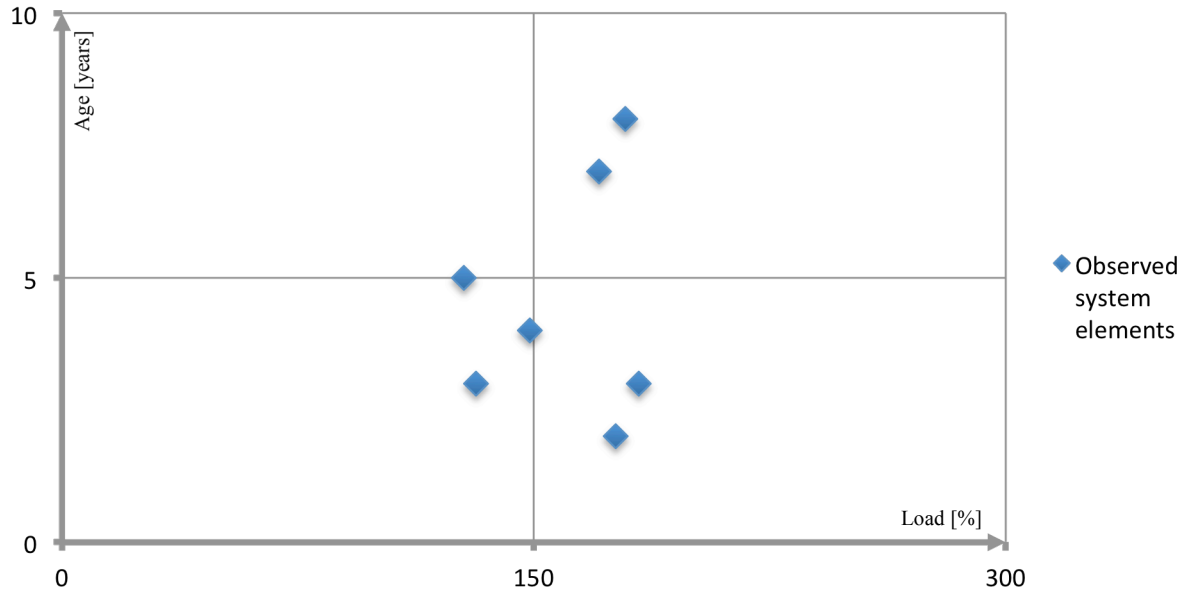


Figure 3: Exploration graph for the networking equipment from the test case System A

accordingly.

Threshold for axis x is defined as a value above which is a parameter considered to have increased value. Each time when a parameter has increased value, it should urge for an action. Such action can be from just informing about the increased value, or alarming and urging for reaction. Resulting actions based on such alarm can also be different, and depend solely on the persons responsible for the system - maintainers, administrators and/or owners of the system. Depending on the number of examined parameters, the threshold line is set at the N*p level, where N is the number of parameters, and p is the defined level above which parameter is considered to have increased value.

From Figure 2 it can be seen that most servers fall into lower-left quadrant, which represents low-usage and young equipment. This is desired position for all equipment, since it means that such equipment should have support from the manufacturer (as "young equipment"), and also that it can support additional burden.

Two servers are positioned into upper-left field which represents older equipment and low-usage. Since the threshold level is set relatively low (5 years), that might mean from one side that the system should last for some time more, however that it is a time to start acquiring or at least plan acquisition of new equipment. As the time passes by, the manufacturers of the equipment lower their support level and new version of software (with newer features) might not be applicable or supported for this type of hardware, so it would be wise to consider replacement for such equipment. The importance of replacement (or at least insurance of redundant parts) is more suggested as the system contains more important data and services.

Lower-right field of the graph represent equipment that is "young", but loaded more that wanted. The more the values are near the upper limit (N*100%), the bigger risk it might impose to partial or total malfunction of a service, or system in the whole. In such cases, it could be wise to revise established topology, additionally examine the source of increased load, or possible interaction between system components.

Positioning of elements in upper-right field is the most critical position, since this signals old equipment with extensive load. Possible risks for such equipment is lack of support from the manufacturer, lack of spare replacement parts (if the equipment is too old, and it was not thought on time), with a little instability in the system or additional load (planned or unplanned) the performance of the system might be raised even more and lead to the instability of the system, as well as partial or total malfunctioning.

However, in the presented scenario, there is only one server in this field, and one on the border between two upper fields.

The situation is a little different for the networking equipment. In this scenario, there are less networking equipment than servers. However, only one element is in the lower-left field, while all the others lie either in the area of older devices, in the area of more-loaded devices or both, with two elements at the borders, but still not in the lower-left field.

All the other analysis and risks presented for the servers part applies also for the networking equipment. The risks that older equipment might not be supported, that it might not be able to adjust to needed changes (e.g. with newer version of the operating systems and available hardware), as well as that more-loaded equipment might be a cause of the limited functioning, or total malfunctioning.

Comparing therefore the state of the servers and networking equipment, it can be seen that bigger point of risk for this system represent used networking equipment than servers.

It is important to outline that this analysis is based on the performance parameters from the monitoring systems. However, all other possible causes of malfunctioning are not examined here and can still be the reason for the unavailability of services and/or network. Such examples, not related to the analysis above are physical damage or malfunction of any of the components, security intrusions that might cause other failures (apart from performance violations), etc.

IV. Related Work

Network management models and techniques are well defined and documented in the literature (such as [1] and a set of RFCs that define SNMP protocol and it subcomponents), and there are a lot of network monitoring systems [2], [3], [4], [6], that are providing one or more of the listed functions, as well as represent specific solutions for a selected system. All of them explain either what to monitor or how to monitor. Several recent papers suggest distributed infrastructure for gathering data for the monitoring purposes [7], [8], However, a tools how to use existing data from monitoring systems for planning of future development of those systems are very hard to find. There are suggestions that it would be useful to use data from information systems and combine economic and mathematical models to introduce new models. The research presented in this paper has the same premises, although with the implementation in different area [5].

V. Conclusion

Presented method combines information about the age of an information system infrastructure element from an asset/inventory management system and the information about the usage of the same element(s) gathered from network/server monitoring systems, in order to determine

the state of firstly of the element itself and then the entire system.

Such information could be used for many different uses including (and not limited to): risk assessments, infrastructure planning, equipment acquisition planning, as well as further examination and analysis of possible and suggested actions that could and should be done in order to keep system in the lower-risks area.

By positioning of the system components in such areas allows the owners and system maintainers to think months in advance and plan proactively instead of reactively.

Moreover this method allows technical personnel and/or middle management to visually, easily and clearly demonstrate to the upper management current status of the entire system and the need and risks of not going into the investment in new equipment on time.

By doing this on a repetitive basis (once a month, quarter or year) it is also possible to get an overview of the observed system through the time and the how fast its crucial components are moving from good status, to acceptable and eventually to unacceptable status. This can become a very strong tool for a long-term systems analysis and in the end investment justification procedure.

LITERATURE

[1] ITU-T Recommendations M.3400, "TMN management functions", Telecommunications management network, M series: TELECOMMUNICATION MANAGEMENT, INCLUDING TMN AND NETWORK MAINTENANCE, 2000-02-04

[2] P. Barlet-Ros, G. Iannaccone, J. Sanjuàs-Cuxart, J. Solé-Pareta, "Robust network monitoring in the presence of non-cooperative traffic queries", Computer Networks, Volume 53, Issue 3, Pages 310-321, 2009,

[3] B. X. Lu, "An Architecture for Web Based and Distributed Telecommunication Network Management System", 2009 Third International Symposium on Intelligent Information Technology Application, :November 2009, pp. 152-155

[4] R. Pugazendi, K. Duraiswamy, "Mobile Agents - A Solution for Network Monitoring", 2009 International Conference on Advances in Recent Technologies in Communication and Computing, October 2009, pp. 579-584

[5] Y. Beres, D. Pym, S. Shin, "Decision Support for Systems Security Investment", Manuscript, HP Labs , 2010

[6] G. Fuxiang, L. Lantao, L. Miao, "Design and Implementation of Network Resource Management System Based on SNMP", 2009 Ninth International Conference on Hybrid Intelligent Systems , August 2009, pp. 341-344

[7] R. Boutaba, J. Xiao, "Network Management: State of the Art", available (Feb 2013) at: http://ecs.victoria.ac.nz/twiki/pub/Courses/NWEN402_2012T1/AdditionalReading/overview_network_management.pdf

[8] J. Famaey, S. Latre, J. Strassner, F. De Turck, "A Hierarchical Approach to Autonomic Network Management", available (Feb 2013) at: https://biblio.ugent.be/input/download?func=downloadFile&recordOId=990767&fileOId=990776

REGULATORY ASPECTS

Regulatory aspects of mobile broadband technologies

Almedin Džebo, dipl.inž.saob. i kom.; M.Sc. Harun Mutapčić, dipl.inž.saob. i kom.
Communications Regulatory Agency, Sarajevo, Bosnia and Herzegovina
adzebo@rak.ba; hmutapcic@rak.ba

Abstract – Global demand for mobile data services is rapidly growing. Numerous forecasts indicate exponential data traffic growth as mobile data usage has more than doubled each of the last five years. Over this past year, the need for additional spectrum has become particularly urgent, resulting in a number of new initiatives by industry and government. This is, consequently, driving the need for continued innovations in wireless and mobile data technologies towards mobile broadband, and new approaches in spectrum management and regulation. Regulators always face challenges when new technologies are introduced. The legislation is always a little bit late. The regulators need to either define how they apply the existing laws to new technologies and businesses either to quickliy adopt and apply new rules. In this paper we will consider regulatory aspects of major mobile broadband technologies such as LTE and LTE-Advanced, Mobile WiMAX and WLAN.

Key words: LTE, WiMAX, WLAN

I. INTRODUCTION

We are witnessing a very high growth of demand for mobile data services. Nearly everyone who uses a mobile phone or device anywhere in the world is demanding faster access to more multimedia data. They are demanding all applications they use on their computers to be availible on their mobile devices. From the operators perspective that demands more bandwidth. The network capacity overload and the potential crisis for mobile communications are guided by two primary factors: limited spectrum and growing data demands. Over this past year, the need for additional spectrum has become particulary urgent, resulting in a number of new iniatives by industry and government. Therefore is constant need for innovation in wireless and mobile technologies in order to provide more capacity and higher Quality of Service (QoS).

Users need fast and convenient data services trough their terminals, and operators need all-IP networks for high speed data conectivity. Competition between fixed line and mobile operators has also driven the evolution of radio technologies. With the knowledge that spectrum is a limited resource, the industries is undertaking the folowing initiatives to leverage all availble spectrum:

- Incerasing the spectral efficiency of technologies to continualy increase the bit per second of data bandwidth for every availble Hertz;
- Adapting specifications to enable opertion of UMTS-HSPA and LTE/LTE Advanced in all availble bands;
- Designing both FDD and TDD versions of techology to allow operation in both paired and unpaired bands;
- Designing carrier aggregation techniques in HSPA+ and LTE Advanced that band together multiple channels to improve both peak data rates and efficiency;
- Deploying as many new cells as is feasible.

In this process very important role has national regulator wich has to give support for developing and implementing new technologies and create environment for successfull market.
The biggest challenge in delivery of broadband services directly to customers terminals represent the “last mile” problem. It is solved with a number of wired or wireless access technologies, depending of the population density, required services and coverage area, existing infrastructure, configuration of terrain, etc.

II. STANDARDISATION OF TECHNOLOGY

Standardisation is very important factor in process of developing new technologies. The 3rd Generation Partnership Project (3GPP) unites telecommunications standard development organizations (ARIB, ATIS, CCSA, ETSI, TTA,

TTC) and provides their members with a stable environment to produce the highly successful Reports and Specifications that define 3GPP technologies. 3GPP technologies evolved from GSM-EDGE, over UMTS-HSPA-HSPA+, to now initial LTE/LTE Advanced deployments, in order to provide increased broadband capacity and user experience (Table 1.). The following terms are also used to describe networks using the 3G specifications: UTRAN, UMTS (in Europe) and FOMA (in Japan). LTE is one of the most promising wireless-technology platforms for the future. The version being deployed today is just the beginning of a series of innovations that will increase performance, efficiency, and capabilities [1].

Table 1. Comparison of data rate and channel bandwidth for wireless standards

Air Interface	Peak data rate	Channel Bandwidth	Type
GSM	9.6-14.4 kbps	200 kHz	TDMA
GPRS	171 kbps	200 kHz	TDMA
EDGE	474 kbps	200 kHz	TDMA
W-CDMA	2 Mbps	5 MHz	CDMA
HSDPA	10 Mbps	5 MHz	CDMA
HSDPA+ (64QAM&2x2 MIMO)	42 Mbps	5 MHz	CDMA
LTE	100 Mbps	20 MHz	OFDMA
LTE 2x2 MIMO	172.8 Mbps	20 MHz	OFDMA
LTE 4x4 MIMO	326.4 Mbps	20 MHz	OFDMA
IMT-Advanced targets	1 Gbps	100 MHz	OFDMA

In addition to 3GPP organization also exists the IEEE association engaged in standardization. The IEEE Standards Association, a globaly recognized standard-setting body within IEEE, develops consensus standards through an open process that engages industry and brings together a broad stakeholder community. IEEE standards set specifications and best practices based on current scientific and tecnological knowledge. IEEE 802 LAN/MAN Standards Comittee has developed three main standardization branches:

- IEEE 802.16 and IEEE 802.20 wireless MAN (Metropolian Area Network) standards will support high-rate broadband wireless access services to buildings, mostly through rooftop anntenas, from central base stations;
- IEEE 802.11 wireless LAN (Local Area Network) standards support users reoming within homes, office buildings, campuses, hotels, airports, etc.;
- And IEEE 802.15 wireless PAN (Personal Area Network) standards will support short-range links amonf computers, mobile phones and other consumer electronics devices that are worn or carried.

IEEE 802.11 and 802.15 have worked particulary closley since they both address unlicensed bands. The purpose of IEEE 802.11 standard is to provide wireless conectivity for fixed, portable, and moving stations within a local area. This standard also offers to regulatory bodies a means of standardizing access to one or more frequncy bands for the purpose of local area communication. Regarding the technical novelties delivered by the new 802.11-2012 we can emphasize several of them such as, new support for 3.65 and 3.7 GHz band, to aviod clashing with 2.4 GHz or 5 GHz networks, as well as better support for direct linking, faster cellular hand-offs, in-car network, roaming and mash networking. The newest IEEE standard in WLAN category is 802.11n. It was designed to improve on 802.11g in the amount of bandwidth supported by utilizing multiple wireless signals and anntenas (MIMO technology) instead of one. 802.11n shoud support data rates of over 100 MBps and also offers somewhat better range over earlier WLAN standards due to its increased signal intesity. This standars is one of main IEEE standards for mobile broadband.

The second main IEEE mobile broadband standard is IEEE 802.16, which defines the wireless metropolitan area network (MAN) tecnology which is breanded as a WiMAX. IEEE.16 has historically dealt with licensed bands and been more independent. The 802.16 includes two sets of standards, 8002.16-2004 (802.16d) for fixed WiMAX, 802.16-2005 (802.16e) for 3G mobile WiMAX and newest IEEE 802.16m-2011 for 4G Mobile WiMAX. The WiMAX wireless broadband access standard provide missing link for the “last mile” connection in MAN networks where DSL, Cable and other broadband access methods are not available or too expensive. The IEEE 802.20 has main technical caracteristics like a vehicular mobility up to 250km/h and larger cells up to 15 km. However, on June 8, 2006, the IEEE-SA Standards Board directed that all activities of the 802.20 Working Group be temporarily suspended until October 1, 2006. The decision came from complaints of a lack of transparency. Later, IEEE 802.20 standard was put to hibernation on March 2011 due to lack of activity [2].

The International Mobile Telecommunications – Advanced (IMT-Advanced) systems, which represent ITU standard for 4G, are mobile systems that include the new capabilities of IMT that go beyond those of well known IMT-2000. Such systems provide access to a wide range of telecommunications services including advanced

mobile services, supported by mobile and fixed networks, which are increasingly packet-based and moves the broadband Internet truly to mobile. Depending on the services for which the technology (or technologies, such: LTE, LTE-Advanced, 802.11n, 802.16m) will be used, continuous radio coverage may not be needed in order to meet the service requirements.

III. RADIOCOMMUNICATION BANDS

To deploy mobile broadband technologies appropriate Radio Frequency bands are needed. These bands need to accommodate new technologies and make it available to end users. New spectrum planned for mobile broadband is spread across additional five bands in portions of the 400 to 700 MHz and 2.3, 2.5 to 3.5 GHz bands. 4G systems and their backward compatibility to 3G will force multiple band and multiple front end products to meet the diverse requirments of regional carriers.

Table 2. Frequncy bands identified for mobile broadband

Frequncy band	Bandwidth (MHz)
450 – 470 MHz	20
790 – 862 MHZ	72
2300 – 2400 MHz	100
2500 – 2690 MHz	190
3400 – 3600 MHz	200

The bands below 1 GHz are cost-effective way to provide IMT services in sparsely populated regions in developed and undeveloped countries. The bands above 1 GHz are preferable for providing continuous blocks of spectrum for future broadband wireless systems such as IMT-Advanced (4G). Among the bands being proposed, the newly identified 3.4 to 3.6 GHz band could prove to be the most attractive for implementing 4G in the future.

The Mobile WiMAX and LTE/LTE Advanced are well positioned to drive the global evolution towards pervasive mobile broadband internet communications with market acceptence, rich ecosystems and promising economies of scale. The operators who succeed in 4G deployment and delivery will be those that can bulid and augment sufficient spectrum holdings in both lower and upper frequncy bands. As a consequence vendors must evolve equipment to flexibly support multi-mode and multi-band funcionality so that service providers can present a view of the spectrum combinations that will characterize the emerging 4G regulation landscape [2].

There are considered a combination of the following bands:

- 800 MHz and 2.5 GHz;
- 800 MHz, 2.5 GHz and tecnology flexibility in 2G bands;
- 900 MHz, 2.3 GHz and 3.5 GHz;
- 800 MHz and 1800 MHz;
- 2.5 GHz and 3.5 GHz;
- 900 MHz, 2.3 GHz and 3.5 GHz.

In a accordance with the spectrum allocation formats, spectrum reform requires planning for the allocation and reassignment of spectrum across multiple bands. This is irrespective of license renewal dates, new licensing timeframes or the availability of new spectrum for mobile services.

A detailed spectrum reform roadmap will recognize the interdependencies between different spectrum bands and increase the level of certainty for operators in understanding how much spectrum is prospectively available, under what conditions, and in which bands.

IV. REGULATORY APPROACH

Regulators have always been interested in the development of the telecommunications industry and infrastructure. In the past, the main role of the telecommunications regulation was in protecting customers from the monopoly power of vertically integrated operators. Deregulation in the 90s changed the industry and the role of the regulators. The technological convergence on the other hand is bringing together the telecom, broadcasting, and information services regulation. The main rationales behind regulation are: effective use of resources, competitive markets, customer rights, preventing abuses like monopolies and cartels, redistribution of wealth.

The regulators have a so called narrow window to guide business models – they can't make any drastic decisions which could affect the industry overnight. New business models may arise either accidentally or deliberately. In the worst case regulation may seriously hinder business activities, if planned negligently. Regulators face always challenges when new technologies are introduced. The legislation is always a little bit late. The regulators need to define how they apply the existing laws to new technologies and businesses [3].

EU does not impose specific legal provisions on mobile broadband technologies. It establishes a regulatory framework and defines the tasks of National Regulatory Agencies (NRA). The telecommunications regulatory framework recognizes that much of telecommunications regulation exists as a means of addressing potential and actual abuses of

market power. With that in mind, the EU attempts a comprehensive, technology neutral approach to regulation.

The European Commission defines a series of relevant telecommunications markets, and provides a set of guidelines for determining the presence or absence of market power. Within each market the NRA determines whether one or more parties possess Significant Market Power (SMP). If SMP exists, the NRA will impose appropriate obligations. Basically EU seeks to move completely away from technology-specific and service-specific legislation.

Many countries have not considered the legal status of WLAN Networks because WLAN has not yet become sufficiently common. Some countries have dealt with the matter only from the viewpoint of frequencies and licenses. Most countries that have considered the matter further, share the opinion Finland has taken. For example, Spain, Italy, Switzerland, Turkey, Hungary and Estonia considered that service offered by a café or a hotel to its customers is not public telecommunications. On the other hand, there may be obligations imposed on service providers on the basis of other laws. For example in Italy, service providers must authenticate users on the basis of the anti-terrorism law [2].

Before regulators start with spectrum reform or the allocation of new 4G spectrum sources suitable for mobile services, a granular understanding of the status quo is critical. Where a registry of spectrum distribution and use is not in place this action becomes an urgent priority. Such a spectrum audit must take account of spectrum that is used or reserved for public authorities along with resources allocated for commercial purposes other than communications.

Globally, provisioned network capacity to handle the explosive growth of broadband data traffic is forecasted to increase by forty times between 2010 and 2015 [4]. As operators evolve towards meeting increasing capacity demands and less are coverage limited, narrow band allocations in lower frequency ranges will attract a lower market premium price that has been the case in the past. This is not to imply that lower frequencies are no longer valuable. However, their optimized use will be in combination with higher band frequencies such as within 2.6 GHz range, where significant bandwidth resources are available. There is a discernible trend of regulators moving to combined spectrum allocation models.

Within Europe, discussions on the digital dividend (band 790 – 862 MHz) started in 2006 with the adoption of a first RSPG (Radio Spectrum Policy Group) opinion and a first Mandate from EC to CEPT in early 2007. In response, CEPT identified the upper part of the UHF band as the preferred band for the purpose of a mobile allocation as part of the digital dividend.

After WRC-07 (World Radiocommunication Conference 2007) decision, in April 2008, the EC issued a second Mandate to CEPT on the technical considerations regarding "harmonization options for the digital dividend in the European Union". On the basis of the CEPT response, the EC adopted the following [5]:

- European Commission Recommendation 2009/848/EC on "Facilitating the release of the digital dividend in the European Union", in October 2009",
- Commission Decision 2010/267/EU20 on "harmonized technical conditions of use in the 790-862 MHz frequency bands for terrestrial systems capable of providing electronic communications services in the European Union", in May 2010.

The technical considerations regarding harmonization options for the digital dividend in the European Union are described in CEPT reports. Four CEPT reports have been adopted, see Table 2.

Table 3. CEPT reports related to digital dividend and new mobile broadband

CEPT Report	Title
CEPT Report 29, 26 June 2009	Guideline on cross border coordination issues between mobile services in one country and broadcasting services in another country
CEPT Report 30, 30 October 2009	The identification on common and minimal (least restrictive) technical conditions for 790 – 862 MHz for the digital dividend in the European Union
CEPT Report 31, 30 October 2009	Frequency channeling arrangements for the 790 – 862 MHz band
CEPT Report 32, 30 October 2009	Recommendation on the best approach to ensure the continuation of existing Program Making and Special Events (PMSE) services operating in the UHF (470 – 862 MHz), including the assessment of the advantage of an EU-level approach

CEPT Report 31 in particular concludes that the preferred frequency arrangement for the frequency range 790 to 862 MHz should be based on the FDD mode in order to facilitate cross-border coordination with broadcasting services, noting that such an arrangement would not discriminate in favour of or against any currently envisaged technology.

The decision of WRC-12 to allocate the band 694-790 MHz to the mobile, except aeronautical mobile, service in Region 1 opens the way for the worldwide harmonization of both the

700, 800 MHz and 900 MHz bands for IMT. Although it should be recognized that the current use of the 700 MHz by broadcasting may prevent its availability to the mobile service in many countries in the years to come, this band will however become available soon for the latter service in many countries, which have lesser requirements for broadcasting and will therefore benefit from worldwide harmonization of the frequency band plans for IMT [6].

However, making a new allocation to mobile does not necessarily mean that the band will become available. Regulators will face major policy decisions on the relative priority of mobile and broadcasting 4G services. Even if the decisions come down in favor of mobile, it will need the concerted effort of a number of countries within a region to bring about substantial change. Where there are existing broadcasting services – as throughout Europe – these would need to be re-arranged to clear the spectrum. This would require difficult negotiations to be held with neighboring countries. And there will be significant cross-border constraints imposed by those countries that keep broadcasting and other services in the band (including aeronautical radars in some places).

Furthermore, successful conclusions were reached on over 30 separate items on the latest WRC-12 (World Radiocommunication Conference 2012), even if in some cases that conclusion was not to change the Radio Regulations. The resulting compromise is a very unusual arrangement. The new allocation, by way of a WRC Resolution, becomes effective from the end of the next WRC (scheduled for late 2015) and will be subject to technical and regulatory conditions to be developed in the intervening period. So it remains to be seen for how long the apparent success of the WRC remains only on paper.

V. CONCLUSION

On the end, we can conclude that, the spectrum reform is a priority globally. This is driven in large part by the burgeoning of mobile broadband and the progression towards 4G (Mobile WiMAX and LTE/LTE-Advanced). Greater regulatory certainty around band re-planning and the structure of new spectrum allocations is demanded as operators seek a firm basis upon which to assess future bandwidth requirements and how they can be met. Rearrangement of current bands allocated to mobile services and the release of unallocated spectrum will present different challenges and degrees of challenges in different nations. However, at the very least, all nations need to move to auditing spectrum distribution and use across multiple bands and public and private sector players. This view then provides the basis for band re-planning as deemed necessary and the development of approaches to the allocation of liberated or new spectrum. Regulators have an urgent demand to develop a multi-band plan-irrespective of existing license expiry dates and the timing of availability of new spectrum earmarked for mobile services.

Regulators are confronted with a daunting set of legacy issues that need to be addressed, along with priorities for 4G spectrum reform priorities and the need to set critical conditions related to new spectrum allocations. As 4G spectrum agendas are developed, these often overlapping issues, priorities and policy approaches must all be considered.

VI. REFERENCES

[1] Savo Glisic, “Advanced Wireless Networks: 4G Technology”, John Wiley & Sons Ltd, England, 2006.

[2] Internationl Telecommunication Union (ITU) – Distance Learning Course on "Mobile Broadband: LTE/LTE-Advanced, WiMAX and WLAN" held from 20 November – 17 December 2012 (http://academy.itu.int/moodle/course/view.php?id=555)

[3] Heikki Hämmäinen, Business models for wireless internet access, Helsinki University of Technology Department of Electrical and Communications Engineering Networking Laboratory, 2006 (http://www.netlab.tkk.fi/opetus/s383042/2006/S38_3042_Seminar_report_final.pdf)

[4] UMTS Forum, Report 44 – Mobile traffic forecasts 2010-2020 report, January 2011. (http://www.umts-forum.org/component/option,com_docman/task,doc_download/gid,2537/Itemid,213/)

[5] ITU, Digital Dividend: insights for spectrum decisions, ITU – Telecommunication Development Sector, August 2012 (http://www.itu.int/ITU-D/tech/digital_broadcasting/Reports/DigitalDividend.pdf)

[6] ITU, World Radiocommunication Conference 2012, Final acts – Agenda and References (Resolutions and Recommendations) (http://www.itu.int/dms_pub/itu-r/oth/0C/04/R0C040000070001PDFE.pdf)

Role of Infrastructure Atlas in NGA Deployment

Suada Hadžović[1], Saša Kovačević[2]
[1, 2] Communications Regulatory Agency BiH, Mehmeda Spahe 1, 71 000 Sarajevo, Bosnia and Herzegovina
shadzovic@rak.ba

***Abstract*— EU Initiative to Reduce the Cost of Rolling out High Speed Communication Infrastructure discusses different methods of reducing costs. In this paper, we present main aspects of centralized atlas of passive infrastructure as one of regulatory measures on reducing the cost of deploying Next Generation Access Network (NGA).**

Keywords—Next Generation Access Network (NGA), infrastructure atlas, cost reduction, broadband network, regulation.

I. INTRODUCTION

Despite the progress which was made, the EU is still suffering from underinvestment in the deployment of high-speed broadband networks. The key underlying cause is unnecessarily high civil engineering costs, which can make up to 80% of total costs in this field and could have been reduce by up to one quarter simply by cross-utility re-use of existing infrastructure [1].

The European Commission has started a work on exploring ways to achieve such costs savings and in April 2012 announced Public Consultation on an EU Initiative to Reduce the Cost of Rolling out High Speed Communication Infrastructure [2]. In this EU Initiative the European Commission discusses different methods of reducing costs by:

- Making a better use of existing infrastructure;
- Enhancing transparency and coordination of civil engineering works;
- Handling requests to roll-out networks in a more efficient way;
- Ensuring "ready for NGA access" buildings.

Recently, after comprehensive policy review, in December 2012 the European Commission adopted seven new priorities for the digital economy and society and updated the original 2010 Digital Agenda for Europe (DAE). Top digital priority for 2013 is therefore creating a new and stable broadband regulatory environment. A package of ten actions in 2013 will include among others, and mechanisms for reducing the civil engineering costs of broadband roll-out. Full implementation of this updated Digital Agenda would increase European GDP by 5%, or 1500€ per person, over the next eight years [3].

The digital economy is growing at seven times the rate of the rest of the economy, but unfortunately, a patchy pan-European policy framework currently holds this potential back. In October 2012, the Commission proposed a second set of actions (Single Market Act II) to further develop the Single Market. One of Single Market Act II key actions is adopting common rules to reduce cost and increase efficiency in the deployment of high-speed broadband (legislative proposal have to be presented in Q1 2013) [1].

These actions are targeted to close the gap between the amount that operators plan to invest in network infrastructure and investment required for achieving the DAE broadband targets.

This article is structured as follows. After this introduction, Section II gives NGA definition. Section III provides an overview of main aspects of infrastructure atlas, Section IV deals with infrastructure atlas impact on the local NGA deployment, while in section IV short conclusion is drawn.

II. NGA DEFINITION

There are many different definitions of Next Generation Access (NGA) networks. According to the European Commission NGA Recommendation, NGA networks means wired access networks which consist wholly or in part of optical elements and which are capable of delivering broadband access services with enhanced characteristics (such as higher throughput) as compared to those provided over already existing copper networks [4].

At the current stage of market and technological development, NGA networks are:

- fibre-based access networks (FTTx),
- advanced upgraded cable networks and
- certain advanced wireless access networks capable of delivering reliable high-speeds per subscriber.

Therefore, NGA networks are understood to have at least the following characteristics [5]:

- deliver services reliably at a very high speed per subscriber through optical (or equivalent technology) backhaul sufficiently close to user premises to guarantee the actual delivery of the very high speed;
- support a variety of advanced digital services including converged all-IP services, and
- have substantially higher upload speeds (compared to basic broadband networks- i.e. networks up to ADSL2+ networks, non-enhanced cable DOCSIS 2.0, mobile networks UMTS and satellite systems).

NGA networks present a number of regulatory, administrative and financial challenges which require the concerted effort of public and private stakeholders. The fact is that levels of investment required to deploy new access lines is a bottleneck for new entrants, but a number of community projects around Europe are today demonstrating that if the cost of passive infrastructure is lowered to the right threshold, the

private sector, even the smaller new entrants, will invest in NGA networks.

III. INFRASTRUCTURE ATLAS

Infrastructure atlas or map could range from being a database that contains information on which infrastructure operators are active in what region to a map that details the exact route of infrastructure as well as details of ownership and capacity for infrastructure sharing. Generally, a centralized database of the existing infrastructure not only concerns telecommunication infrastructure, such as wired, wireless or satellite infrastructure, but also alternative infrastructures of other industries (utilities such as electricity, gas, water and sewers; utility poles, including antenna masts; potential antenna sites on tall buildings). Among information about existing infrastructure, database could include information about planned works too.

Besides data about the location of ducts, it would be useful to know who owns the duct, what is the administrative procedures for granting rights of way to the existing infrastructure, and, whether the ducts are suitable for the deployment of additional infrastructure (e.g. whether there is sufficient space in a duct for more fibre). In fact, the centralised atlas of passive infrastructure, the one-stop shop on rights of way and administrative procedures, and the database of planned civil works (i.e. methods of reducing costs mentioned in EU Initiative to Reduce the Cost of Rolling out High Speed Communication Infrastructure) are interlinked. All of these measures require a similar database so if implemented in parallel, it is likely that much of the IT implementation costs would overlap between these measures and the resulting system would enable the implementation of mandated access to passive infrastructure, and thus encourage both deployment in shared ducts and self-deployment.

There are two principal advantages of centralised infrastructure atlas:

- Operators and utility companies that are due to carry out civil works are more likely to be informed about where existing infrastructure is located, and hence are less likely to cause damage to that infrastructure when carrying out their own excavation works for new deployments. The reduction in damage to existing infrastructure during excavation work, could be between EUR10 million and EUR50 million per annum in some EU countries. It is therefore possible that the cost savings from damage to existing infrastructure alone could equate the cost of implementing an infrastructure atlas in perhaps two to three years [6].
- Such an atlas would be an enabler of passive infrastructure sharing, which could reduce the cost of NGA deployment. Operators would be able to find out exactly where existing ducts lie, and may be able to place new cables and fibres within these, rather than carrying out their own excavation works and installing their own ducts, thus saving time, money and reducing unnecessary civil disruption.

Fig.1. show estimation of the costs and overall benefits of implementing regulatory measures for reducing costs of NGA deployment - infrastructure atlas, in-building wiring, mandated access to infrastructure, coordination of civil works and one stop shop on rights of way and administrative procedures [6].

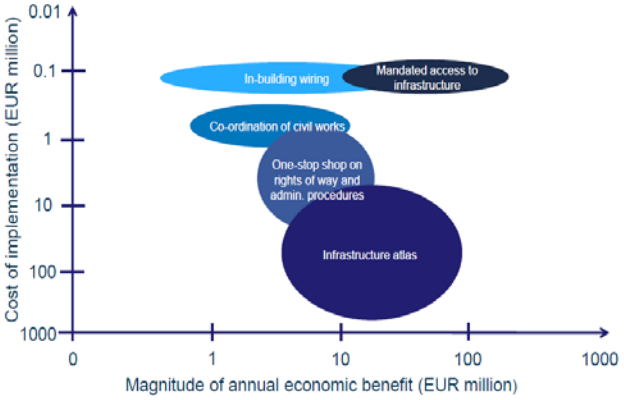


Fig.1. Estimate of the cost and overall benefits of five regulatory measures

Regulatory authorities have been entrusted with developing the tool, the collection of data and the management of access rights and requests. So far, about half of the Member States have started projects to set up a civil infrastructure mapping, that is, register of civil infrastructure that can be used in broadband network roll-out. Some are already operational and responding to information requests [7].

But, approaches differ, while some authorities try to gather as much data as possible to have information which is as complete as possible, others begin by placing a higher priority on passive infrastructure assets in the public domain.

However, infrastructure atlases face same obstacles:

- Problems with collecting information due to data owners' lack of incentive to coordinate. The revised regulatory framework for electronic communications partially solves this problem in that it imposes an obligation on electronic communications network providers to give information on the location of their infrastructure for the purposes of network roll-out. Some countries have extended disclosure obligations to include all relevant infrastructure assets (including those outside the telecommunications sector).
- Security and competitive issues. Concerns about the security of the information disclosed are considered to inhibit the development of infrastructure mapping. It is argued that there is a need to protect sensitive information, either on grounds of public safety or for commercial reasons. However, demands for tighter security can be excessive and consequently restrict infrastructure mapping projects. These difficulties can be solved by the sophisticated definition and management of access rights, redacting confidential details and limiting disclosures to the subset of information that is necessary for the tool to be useful.

- Probably the biggest problem is that mapping projects are very expensive, with the cost being heavily dependent on the detail and scale of the mapping project implemented, as well as the amount of prior knowledge regarding the location of infrastructure.

Many of the existing maps have multiple purposes, like providing a portal for announcing planned civil works, or reducing the administrative burden associated with the planning or permit application process. Thus, it is likely that some implementation costs would overlap across these different measures, in particular IT costs (which have typically been found to be in the EUR several millions range), and the collection and processing of data, which could amount to many hundreds of staff hours each year [6].

As the cost of implementing an infrastructure atlas is largely dependent on the detail of the data included in the database, it is reasonably to use a two-phase approach:

- The first phase could contain geographical information of existing passive infrastructure, populated by requesting the information from the operators and utility companies;
- The second phase may provide more detailed information about the share ability of each duct, from the results of a ground survey.

With this approach it might be possible to implement the first phase fairly quickly and at a reasonable cost to the regulatory authority, supposing the information is readily available from operators and utility companies and no surveys are required. Therefore, this might allow telecom operators to identify expansion opportunities that they did not originally believe to be economically viable.

Set-up and maintenance costs might be relevant, depending on the inclusion of old infrastructures whose information might not be available, the level of detail of the information included, the need to update it, as well as on the need to adapt to a standard format in view of the different mapping systems used by each operator or across sectors and countries. A few respondents to the public consultation on EU Initiative to Reduce the Cost of Rolling out High Speed Communication Infrastructure considered that, at least for old passive infrastructures, costs would outweigh the benefits, while confirming its feasibility for new facilities. The vast majority of respondents, on the contrary, stressed the importance to find the right balance in defining the level of detail of the information, also on the basis of the available existing information, in order to reduce the costs of the exercise, while at the same time ensuring most of the benefits [8]. In particular, it was concluded that the systematic information needed at an early stage, such as in planning and negotiation phases, is significantly different from the more granular and detailed information needed in the execution phase. Moreover, continuous inspections are in any case needed in order to assess the current state of the facilities. In conclusion, while standardised and easily accessible basic information appears to be highly valuable at an early stage, systematic high level of detail might not bring significant added value, while it has a significant impact on the overall costs of the system. Respondents mentioned several estimation infrastructure atlas costs, either on a per unit basis (1-4€ per squared meter mapped), or based on existing experiences (77mln€ Capex in Flanders, approx 9-10mln€ Opex in Netherlands; 4mln€ contract tendered in Portugal; approx 1230 mln€ Capex + approx 123 mln€ administrative costs in Poland) [8].

IV. Infrastructure atlas impact on the local NGA deployment

Generally, many national and local authorities think that systematic mapping enhances urban planning and soil management. Both operators and public authorities also see environmental benefits, in terms of reduction of need for civil works and better coordination, as well as administrative benefits with regard to the management of permit granting procedures. Other utilities companies and public authorities mentioned the benefits of systematic knowledge of networks infrastructures in order to improve disaster management [8].

V. Conclusion

A centralised infrastructure atlas could be a significant enabler of mandated access to passive infrastructure, but depending on the detail of the mapping and the amount of prior infrastructure knowledge, price of implementation could be very expensive. However, infrastructure atlas could reduce both the cost and construction timescale of NGA deployment by exploiting pre-existing infrastructure, and if the additional social and economic benefits of reduced damage to existing infrastructure during civil works, as well as increasing the opportunity for the coordination of civil works are taken into account, such a mapping project could be worthwhile.

References

[1] Communication from the Commission: Single Market Act II Together for new growth, COM (2012) 573 final, http://ec.europa.eu /internal_market /smact/docs/single-market-act2_en.pdf

[2] Public Consultation on an EU Initiative to Reduce the Cost of Rolling Out High Speed Communication Infrastructure in Europe, http://ec.europa.eu/information_society/policy/ecomm/doc/library/public_consult/cost_reduction_hsi/cost_reduction.pdf

[3] EC Press release, 18 Dec 2012 Digital "to-do" list: new digital priorities for 2013-2014 http://europa.eu/rapid/press-release_IP-12 1389_en.htm

[4] Commission Recommendation of 20 September 2010 on regulated access to Next Generation Access Networks (NGA) (2010/572/EU) http://eur-lex.europa.eu/

[5] Communication from the Commission EU Guidelines for the application of state aid rules in relation to the rapid deployment of broadband networks (2013/C 25/01) 26.01.2013. http://eur-lex.europa.eu /LexUriServ/LexUriServ.do?uri=OJ:C:2013:025:0001:0026:EN:PDF

[6] Matt Yardley, Rod Parker, Mike Vroobel: 'Support for the preparation of an impact assessment to accompany an EU initiative on reducing the costs of high-speed broadband infrastructure deployment'', Analysys Mason study, November 2012, https://ec.europa.eu/digital-agenda/en/news/support-preparation-impact-assessment-accompany-eu-initiative-reducing-costs-high-speed

[7] Commission Staff Working document on the implementation of national broadband plans, Brussels, 23.3.2012. SWD (2012) 68 final/2 http://ec.europa.eu/information_society/newsroom

[8] Results of the public consultation on how to reduce the cost of rollout of high speed broadband, ec.europa.eu/information_society/newsroom

Trendovi i mogućnosti primjene mobilnog širokopojasnog pristupa

Amina Krivošić*, Almedin Džebo *

* Regulatorna agencija za komunikacije/Sarajevo, Bosna i Hercegovina

akrivosic@rak.ba, adzebo@rak.ba

Sažetak

Rad ukazuje na potrebu razvoja politike uvođenja mobilnog širokopojasnog pristupa na globalnom i nacionalnom nivou kao osnovu za ostvarivanje mogućnosti konekcije u pokretu. 24 satna konekcija i veliki broj uređaja preko kojih se može povezati na internet današnja su realnost. Razvoj tržišta zahtijeva kreiranje politike za promovisanje pristupačnog univerzalnog pristupa širokopojasnim servisima. Mnoge zemlje danas imaju nacionalne planove razvoja širokopojasnog pristupa, ali se uglavnom fokusiraju na izgradnju i održavanje robusne fiksne infrastrukture. Mobilni broadband vodi boljem razvoju socio ekonomskih aspekata društva kroz pristupačniju edukaciju, unaprjeđenje zdravstvene njege, efikasnije uključivanje u finansije, omogućavanje e-governmenta, ruralnom razvoju i dr. Sa stanovišta upravljanja spektrom cilj je: više spektra za komercijalne mobilne broadband servise koji se može postići prije svega kroz harmonizaciju spektra koja zajedno sa međunarodno prihvaćenim standardima vodi ka znatnom porastu korištenja i shvatanju mobilnih tehnologija. Provođenje i implementacija regulatornog okvira Evropske unije doprinosi ekonomičnijem pokrivanju bežičnim širokopojasnim sistemima i izbjegavanje fragmentacije tržišta. Bežične elektronske komunikacije kao integralni dio komunikacijske infrastrukture i promocija evropske konkurentnosti ostvareni su kroz regulatorni okvir Evropske unije za elektronske komunikacije, a politika upravljanja radiofrekvencijskim spektrom podržava ga i dopunjava.

Ključne riječi: širokopojasni pristup, mobilne komunikacije, upavljanje spektrom, regulatorni okvir

I. UVOD

Širokopojasne radiokomunikacijske mreže imaju važnu ulogu u omogućavanju pristupa internetu visokog kapaciteta u pokretu kako u urbanim tako i u ruralnim područjima. Bežične komunikacije često su najefikasnija opcija za brzo rješavanje potreba pristupne infrastrukture. Ekspanzija u mobilnim širokopojasnim mrežama prouzrokovala je razvoj bržeg širokopojasnog interneta na udaljenim i ruralnim područjima. Protok podataka putem mobilnih komunikacija u 2010. godini veći je od predviđenog i do pet puta [1], tako da je vrlo nezahvalno vršiti takve vrste procjena jer se mobilne komunikacije nezaustavljivo razvijaju i šire, a prije svega zbog sve većih zahtjeva krajnjih korisnika. Prema podacima iz februara 2011. godine, više od 90% svjetske populacije je pokriveno signalom mobilne telefonije i dostupne su im mobilne mreže, što poredeći sa globalnom mobilnom internet penetracijom od 30% predstavlja veliki potencijal da mobilni broadband postane glavna mogućnost pristupa internetu. Zapravo, za većinu zemalja u razvoju prvi pristup internetu je ostvaren upravo preko mobilne IMT mreže [1].

II. DIGITALNA AGENDA ZA EVROPU

Evropska unija je razvila strategiju za razvoj ekonomije baziran na informaciono komunikacijskim tehnologijama (ICT) pod nazivom Digitalna agenda 2010. Ukoliko se žele u potpunosti iskoristiti potencijalne koristi od digitalne ekonomije, potrebno je unaprijediti opremu i omogućiti pružanje znatno bržeg širokopojasnog interneta, povećati povjerenje ljudi u internet, poboljšati vještine i potaknuti inovacije u informaciono komunikacijskim tehnologijama. Prijedlog konkretnih mjera u tim područjima je razvijen od strane Evropske komisije putem Digitalne agende za Evropu[1] koja uspostavlja sedam oblasti djelovanja:

1. jedinstveno digitalno tržište,
2. poboljšanje interoperabilnosti proizvoda i usluga koje se zasnivaju na ICT tehnologijama,
3. jačanje povjerenja i sigurnosti u internet,
4. osiguranje bržeg pristupa internetu,
5. podrška ulaganju u istraživanje i razvoj,
6. poboljšanje digitalne pismenosti, znanja i vještina,
7. omogućavanje koristi od ICT tehnologija u oblastima kao što su: zaštita okoline, zdravstvo, kultura, transport i sl.

37 evropskih zemalja je kroz različite dokumente usvojilo nacionalne politike promocije broadbanda[2]. Kad kažemo politika podrazumijeva se nacionalni broadband plan, broadband strategija, inicijativa vlade, ICT plan koji uključuje razvoj broadband infrastrukture, broadband politika i sl. Pretpostavlja se da je većina zemalja krenula u realizaciju planova implementacije širokopojasnog pristupa kao neminovnosti savremenog društva. Cilj je postići univerzalnost kroz kreiranje strategije razvoja širopojasnosnosti.

[1] Dostupno na http://ec.europa.eu/information_society/digital-agenda

[2] ITU World Telecommunication Regulatory Database, Broadband Policies Worldwide, year 2011

Evropska komisija je krajem 2012. godine usvojila sedam novih prioriteta za digitalnu ekonomiju i društvo. Značajno je spomenuti da je porast digitalne ekonomije sedam puta veći od porasta ostalih ekonomija. Današnji prioriteti slijede sveobuhvatan pregled politika i nadopunjavaju Digitalnu agendu za Evropu 2010. Prioriteti su usmjereni na povećanje ulaganja u širokopojasni pristup i cilj je maksimizirati doprinos digitalnog sektora u ekonomiji[3].

Puna provedba ove obnovljene Digitalne agende bi povećala evropski GDP za 5%, ili 1500 € po osobi, u idućih osam godina, povećanjem ulaganja u ICT, poboljšanjem nivoa e-vještina kod zaposlenog stanovništva, omogućavajući inovacije u javnom sektoru i reformom okvirnih uvjeta za internet ekonomiju.

Novi prioriteti za 2013-2014 su:

1. Stvaranje novog i stabilnog regulatornog okruženja za širokopojasni pristup
2. Nova infrastruktura javnog digitalnog servisa
3. Pokrenuti koaliciju digitalnih vještina i radnih mjesta
4. Predložiti strategiju i direktivu Evropske unije za sigurnost u kompjuterskim mrežama
5. Obnoviti direktivu EU o autorskim pravima
6. Ubrzati cloud computing kroz kupovnu moć javnog sektora
7. Pokrenuti novu strategiju elektronske industrije

A. Komisija za širokopojasni pristup

ITU i UNESCO su u maju 2010. godine osnovali Komisiju za širokopojasni pristup za digitalni razvoj na zahtjev generalnog sektretara UN-a Ban Ki-moon-a koji poziva za ubrzanje napora UN-a na ispunjavanju Milenijskih ciljeva razvoja. Njen cilj je da jača važnost širokopojasnosti kao međunarodni politički program. Vjeruje se da je širenje širokopojasnog pristupa, u svakoj zemlji ključ za ubrzanje napretka prema tim ciljevima sa 2015. godinom kao zadatim rokom za provedbu. Definirani su praktični načini na koji se to može postići u saradnji s privatnim sektorom u svim fazama razvoja. Sastoji se od uticajnih organizacija, uključujući izvršna rukovodstva i vodeće industrije, značajne kreatore politike i predstavnike vlasti, međunarodne agencije, akademske zajednice i organizacije koje se bave razvojem. Čelnici Komisije, svako u svom području, vjeruju u budućnost na temelju širokopojasnosti i nude bogate uvide i iskustva u to kako implementirati i koristiti širokopojasne mreže i usluge na dobrobit zajednice i krajnjih korisnika [2].

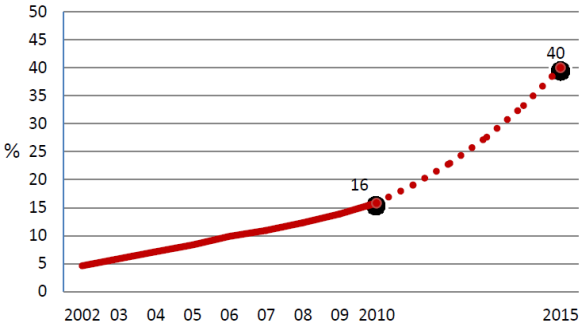
Komisija za širokopojasni pristup zagovara i pokazuje da su širokopojasne mreže:

- osnovna infrastruktura u savremenom društvu baš kao i cesta, struja i voda;
- jedinstveno moćan alat za ubrzanje napretka prema Milenijskim ciljevima,
- iznimno ekonomične i da nude impresivan povratak investicija u ekonomiji u razvoju i razvijenoj ekonomiji podjednako;
- podupiru sve industrijske sektore i poboljšavaju javne usluge i socijalni napredak;
- treba da budu podržane od strane vlada u zajedničkoj saradnji s industrijom, kako bi se u potpunosti iskoristile prednosti širokopojasnih mreža i usluga.

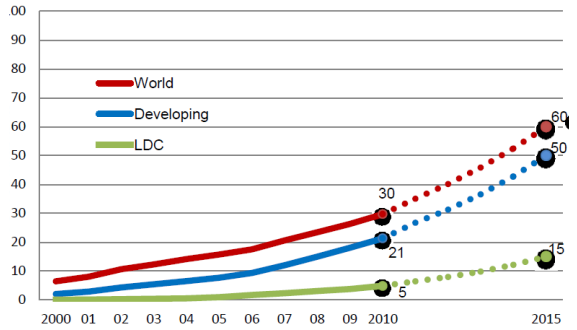
Komisija smatra da je brza širokopojasna konekcija velikog kapaciteta bitan element u modernom društvu, pripisujući joj široke društvene i ekonomske koristi [2]. Komisija za širokopojasni pristup postavila je četiri jasna, nova cilja za izradu univerzalne širokopojasne politike u svrhu jačanja dostupnosti širokopojasnostog pristupa i boljem shvatanju širokopojasnosti uopće:

- Cilj 1: Izrada univerzalne politike širokopojasnosti. Do 2015. god., sve zemlje trebale bi imati nacionalni plan ili strategiju širokopojasnosti ili uključiti širokopojasnost u definicije univerzalnog pristupa, odnosno univerzalnih usluga.
- Cilj 2: Učiniti dostupnim širokopojasni pristup. Do 2015. god., osnovni model širokopojasne usluge treba biti dostupan u zemljama u razvoju kroz adekvatno reguliranje i tržišnu održivost (u iznosu od najmanje 5% od prosječnog mjesečnog prihoda).
- Cilj 3: Povezati domaćinstva na širokopojasni pristup internetu. Do 2015, 40% domaćinstava u zemljama u razvoju treba imati pristup internetu. U razvijenim zemljama, više od dvije trećine domaćinstava već ima pristup internetu prema mjerenjima na kraju 2010. godine, u odnosu na oko 16% domaćinstava u zemljama u razvoju. Vjerovatno je da će se ovo značajno povećati do 2015. god., a posebno s razvojem mobilnog interneta. Ovaj cilj uključuje pristup internetu i fiksnim i mobilnim mrežama.
- Cilj 4: Biti online. Do 2015., penetracija internet korisnika treba doći do 60% u svijetu, 50% u zemljama u razvoju i 15% u nerazvijenim zemljama. Ove prognoze su ambicioznije od ITU prognoza do 2015. god. Korištenje interneta može biti putem fiksne ili mobilne mreže, te obuhvaća korištenje interneta kod kuće, na poslu, u školi, kod prijatelja, u internet kafićima, na javnim mjestima - drugim riječima internet koristiti bilo gdje.

[3] http://ec.europa.eu/digital-agenda/en/news/digital-do-list-new-digital-priorities-2013-2014



Slika 1 – Postotak domaćinstava sa internet pristupom (zemlje u razvoju) – izvor Broadband Commission



Slika 2 – Penetracija interneta - izvor Broadband Commission

Na slici 2 je prikazana penetracija interneta u 2010. godini i procjena povećanja do 2015. godine za svjetsku populaciju (World), zemlje u razvoju (Developing) i najmanje razvijene zemlje (LDC - Least Developed Countries).

III. VIŠE SPEKTRA ZA MOBILNI ŠIROKOPOJASNI PRISTUP

U cilju osiguravanja da konkurentne aplikacije koje koriste radio spektar ne ometaju jedni druge, bitno je da raspodjela frekvencija bude koordinirana i pravilno regulisana na nacionalnom i međunarodnom nivou, a pravila za korištenje spektra moraju biti usklađena na širem području od nacionalnog. Za jedinstveno evropsko tržište ova koordinacija upravljanja spektrom će biti na evropskom nivou kako bi se postiglo da je svrha efikasnije regulacije evropska korist. Cilj politike Evopske unije u ovom području je koordinacija upravljanja radio spektrom u Uniji kako bi se osiguralo da Evropa dobije maksimalnu korist iz korištenja ovog ograničenog resursa. Na taj način će efikasna regulacija predstavljati korist za ovaj sektor.

Na globalnom nivou sektor radiokomunikacija Međunarodne unije za telekomunikacije (ITU-R) nadzire globalnu raspodjelu i upravljanje radiofrekvencijskog spektra, identificira, upravlja i nadgleda radiofrekvencijski spektar kako bi se osigurao efikasan rad širokopojasnih mreža bez smetnji prema drugim radio sistemima i obrnuto. ITU-R također radi na razvoju globalno prihvaćenih tehničkih standarda za bežične širokopojasne sisteme. ITU-R studijske grupe ujedinjuju vodeće stručnjake mrežnih operatora, Internet servis provajdera, emitera, regulatora i istraživačko razvojnih organizacija za razvoj tehničkih standarda za interoperabilnost i performanse zemaljskih i satelitskih širokopojasnih sistema. ITU-R je razvio prvu grupu standarda za međunarodni globalni širokopojasni multimedijski mobilni sistem poznat kao IMT-2000, odnosno 3G. Sada se radi na slijedećoj generaciji IMT-Advanced mobilnih sistema.

IMT-Advanced sistem uključuje nove mogućnosti koje nadilaze IMT-2000, koji je rasprostranjen još od 2000. ITU je speificirao standarde za IMT-Advanced, iduću generaciju globalnih bežičnih širokopojasnih komunikacija koje pružaju pristup širokom spektru telekomunikacijskih usluga baziranih na paketskim uslugama koje mobilne i fiksne mreže podržavaju. ITU-R također nadgleda dodjelu orbitalnih pozicija za satelitske širokopojasne sisteme, koji nude perspektivu povezivanja cijele zemlje brzim internet pristupom kroz jednu satelitsku mrežu. Pored toga poduzeta su istraživanja potencijalnih opsega za nove dodjele za mobilne satelitske službe (MSS) sisteme. Očekuje se u bliskoj budućnosti razvijanje ažuriranih standarda za fiksne satelitske širokopojasne sisteme.

Tokom Svjetske radio konferencije 2012 (WRC-12) pregledana je i potvrđena globalna raspodjela spektra kako bi se osiguralo da se ovaj dragocjeni resurs koristi učinkovito u korist ITU članica i korisnika radiofrekvencijskog spektra i mobilne telefonije bilo gdje u svijetu. Kroz svoje brojne aktivnosti, ITU-R omogućuje isporuku pravednog, pristupačnog širokopojasnog pristupa internetu za sve ljude, bez obzira gdje žive.

A. Potražnja za spektrom

Širenje tržišta mobilnog širokopojsnog pristupa rezultiralo je povećanom potražnjom za širinom opsega, smanjenje granica za mobilne operatore i poboljšanje virtualnih iskustava za korisnike. Rast širokopojasnih usluga preko bežičnih mreža povećao je pritisak na radiofrekvencijski spektar. Povećan je broj regulatora koji su uveli tržišno bazirane mehanizme kao što su: trgovina spektrom, dijelenje spektra kako bi se raspodijelio spektar u svrhu brzog i efikasnog odgovora na zahtjeve za novim opsezima za 3G i 4G usluge.

Zbog sve većih potreba za spektrom i vrlo dinamičnih promjena u njegovoj regulaciji, regulatori bi trebali prestati davati licence sa ekskluzivnim pravima za određene opsege. Neki regulatori su dodijelili tzv. nelicencne opsege, čime se daje više slobode učesnicima na tržištu da sami između sebe upravljaju spektrom. Tržišno upravljan pristup, baziran na poticaju da bi se došlo do više raspoloživog spektra za mobilne širokopojsne usluge danas se smatra najboljom praksom omogućavajući konkurenciju

i poticanje inovacija. Prema zaključcima WRC-12, digitalna dividenda vjerovatno će omogućiti spektar za LTE (Long-Term Evolution) mreže. Gotovo polovina zemalja u svijetu koriste aukcije kao primarni način dodjele spektra za najpopularnije mobilne širokopojasne usluge, 3G (40 %) i WiMAX (45 %), sistem "Prvi došao, prvi uslužen" koristi se u trećini dodjela, a "beauty contest" u jednoj petini svih zemalja. Dodjela spektra za LTE usluge razlikuje se od tih utvrđenih obrazaca nakon što je uvedena tehnologija neutralnog licenciranja, a refarming frekvencija igra sve veću ulogu.

B. Radio frekvencije za mobilni širokopojsani pristup

Evropski Parlament sklopio je sporazum sa državama članicama o tome kako najbolje preraspodijeliti frekvencije radio spektra koje su oslobođene prilikom prelaska sa analognog na digitalno emitovanje i taj spektar dodijeliti za novu namjenu, poput bežičnih širokopojasnih usluga. To će pomoći da se riješi pitanje porasta korištenja bežičnih širokopojasnih usluga. Prva politika Evropske unije o radio spektru omogućit će da EU povrati vodeću ulogu kad su u pitanju bežične komunikacije. Ona će optimizirati korištenje frekvencija za dobrobit svih korisnika pametnih telefona i drugih mobilnih digitalnih platformi. Dodjela radijskih frekvencija je stvar nacionalnih tijela, a ne EU, ali pravila o tome kako podijeliti spektar među pružaocima usluga interneta i korisnicima trebaju biti dogovoreni na nivou EU-a.

Specifičan regulatorni instrument za dalje poboljšanje saradnje i koordinacije između država članica u vezi politike upravljanja spektrom nazvan program politike radio spektra (RSPP – Radio Spectrum Policy Programme) u EU uveden je kao dio novog regulatornog okvira.

Na osnovu odredbi Okvirne direktive, Evropska komisija sada može podnositi zakonske prijedloge Evropskom parlamentu i Vijeću za uspostavljanje višegodišnjeg programa politike radio spektra za utvrđivanje političke orijentacije i ciljeva za strateško planiranje i harmonizaciju korištenja radio spektra u Evropskoj uniji. To predstavlja priliku da se uključe važni donosioci odluka u postizanju šireg konsenzusa o smjeru politike spektra.

Program postavlja opća regulatorna načela i političke ciljeve koji će se primjenjivati za spektar za sve sektore unutarnjeg tržišta, definiše aktivnosti i zajedničke principe kako bi se poboljšala efikasnost i fleksibilnost, očuvanje i promocija konkurencije, daje podršku bežičnim širokopojasnim komunikacijama, kao i druge politike EU, kao što su promet, zaštita okoline, nadgledanje Zemljine površine ili istraživanje svemira. Također, inicira definisanje i praćenje postojećih i novih potreba za spektrom, utvrđuje načela koja brane interese EU-a na međunarodnom nivou i potiče napredak u procesu standardizacije [5]. RSPP postavlja konkretne prioritetne inicijative za bolju koordinaciju, fleksibilnost i dostupnost spektra za bežične širokopojasne komunikacije i druge specifične politike EU, poboljšava zaštitu interesa EU-a u međunarodnim pregovorima te pomaže državama članicama u bilateralnim pregovorima. Evropski parlament je 14. marta 2012. godine usvojio Odluku o prvom višegodišnjem programu poliltike radio spektra koja stvara sveobuhvatan program politike radio spektra u EU do 2015. godine za kompletiranje unutrašnjeg tržišta, posebno u skladu s evropskom inicijativom 2020 i digitalnom agendom za Evropu[5].

Prema dogovoru, frekvencijski opseg 800 MHz korišten za TV, trebao bi biti dostupan za bežične širokopojasne usluge u svim državama članicama EU od 1. januara 2013. To će pomoći u stvaranju pan-Evropskog telekomunikacijskog tržišta prilikom čega nove usluge otvaraju nove mogućnosti i porast za oko 500 milijuna potrošača EU. Na zahtjev Parlamenta, treba dodijeliti najmanje 1200 MHz, za mobilni podatkovni promet do 2015. godine, prema legislativi već odobrenoj od strane država članica. Nadalje, Komisija treba procijeniti, najkasnije do 1. januara 2015., da li postoji potreba za usklađivanjem dodatnih frekvencijskih opsega kako bi se upravljalo eksponencijalnim rastom bežičnog podatkovnog prometa.

Omogućavanje brzog širokopojasnog web pristupa svim evropskim građanima i poslovnim subjektima, uključujući i one koje se nalaze u ruralnim područjima ili na udaljenim otocima, nemoguće je bez bežičnog interneta, koji treba radio frekvencije za neprekidne veze. To se može postići dodjelom "digitalne dividende" – tj frekvencija koje će biti oslobođene prilikom prelaska sa analogno na digitalno emitiranje televizijskih programa.

Mada bi prema planovima Evropske komisije, zemlje EU trebale omogućiti da opseg 800 MHz bude dostupan za usklađeno korištenje bežičnih širokopojasnih usluga od 1. januara 2013, Parlament je donio izmjene i dopune koje će omogućiti državama članicama da traže odgodu do kraja 2015., ili čak i više, ako imaju problema u prekograničnoj koordinaciji frekvencija sa susjednim trećim zemljama. Nakon 2013. godine opseg 2,3 GHz također bi trebao biti dostupan za mobilni širokopojasni pristup, dok namijenjeni spektar za mobilni podatkovni promet treba doći do barem 1200 MHz do 2015. godine. Osim toga, Komisija bi trebala pratiti tehnološki razvoj i učinkovitost upotrebe spektra te ocijeniti da li dodatne frekvencije (kao što je 700 MHz) također trebaju biti usklađene kako bi napravili mjesta za nove korisnike i usluge.

Novi spektar frekvencija kao ultimativni faktor razvoja mobilnih širokopojasnih usluga, postavlja pred regulatore veoma složen zadatak reforme spektra. Reforma zahtjeva replaniranje dodijeljnih frekvencijskih resursa i to uzimajući u obzir više radiofrekvencijskih opsega koji predstavljaju potencijalne RF opsege za implementaciju mobilnih širokopojasnih tehnologija.

Dodatni frekvencijski opsegi koji su određeni kao pogodni za implementaciju mobilnih širokopojasnih tehnologija 4G su:

- 450 – 470 MHz
- 790 – 862 MHz
- 2300 – 2400 MHz
- 2500 – 2690 MHz
- 3400 – 3600 MHz

C. BDA2GC

Potreba za mobinim širokopojasnim pristupom je osim u udaljenim i nepristupačnim dijelovima te drugim mjestima gdje je fiksna infrastruktura nedostupna, dolazi do izražaja i kod pomorskih i zračnih komunikacija, odnosno kao vid komunikacije na letjelicama i plovilima.

U toku je izrada ECC izvještaja o sistemu širokopojasne direktne komunikacije između zraka i zemlje (DA2GC - Direct Air-to-Ground Communication) koji predstavlja zahtjev za različite vrste telekomunikacijskih usluga, kao što su pristup internetu i mobilnim multimedijalnim uslugama [5]. Cilj je omogućiti pristup širokopojasnim komunikacijskim uslugama na nivou cijele Evrope tokom kontinentalnih letova. Zahtjev za spektrom se odnosi na direketno radio rješenje zrak-zemlja.

Korisnici mobilnih komunikacija očekuju konekciju gdje god se nalaze sa bilo kojom vrstom mobilnog uređaja. To uključuje obezbjeđenje brodband usluga na letjelicama, a evropske aviokompanije imaju veliki interes da ponude internet usluge korisnicima za vrijeme kontinentalnih letova što je prije moguće.

Uspostavljanje pan-Evropskog regulatornog okruženja za broadband DA2GC obezbijediće za korisnike, avio kompanije i putnike koji lete u Evropi:

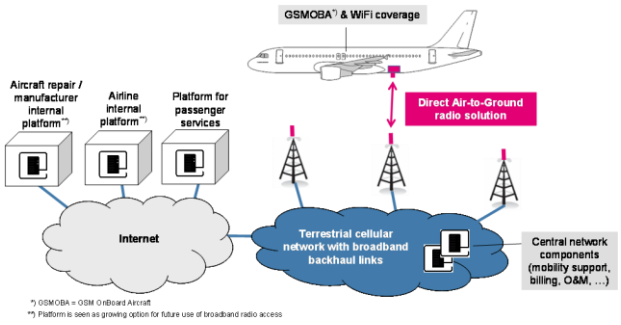
- alternativu koja ubrzava konkurenciju i vodi ka manjim troškovima za kompaniju i putnike,
- tehničku implementaciju koja stimuliše konkurenciju,
- putem DA2GC izbjegavaju se kašnjenja koja se dešavaju kod geostacionarnih satelitskih veza,
- cijena instalacija i održavanje su ključni za aviokompanije jer činjenica da se DA2GC oprema može instalirati preko noći na avionu kao i specifikacija antena, zemaljskoj soluciji daje prednost u odnosu na satelitsku.

Konekcija između zemlje i letjelice može biti putem satelita ili putem DA2GC. Što se tiče satelitskog broadbanda u pripremi je puštanje u rad Ka opsega za koji se pretpostavlja da će zadovoljiti potrebe za širokopojsnim uslugama tokom leta[4]. ECC također radi na odluci koja podržava ESOMP (Earth Stations on Mobile Platforms). Istovremeno DA2GC usluge su već uspostavljene u Sjevernoj Americi (USA i Kanada), s više od 1.300 komercijalnih zrakoplova. Također, uprava civilnog zrakoplovstva u Kini (CAAC) testirala je širokopojasni komunikacijski sistem zemlja-zrak i ima ga u planu pustiti u komercijalnu upotrebu.

Frekvencijski opsezi za DA2GC su podijeljeni u 4 kategorije:

I kategorija (2400-2483.5 MHz, 3400-3600 MHz, 5855-5875 MHz)
II kategorija (2483.5-2500 MHz, 5150-5170 MHz)
III kategorija (1452-1492, 1900-1920/2010-2025, 3600-3800, 5905-5925)
IV kategorija (2300-2400 MHz)

Na slici je prikazana arhitektura sistema za širokopojasne direktne komunikacije između zraka i zemlje [5]. Kod uspostavljanja ovog sistema, a što se tiče licenciranja, još uvijek se razmatraju opcije individualne ili opće autorizacije. Također, radi se na proračunima scenarija interferencije između pojedinih radiofrekvencijskih opsega kako bi se zadovoljila međusobna kompatibilnost zajedničkog rada pojedinih radiokomunikacijskih servisa .



Slika 3. Arhitektrura sistema za širokopojasni DA2GC, izvor draft ECC izvještaj XX/13

[4] U toku je priprema eksperimentalnog puštanja u rad EUTELSAT Ka band satelita koji bi u drugoj polovini 2013. godine počeli sa pružanjem dvosmjernog satelitskog interneta za vrijeme kontinentalnih letova u Evropi

IV. ZAKLJUČAK

Bez širokopojasne infrastrukture i usluga, zemlje u razvoju riskiraju isključenje iz sudjelovanja u nadolazećoj globalnoj digitalnoj ekonomiji. Cilj koji je postavila Komisija za širokopojasni pristup, a koji se odnosi na promoviranje usvajanja prihvatljive širokopojasne prakse i politike, tako da svi ljudi mogu iskoristiti prednosti širokopojasnog pristupa, potrebno je iskoristiti i obavezno uvrstiti u strateške ciljeve razvoja zemlje. Bosna i Hercegovina je jedna od rijetkih evropskih zemalja koja do sad nije usvojila okvir za uspostavu strategije razvoja širokopojasnog pristupa. Prilika je da upravo mobilni širokopojasni pristup, njegovo širenje i razvoj bude odrednica za usvajanje strateškog cilja širokopojasnosti u periodu do 2015. godine i na taj način bude oslonac, a ne kočnica u stvaranju pan-Evropskog telekomunikacijskog tržišta prilikom čega nove usluge otvaraju nove mogućnosti za milione evropskih potrošača.

LITERATURA

[1] Report ITU-R M.2243 - Assessment of the global mobile broadband deployments and forecasts for International Mobile Telecommunications, 11/2011

[2] The State of Broadband 2012: Achieving digital inclusion for all – Report by Broadband Commission, Sept. 2012

[3] ITU. Regulation of Global Broadband Satellite Communications, Sept. 2011

[4] Universal Service: The trends, opportunities and best practices for universal access to broadband services, Intelecon Research & Consultancy Ltd., Vancouver BC, Canada

[5] DECISION No 243/2012/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 March 2012 establishing a multiannual radio spectrum policy programme

[6] Draft ECC Report on Broadband Direct-Air-to-Ground Communications (BDA2GC), doc. FM 48 (13)004 dostupno na www.eco.dk

[7] Ernst & Young (2012). Mobile Broadband readiness, Key drivers to create a positive environment, GSMA Mobile Asia Expo, June 2012

Prenosivost telefonskih brojeva

Adnan Gabela, Feriz Gabela, Jasmina Gabela
Dioničko društvo BH Telecom Sarajevo
Obala Kulina bana broj 8
Sarajevo
Telefoni: 387 – 33 586 508; 387 – 61 507 530
E-mail: adnan.gabela@bhtelecom.ba;feriz.gabela@bhtelecom.ba;

Sažetak - Ovaj rad daje pregled pravnog uređivanja prenosivosti telefonskih brojeva. Radi se o pravu telefonskog pretplatnika da zadrži stari telefonski broj nakon promjene operatera fiksne ili mobilne mreže i promjeni vrste korištenja telefonskih usluga. U prvom dijelu rada se utvrđuje pojam prenosivosti telefonskog broja i njegov značaj. Dalje se analizira servis prenosivosti telefonskog broja u svijetu i državama Europske unije. Na samom kraju je analizirano i pravno uređenje prenosivosti telefonskih brojeva u Bosni i Hercegovini i problemima u vezi s prenosivošću telefonskih brojeva.

Summary - This paper provides an overview of legal regulation of number portability. It is the right to retain old phone number after changing providers of fixed or mobile networks or changing uses of telephone services. The first part of the paper establishes the notion of telephone number portability and its significance. The second part analyzes telephone number portability service in the world and the countries of the European Union. At the end the legal system of number portability in Bosnia and Herzegovina and its problems is analyzed.

Ključne riječi – prenosivost broja, javna telefonska mreža, javna telefonska usluga, operater davatelj broja, fiksni operater, mobilni operater.

Keywords - number portability, public telephone networks, public telephone services, operator service number, fixed operator, mobile operator.

I. UVOD

Telefonski broj predstavlja jedan od elemenata identifikacije osoba, te pretplatnici telefonskog broja prihvaćaju promjenu telefonskog broja samo kada je to potrebno. Na svijetu ne postoje dva krajnja pretplatnika s istim telefonskim brojem. Cilj uvođenja usluge prenosivosti broja je poticanje konkurentnog tržišta, što se ogleda kroz smanjenje barijere pri promjeni operatera, odnosno mogućnost zadržavanja postojećeg broja, povećanu konkurentnost i smanjenje cijena javnih telefonskih usluga.

Prijenos broja je skup tehničkih i organizacijskih mjera kojima se izvršava usluga prenosivosti broja na propisan način, a obavlja se tako što pretplatnik bilo kojeg operatera telekomunikacija podnosi zahtjev svom operateru za raskid ugovora i prijenos broja na željenog operatera s kojim će zaključiti novi ugovor. Mogućnošću izbora operatera praktično počinje prava konkurencija na tržištu telekomunikacija. To je dobar signal za potencijalne korisnike telefonskih usluga, koji će svakako biti na dobitku, jer pojačavanje konkurencije znači i snižavanje cijena javnih telefonskih usluga.

Pravilom o prenosivosti broja, koje omogućava izbor telekom-operatera bez promjene telefonskog broja, usvojeno od strane Vijeća Regulatorne agencije za komunikacije, u Bosni i Hercegovini kreirane su obaveze koje proizilaze iz europskih preporuka. Zbog tehničkih problema koje je imao jedan od dominantnih operatera, usluga prenosivosti broja u mobilnoj telefoniji zaživljela je u Bosni i Hercegovini s početkom 2013. godine.

II. PRENOSIVOSTTELEFONSKOGBROJA

Prenosivost telefonskog broja je usluga koja omogućava pretplatniku da na vlastiti zahtjev, uz ispunjenje zadanih uvjeta, u okviru javno raspoloživog telefonskog servisa, zadrži svoj broj koji mu je dodijeljen u javnoj telefonskoj mreži, nezavisno o promjeni operatera telekomunikacija ili davatelja usluga. [1] Prenosivost telefonskog broja je institut koji se primjenjuje isključivo na telefonske mreže, bilo fiksne bilo mobilne. Pretplatnici mogu odabrati za sebe optimalnog operatera u skladu sa svojim potrebama i navikama.

Operateri i davatelji usluga moraju omogućiti pretplatnicima zadržavanje istog broja i predbroja, bez obzira na mrežu u koju prelaze. Prijenos telefonskog broja se može vršiti kako u fiksnim telefonskim linijama tako i u mobilnim telefonskim linijama.

Značaj prenosivosti telefonskog broja za konkurenciju je osnovni argument koji je upotrebljavan u zalaganjima za njegovo pravno uređenje. Ističe se da je telefonski broj jedan od elemenata identifikacije pretplatnika i da se stoga u njihovom interesu mora omogućiti zadržavanje ranijeg telefonskog broja u što više slučajeva. Takve ideje su i dovele do pomjeranja mjesta uređenja ovog instituta u dijelove zakona, koji se bave zaštitom pretplatnika, a u Europskoj uniji u Direktivu 2002/22/EZ o univerzalnoj usluzi.[3]

Kod prenosivosti telefonskog broja postoje tri situacije.

- Pretplatnik zadrži isti telefonski broj kad prelazi od jednog fiksnog operatera ka drugom fiksnom operateru.
- U nekim slučajevima pretplatnik može zadržati isti telefonski broj dok se prijelaz vrši između fiksnih i mobilnih operatera.
- Pretplatnik zadrži isti telefonski broj kad prelazi s jednog mobilnog operatera na drugog mobilnog operatera. [2]

Postoji prividna i stvarna prenosivost telefonskog broja između dva operatera.

Kod prividne prenosivosti, pravo na korištenje broja nije prenijeto na novog operatera, već ga zadržava prvi operater. Pošto prvi operater zadržava pravo korištenja broja, svi telefonski pozivi ka tom broju će biti usmjeravani ka njegovoj mreži, jer su telefonske centrale programirane da pozive usmjeravaju ka onoj mreži čijem je operateru broj dodijeljen na korištenje. Uloga prvog operatera se sastoji u tome da poziv koji stigne na njegovu mrežu preusmjeri na mrežu novog operatera, kako bi ga ovaj proslijedio do krajnjeg pretplatnika.

Kod stvarne prenosivosti, pravo na korištenje telefonskog broja se prenosi sa prvog na novog operatera u korist krajnjeg pretplatnika.

Proces omogućavanja stalnog ažuriranja telefonskih centrala s prenijetim pravima korištenja telefonskih brojeva čini nužnim postojanje centralne baze tih podataka, iz koje bi se centrale upoznavale s njima. Prijenos dodijeljenog prava korištenja broja je odnos između dva operatera i moguć je samo ako zakonom nije zabranjen ili ograničen. U tehničkom smislu se takav prijenos sprovodi na nacionalnom nivou i uključuje formiranje i održavanje centralne baze podataka i svaku telefonsku centralu u državi. Zbog toga je ovaj postupak, iako pravno jednostavniji, tehnički veoma skup i zahtijeva dugu pripremu i ispunjenje brojnih pretpostavki za njegovo ostvarenje.

U europskim državama se koristi prvi način (putem međupovezivanja, odnosno prividnog prijenosa), dok se u SAD intenzivno radi na prelasku na stvarni prijenos, formiranjem centralne baze podataka o prenijetim brojevima.

Razlikuju se tri grupe prenosivosti telefonskog broja:

- geografska prenosivost,
- prenosivost u pogledu usluge i
- prenosivost u pogledu davatelja usluga. [2]

Mnogi autori geografsku prenosivost broja pri zadržavanju istog davatelja usluga nazivaju mobilnošću broja i ne smatraju je vrstom prenosivosti, dok prenosivost u pogledu usluge ne spominju. [2] Prenosivost broja u pogledu usluge se može podijeliti na: prenosivost između različitih podvrsta fiksnih usluga, prenosivost između različitih podvrsta mobilnih telefonskih usluga i prenosivost između fiksnih i mobilnih telefonskih usluga.

Prenosivost u pogledu davatelja usluga se dijeli na: inicijalnu i naknadnu prenosivost broja. Inicijalna prenosivost broja postoji kada se broj prenosi od prvog k novom operateru. Naknadna prenosivost broja postoji kada se broj nakon inicijalnog prenosa prenosi od novog k trećem davatelju usluga koji nije prvi operater. [2]

III. PRAVNO UREĐENJE PRENOSIVOSTI BROJA

Usluga prenosivosti telefonskog broja regulira se zakonima, a detaljno podzakonskim aktima. Prema pravnom okviru Europske unije, prenosivost broja je uređena Direktivom 2002/22/EZ o univerzalnoj usluzi. Osim nacionalnih i regionalnih pravnih pravila, usluga prenosivosti nije uređena na međunarodnom nivou od strane Svjetske trgovinske organizacije i Međunarodne unije za telekomunikacije. [3]

U pravu Europske unije pravna pravila o prenosivosti broja u početku se odnosila isključivo na prenosivost telefonskih brojeva unutar fiksnih telefonskih usluga. Prenosivost između mobilnih telefonskih usluga uvedena je tek kasnije.

S uređenjem prenosivosti između mobilnih usluga se oklijevalo, jer je konkurencija na tržištima mobilne telefonije već bila razvijena, pa se smatralo da uređenje prenosivosti mobilnog telefonskog broja nije potrebno. U pravu Europske unije obavezu omogućavanja prenosivosti telefonskog broja ima svaki operater javne telefonske mreže i/ili javnih telefonskih usluga. Javne telefonske mreže su one mreže preko kojih se pružaju javne telefonske usluge, a javne telefonske usluge su one telefonske usluge koje su dostupne svim zainteresiranim osobama pod unaprijed poznatim uslovima. Pod pojmom telefonska usluga se podrazumjeva usluga upućivanja i primanja nacionalnih i međunarodnih poziva. U telefonske usluge spada usluga klasične govorne telefonije, kao prijenosa govora u realnom vremenu između dvije završne točke u telefonskoj mreži. [4] Ostaje

sporno mogu li se pod telefonsku uslugu podvesti prijenos podataka i Internet.

Pravo na prenosivost broja pripada pretplatniku javne telefonske usluge koji uslugu prima za ispunjenje vlastitih potreba, a ne onaj koji uslugu prima da bi mogao pružati telekomunikacijske usluge drugim osobama. Pravno uređenje prenosivosti telefonskog broja sadrži i druge odredbe kojima se ograničava sloboda postupanja operatera u vezi s ovom uslugom. Operateri često svojim ponašanjem u vezi s prijenosom destimuliraju krajnje korisnike u odluci da promjene operatera. [5]

Odredbama telekomunikacijskog prava se propisuje naplata naknade u vezi s uslugom prenosivosti, kako ne bi predstavljala teret za pretplatnike i umanjila značaj prenosivosti.

Povodom prenosivosti telefonskog broja operateri naplaćuju dvije vrste naknada: naknadu od krajnjeg pretplatnika (direktna naknada) i naknadu od drugih operatora (posredna naknada). Direktiva 2002/22/EZ o univerzalnoj usluzi ograničava visinu direktne naknade pravilom da ona ne smije odvraćati pretplatnike od korištenja usluge prenosivosti. [3] U pojedinim europskim telekomunikacijskim pravima se zabranjuje operaterima da troškove prenosivosti brojeva prevale na svoje pretplatnike, pa te troškove moraju sami snositi.

Posredna naknada je naknada koju naplaćuje prvi operater novom operateru na osnovu ugovora o interkonekciji (međupovezivanju) putem koga se sprovodi prenos i koju naplaćuje operater mreže operateru usluga koji je u odnosu sa pretplatnikom čiji se broj prenosi. [3]

U pravu SAD se izričito propisuje da obaveza prenosivosti telefonskog broja podrazumjeva osiguranje korištenja istog broja kod drugog pružaoca telekomunikacijskih usluga bez umanjenja kvaliteta ili pouzdanosti korištenja usluga. [7] Ovakav zahtjev je veoma strog, a trebalo bi da stupi na snagu tek nakon uvođenja centralne baze podataka o prenijetim brojevima.

Do tada će se prenosivost obavljati putem međupovezivanja, a operateri će biti dužni da osiguraju zadržavanje broja sa što je moguće manje narušavanja funkcionalnosti, kvaliteta, pouzdanosti i lakoće korištenja telekomunikacionih usluga.

Servis prenosivosti telefonskog broja je prvi put uveden prije skoro 20 godina. Velika Britanija je jedna od prvih država koja je implementirala prenosivost broja i to za fiksnu mrežu 1996. godine, a za mobilnu 1999. godine. Finska je implementirala rješenje bazirano na inteligentnoj mreži 1997. godine. Po broju prenijetih brojeva prednjače Španjolska i Italija, kod kojih je samo tokom jedne godine (listopad 2007. – listopad 2008.) preko 4 milijuna mobilnih korisnika prenijelo svoje brojeve u drugu mrežu, a ukupno preko 20 milijuna, što u Španjolskoj predstavlja 38%, u Italiji 20% svih mobilnih korisnika. Ipak, Finska ima najveći postotak prenijetih brojeva – oko 70%, a zatim slijedi Danska s 40%, pa Španjolska. U Francuskoj je od 2003. godine prenosivost iskoristilo više milijuna korisnika. U Rumunjskoj prenosivost funkcionira od listopada 2008. godine, a prenijeto je 500.000 brojeva. Vodafon u Turskoj je zbog nove usluge izgubio oko 600.000 korisnika. U Hrvatskoj i Sloveniji prenosivost funkcionira još od 2006. godine. U Sloveniji je prenijeto oko 6% brojeva. Prenosivost broja u Mađarskoj funkcionira od svibnja 2004. godine i iskorišten je od strane 300.000 korisnika, oko 2% brojeva (od ukupno 12 miliona). U 27 država Europske unije 10,3% svih korisnika mobilne telefonije prenijelo je svoj broj kod drugog operatera. [5]

U SAD-u mogućnost prenošenja brojeva prilikom prelaska kod novog operatera iskoristilo je samo pet od sto korisnika. Japan je prenosivost brojeva uveo u listopadu 2006. godine [7]

Podaci pokazuju da prenosivost brojeva postoji u oko 75 država širom svijeta. Osnovni problemi u prenosivosti brojeva bili su troškovi prijenosa broja koji su u nekim državama išli i do nekoliko stotina eura (Vlada Indije to pitanje riješila je ograničavanjem tarife za prijenos i dug proces prenošenja). I sami operateri učinili su prenosivost broja manje atraktivnim jer vezuju pretplatnike dugoročnim ugovorima koji podrazumijevaju plaćanje penala ako odluče da pređu u drugu mrežu.

IV. PRENOSIVNOST TELEFONSKIH BROJEVA U BOSNI I HERCEGOVINI

Institut prenosivosti telefonskih brojeva nije izričito reguliran Zakonom o komunikacijama Bosne i Hercegovine iz 2003. godine. [7] Međutim, imajući u vidu da je interkonekcija (međupovezivanje) jedan od načina sprovođenja prenosivosti telefonskih broja, regulirana odredbama navedenog zakona, postavlja se pitanje je li regulativa interkonekcije dovoljna da bi se smatralo da je u našem pravu posredno uređen i institut prenosivosti brojeva. Regulatornoj agenciji za komunikacije Bosne i Hercegovine dana je nadležnost za upravljanje planovima numeracije i da po zahtjevu dodjeljuje elemente adresiranja operaterima telekomunikacija na upotrebu. Operateri kojima su od strane Agencije za komunikacije dodijeljeni brojevi, te brojeve dalje daju na korištenje svojim krajnjim pretplatnicima. Kada je riječ o fiksnim telefonskim uslugama, operater dodjeljuje broj u vezi s određenim fiksnim telefonskim priključkom. [7]

Pravilom 32/2008 o prenosivosti telefonskih brojeva, omogućeno je pretplatnicima u javnoj telefonskoj mreži Bosne i Hercegovine da zadržavaju svoj/svoje brojeve nezavisno od promjene lokacije, servisa i operatera telekomunikacija. Broj koji je pretplatnik dobio se prenosi k drugom operateru, tako da osobe koje pozivaju

pretplatnika, koji je prenio broj drugom operateru, biraju isti broj kao i prije toga. [8]

Pravilom o prenosivosti telefonskih brojeva, definiran je regulatorni okvir za uvođenje usluge prenosivosti telefonskih brojeva na tržištu telekomunikacija Bosne i Hercegovine. Ovom uslugom nije omogućen prijenos broja iz fiksne u mobilnu mrežu i obratno. Pravilom je predviđeno da je za uslugu prenosivosti telefonskih brojeva potrebno uspostaviti centralnu administrativnu bazu podataka prenesenih brojeva, što je zajednička obaveza licenciranih operatera telekomunikacija. Primjena ovog Pravila omogućuje svakom pretplatniku, da promjenom lokacije, operatera ili servisa upotrebljava isti telefonski broj. [8]

Usluge Upravitelja centralne administrativne baze podataka prenesenih brojeva obavlja firma QSS d.o.o. Sarajevo, kojoj je 1. lipnja 2010. godine dodijeljena dozvola za pružanje usluge Upravitelja centralne administrativne baze podataka prenesenih brojeva mobilne i fiksne telefonije u Bosni i Hercegovini. Sistem prenosivosti broja omogućava pretplatnicima fiksnih i mobilnih mreža da zadrže svoj broj telefona prilikom promjene operatera. Implementacija Kapsch sistema prenosivosti broja je unaprijedila liberalizaciju telekomunikacijskog tržišta u Bosni i Hercegovini. Odabirom upravitelja centralne administrativne baze podataka prenesenih brojeva, stekli su se uvjeti da svi telekom operateri koji pružaju usluge telefonije u Bosni i Hercegovini izvrše tehničke pripreme i omoguće pretplatnicima da koriste uslugu prenosivosti telefonskih brojeva.

U Bosni i Hercegovini usluga prenosivosti telefonskih brojeva za pretplatnike fiksnih javnih telefonskih usluga, u okviru iste mrežne grupe, u primjeni je od 15. travnja 2011. godine. Usluga prenosivosti telefonskih brojeva omogućava pretplatnicima koji koriste javne govorne telefonske usluge u fiksnoj telefonskoj mreži da, na nivou mrežne grupe uz zadržavanje postojećeg telefonskog broja izaberu željenog telekom operatora čije su usluge dostupne u istoj mrežnoj grupi. Izbor novog operatera se vrši na način da pretplatnik podnosi pismeni zahtjev novoizabranom operateru, koji u istoj mrežnoj grupi ima mogućnost da priključi pretplatnika na svoju govornu telefonsku mrežu. Za svaki prenešeni broj, operater primatelj broja će morati plaćati nadoknadu operateru davaocu broja. [8] Prenosivost brojeva u mobilnoj telefoniji u Bosni i Hercegovini primjenjuje se od 1. siječnja 2013. godine. [7]

V. ZAKLJUČAK

U radu je razmatran značaj usluge prenosivosti brojeva u javnoj telefonskoj mreži i dan je pregled stanja na planu primjene ove usluge u Europi i Bosni i Hercegovini. Institut prenosivosti broja je veoma značajan iz ugla interesa pretplatnika i kao instrument poticanja razvoja konkurencije na liberaliziranim telekomunikacijskim tržištima. Njegovom specifičnom pravnom uređenju ne postoji druga odgovarajuća zamjena. Usluga prenosivosti telefonskog broja u javnoj telefonskoj mreži pruža slobodu izbora korisnicima telefonskih usluga pri čemu oni zadržavaju svoj stari broj obzirom da mijenjanje telefonskog broja usporava ili u potpunosti prekida komunikaciju između ljudi.

Države Europske unije su uvele ovu uslugu tijekom posljednjih petnaestak godina s različitom dinamikom i koristeći različite modele za implementaciju. Međutim, evidentan je proces harmonizacije ovog servisa u državama Europske unije zahvaljujući pomjeranju k zajedničkom tehničkom rješenju i usaglašavanju administrativnih pitanja.

Pravilom o prenosivosti telefonskih brojeva, definiranje regulatorni okvir za uvođenje usluge prenosivosti telefonskih brojeva na tržištu telekomunikacija Bosne i Hercegovine. Ova usluga omogućava korisnicima telekom usluga da biraju željenog telekom operatera uz mogućnost zadržavanja postojećeg telefonskog broja. Prenosivost telefonskih brojeva u javnoj telefonskoj mreži u Bosni i Hercegovini je još jedan od koraka u globalizaciji tržišta u telekomunikacijskom sektoru.

LITERATURA

[1] Sprinzl, M.: Mobile Nummernportierung in Österreich – die Realisiserung eines regulatorischen Konzeptes, u: Lichtenberger Ewald, Ruhle Ernst-Olav (Hrsg.), Die Regulierung der österreichischen Telekommunikationsmärkte neuen Rechtsrahmen, Regulierungsrecht und Regulierungsökonomie, Band 5, herausgegeben von Piepenbrock-Schuster, Düsseldorf, Köln, 1. Auflage, 2005.

[2] Marković, M.: Prenosivost telefonskog broja u telekomunikacionom pravu. Pravni fakultet Univerziteta u Beogradu, 2007.

[3] Universal Service Directive 2002/22/EC, on universal service and users' rights relating to electronic communications networks and services (Universal Service Directive)", Parliamentand the Council, Official Journal L 108/51, March 2002.

[4] "Progress report on the single european electronic communications market 2007", Brussels, 19/03/2008.

[5] "Progress report on the single European electronic communications market 2008, Commission of the European communities, Brussels, 24/03/2009.

[6] "Implementation of mobile number portability in CEPT countries", ECC report 31, CEPT, Updated: October 2005.

[7] Colleen, B.: Number Portability for Consumers: Taking Your Wireless Number With You, Loyola Consumer Law Review, No. 16, 2004.

[8] Zakon o komunikacijama Bosne i Hercegovine („Službeni glasnik BiH", broj 33/02).

[9] Pravilo 32/2008 o prenosivosti telefonskih brojeva ("Službeni glasnik BIH", broj 20/08).

WIRELESS SENSOR NETWORKS

Effective Energy Consumption on Wireless Sensor Networks: Survey and Challenges

Anuparp Boonsongsrikul*, Slavko Kocijancic** and Somjet Suppharangsan*
* Burapha University, Faculty of Engineering, Department of Electrical Engineering, Chonburi, Thailand
** University of Ljubljana, Faculty of Education, Department of Physics and Technology, Slovenia
anuparp@eng.buu.ac.th

Abstract – A wireless sensor network consists of a group of embedded devices called sensor nodes reading, computing and reporting data to the base station. The applications of wireless sensor networks are promising to make our lives more convenient. One of the great issues facing sensor node is the energy constraint due to only having a small battery as their energy. Efficiency of energy consumption and prolonging network lifetime are challenge. This survey paper aims to collect and classify contemporary approaches of effective energy consumption in wireless sensor networks.

I. INTRODUCTION

A wireless sensor network is a modern technology that consists of a number of sensor nodes and the base station. Sensor nodes read and send data to the base station. The base station collects and summarizes data. Earlier, a wireless sensor network was exploited for military surveillance in particular detecting enemies in a battlefield. During the peacetime, wireless sensor networks have become a promising technology that can be applied in multidisciplinary fields. In wireless sensor networks, researchers proposed a variety of communication protocols, topologies and mechanisms for particular applications such as detecting fires in jungles, forecasting weather, monitoring patients' movement and tracking animals.

One of the greatest issues that wireless sensor networks faced is energy constraint. Since sensor nodes have a small size, small memory, small battery and tiny central processor unit. In wireless sensor networks where sensor nodes are stationary, communication protocols between sensor nodes should be efficient in order to prolong network lifetime. Data aggregation [1] [2] [3] [4] is one of effective approaches that reduce redundant data sent by sensor nodes to the base station. In work [5] [6] [7], many researchers proposed a topology construction after sensor nodes are deployed. Their simulation results illustrated that balanced topologies assist to minimize data collision and to prolong network lifetime. In wireless sensor networks, research [8] and [9] have tried to harvest energy from sensor nodes. Harvesting energy in wireless sensor networks is active research. However, wireless sensor networks are resource-constrained, complicated protocols may increase extra burdens to sensor nodes and result in reducing network lifetime. It is challenging to prolong energy-power and network lifetime of wireless sensor networks. This survey paper aims to collect and classify contemporary approaches of effective energy consumption and to discuss the proposed one of effective energy consumption methods in wireless sensor networks.

The rest of this paper can be divided as follows. Section II presents related works. Section III describes classification of efficient energy consumption approaches in wireless sensor networks. Section IV discusses possibilities and challenges. Section V finally concludes our work.

II. RELATED WORK

There are many survey papers related to wireless sensor networks. This section describes existing survey papers and articles related to our survey papers. Sensor networks are often referred as ubiquitous computing [11] that can be applied to many fields therefore requirements of quality and focused aspects are different. For quality of data, a survey paper [12] discussed about outlier detection based on statistical or data mining techniques. The quality of data refers to accuracy and precision [13]. Quality of Services (QoS) is active research as well. For high bandwidth demands, work [14] investigated QoS of wireless sensor networks while work [15] surveyed multimedia sensor networks. Due to the development of CMOS cameras and microphones which can wirelessly interconnect [16], a survey paper [17] focused on the video encoding at the video sensors and the real-time transport of the encoded video to a base station. Latency and throughput in multimedia sensor networks are the first priority for performance evaluation. For media access control (MAC) in wireless communications, sensor nodes possibly send messages at the same time and result in data collision. A survey paper [18] collected effective media access control protocols that reduce data collision. A survey paper [19] intensively reviewed routing protocols in wireless sensor networks. However, aforementioned survey papers focused on other different aspects whereas our work is mainly concerned about effective energy consumption in wireless sensor networks. This motivates us to review literatures relevant to energy consumption in wireless sensor networks.

There is a diversity of works related to efficient energy consumption in wireless sensor networks. In work [1] [2] and [3], authors focused on efficient energy consumption in delivering data from sensor nodes to the base station. In particular, data aggregation protocols eliminate redundant sensor nodes' data sent to the base station. In work [5] [6]

This research was supported by Faculty of Engineering, Burapha University, Thailand under the contract number 15/2555.

[7], authors proposed a new topology construction for sensor network topologies in order to minimize bottleneck of data transmission and reception. In work [8] and [9], authors aimed to develop energy harvesting in wireless sensor networks. Even though there are different methodologies in related works but they have the same ultimate goal which is efficient energy consumption in wireless sensor networks.

III. CLASSIFICATION OF EFFICIENT ENERGY CONSUMPTION APPROACHES IN WIRELESS SENSOR NETWORKS

This section describes efficient energy consumption approaches that can be classified into three groups based on contemporarily active research: data aggregation, topology construction and energy harvesting in wireless sensor networks as shown in Fig. 1.

A. Data Aggregation

Data aggregation is a technique to eliminate redundant data and reduce energy consumption. There are basically three types of data aggregation comprising of tree topologies, cluster topologies and a hybrid of tree and cluster topologies.

An example of tree topologies is the Tiny Aggregation (TAG) [1] in which the base station initiates a session of data collection by broadcasting flooding messages to all sensor nodes. External nodes send sensed data by hop-by-hop to the base station via their internal nodes that aggregate data based on functions such as count, min, max, sum and average. The communication overhead is *O(n)* where *n* is the total number of sensor nodes.

In cluster topologies, the Low-Energy Adaptive Clustering Hierarchy (LEACH) protocol [2] is a representative that sensor nodes are divided into multiple clusters. Each cluster consists of cluster members and a cluster head. Since a cluster head plays an important role to aggregate data and sends aggregated data with a single hop to the base station while cluster members only send sensed data to their cluster head, a cluster head must be first exhausted comparing with its cluster members. This paper proposed an algorithm to select a cluster head and to change a cluster head when a cluster head is exhausted.

In a hybrid of tree and cluster topologies, a Secure Hop by Hop Data Aggregation protocol (SDAP) [3] is an example that sensor nodes are divided into multiple balanced clusters and their topology actually is a tree. This protocol supports secure data aggregation [4] and scalability. When a cluster head completely aggregates data sent by its cluster members, it sends aggregated data to the base station via internal nodes. The communication overhead depends upon the average cluster size, *c*. If *c* is as large as n, the overhead is about *O(n)*. Otherwise, if *c* is small and can be considered as a constant number, the overhead is *O(n log n)*. In either case, the communication overhead is less than the no-aggregation approach and higher than the hop-by-hop aggregation approach. However, data aggregation works normally under the assumption that a wireless sensor network topology is already established and well balanced. The effective design of a topology construction is challenging and described in a next section.

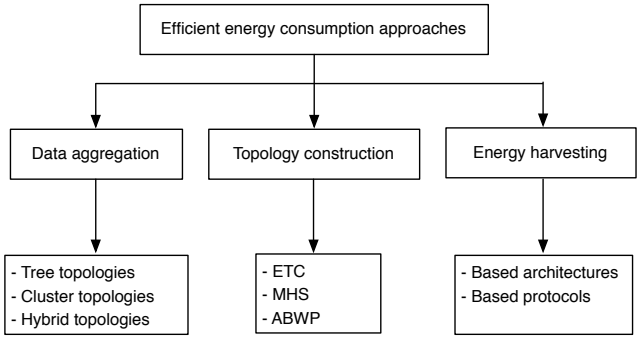


Figure 1. Classification of efficient energy consumption in wireless sensor networks

B. Topology Construction

In some applications such as data acquisition, a construction process of balanced topologies is one of active research. Unlike data aggregation, the applications require all sensor nodes' data to be collected at the base station. In other words, sensed data cannot be integrated as the same data aggregation approaches. Internal nodes have to resend or forward sensor nodes' data. An unbalanced topology leads to increase data collisions. An internal node has to repeatedly send the same data until it succeeds. This results in consuming a lot of energy to resend data and decreasing network lifetime. This section describes papers relevant to a balanced topology construction.

An Energy-driven Tree Construction approach (ETC) [5] was proposed to build a balanced tree that can minimize data collision. It consists of two steps in building the query routing tree. First is to construct an arbitrary tree and; second is to balance node degrees of internal nodes.

In the first phase, the sink node (the base station) broadcasts a query message to build an arbitrary tree. Each node rebroadcasts a query message. Each sensor s_i selects one of its neighbors as a candidate parent.

In the second phase, the sink uses the highest depth of the tree d, and the total number of nodes n, to calculate node degree, $\beta = \sqrt[d]{n}$. For example, *n* equals to 80 and *d* equals to 8 then β will be 1.729. Therefore for an integer number, the node degree can be approximately 2. In this paper, a balanced tree is a tree where the heights of the children of each internal node differ at most by one. A possible balanced tree of ETC can be illustrated in Fig. 2 where $n = 80$, $d = 8$ and $\beta = 2$.

A Minimum-Hot-Spot Query Trees (MHS) approach [6] was introduced to minimize bottleneck in a tree topology. During constructing a tree, the order in which the nodes pick their parent importantly affects the quality of the outcome of minimizing hot spots. Nodes with many incoming streams are primary candidates for network hotspots. Nodes with more candidate parents should pick a parent at a later time slot while nodes with a few candidate parents will pick a parent earlier. The proposed tree outperforms competition and achieves balanced tree. The majority of internal nodes have the degree equal to one.

Even the simulation results of algorithms in work [5] and [6] can construct balanced trees but both works do not have a mechanism to prolong a sensors life when a node runs nearly out of energy. In particular, a node (black node in Fig.2) nearly located at the sink (bottom left in Fig. 2) does not balance the number of descendants with its neighbors. This motivates authors [7] to propose a novel protocol for adaptively balancing workload on query trees for maximizing the network lifetime in sensor networks.

An Adaptively Balancing Workload Protocol (ABWP) [7] was proposed in order to reduce the burdens of a node closed to the base station and result in maximizing lifetime in sensor networks. The protocol consists of two phases: construction of a balanced query tree and reconstruction of the tree during service time.

In the first phase, the base station uses breadth-first search in order to select its neighbors to be its children. Among children the base station selects two nodes which have the first and second maximum number of neighboring nodes. Two children selected by the base station become Subtree Heads (SH) and perform balancing workload.

In the second phase, when internal node *L* has residual energy less than the threshold, then it sends a help-me message to inform its upper internal node M (its parent). Neighboring nodes having residual energy greater than the threshold and overhearing a help-me message will sends a willing message to node *M*. Then node *M* lets a neighboring node which has the greatest residual energy be a new internal node instead of node *L*. The balance index [10], $\beta(E)$ is a fairness index among loads in subtrees defined as follows.

$$\beta(E) = \frac{(\sum K_{SHi})^2}{x\sum K_{SHi}^2}$$

where SH_i is Subtree Head of (top) subtree *i*, K_{SHi} is the total number of workloads of SH_i, and *x* is the number of subtrees.

The simulation result [7] was illustrated that $\beta(E)$ of ETC, MHS and ABWP are 0.77, 0.74 and 1 on average respectively. After running 2,100 rounds of data delivery, the node by the ABWP scheme still remains 20% of initial energy while the one of the MHS scheme is fully exhausted.

C. Energy Havesting

Energy harvesting in wireless sensor networks means gathering energy from surrounding environment or other energy sources and converting it to electrical energy of sensor nodes. Contemporary approaches relevant to energy harvesting can be categorized into two groups: based architectures and based protocols.

In the based architectures, the energy harvesting architecture refers to a sensor node that requires extra hardware such as solar cells, piezoelectric materials to convert other energy sources to harvested electrical energy. In addition, an energy-harvesting sensor node also requires energy storage devices such as recharged batteries or storage capacitors. In work [8], authors studied

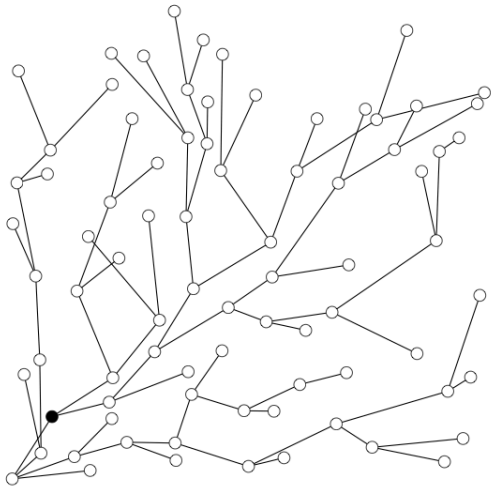


Figure 2. A possible balanced tree of ETC

on lifetime maximization in a single cluster topology in wireless sensor networks. This work uses the same energy model as LEACH [2]. Instead a cluster head sends aggregated data with a single hop to the base station, it sends aggregated data to the base station via an energy-harvesting sensor node having a solar cell and performing like an intermediate node between the base station and a cluster head. Simulation results illustrated that the network lifetime increased up to 8.59% compared with the case without an energy-harvesting sensor node. For a wireless sensor node adding a storage capacitor and a solar cell [20], its storage power is slightly greater than the storage power of a sensor node without a storage capacitor. This topology consists of sensor nodes and a central coordinator. To prolong sensor lifetime, sensor nodes are able to control local scheduling for sleep periods while the only central coordinator can control global scheduling for sleep periods of the entire network. In press [21], a kinetic converter is a power generator of a sensor node applying for tracking animals.

In the based protocols, an energy harvesting protocol refers to a communication protocol relevant to a sensor node being able to harness energy. In work [9], the authors proposed schemes that allow sensor nodes to adjust duty-cycle based on residual energy. Since a high sleep latency causes a large end-to-end delay in cases where multi-hop forwarding of a packet is required and leads to packet collision [9], the sleep latency, duty-cycle should be as small as possible in order to avoid a large end-to-end delay and packet collision. However, each node has different residual energy. Therefore, duty-cycle should be adjusted based on residual energy in order to reduce the sleep latency and to balance energy consumption among sensor nodes.

Even though effective energy consumption approaches have different methodologies and performance indexes, we observe that there are some common aspects such as data types, data delivery from a sensor node to the base station, requirements of extra hardware, balancing topologies and scalability in wireless sensor networks. We summarize and compare the aspects of the effective energy consumption approaches and present their feasible applications in Table I.

TABLE I. CLASSIFICATION AND COMPARISION OF EFFECTIVE ENERGY CONSUMPTION APPROACHES IN WIRELESS SENSOR NETWORKS

Approaches	Classification	Common aspects					Feasible applications
		Data types	Data delivery	Extra hardware	Balancing topologies	Scalability	
Madden [1]	Data aggregation	Aggregate	Multi-hop	No	No	No	Weather monitoring
Heinzelman [2]	Data aggregation	Aggregate	2 hops	No	Yes	Yes	Weather monitoring
Yang [3]	Data aggregation	Aggregate	Multi-hop	No	Yes	Yes	Fire detection
Andreou [5]	Topology construction	Raw data	Multi-hop	No	Yes	Yes	Habitat monitoring [22]
Chatzimilioudis [6]	Topology construction	Raw data	Multi-hop	No	Yes	Yes	Votree environment detection [23]
Boonsongsrikul [7]	Topology construction	Raw data	Multi-hop	No	Yes	Yes	Petroleum industrial [24]
Zhang [8]	Energy harvesting	Raw data	3 hops	Solar cell	Yes	No	Weather monitoring
Gyorke [20]	Energy harvesting	Raw data	A single hop	Solar cell	Yes	No	Home health monitoring [20]
Yoo [9]	Energy harvesting	Raw data	N/A	Solar cell	Yes	No	Habitat monitoring
GutierrezMartin [21]	Energy harvesting	Raw data	N/A	Kinetic convertor	No	Yes	Tracking animals

IV. POSSIBILITY AND CHALLENGE

A wireless sensor network normally has a specific topology to a specific application. When an application is changed, the network topology might have to change corresponding to the application. However, when a topology is reconstructed and unbalanced, sensor nodes waste their energy and their lifetime were shorten. It is challenging to prolong lifetime and minimize energy consumption in reconstruction of a network topology that supports multiple applications such as data aggregation and data acquisition. In the following subsections, we briefly discuss possibilities and challenges of our proposal for effective energy consumption in wireless sensor networks.

A. Energy management

Sensor nodes located nearly at the base station should be replaced with energy-harvesting sensor nodes, because they can recharge energy and prolong their lifetime. Several approaches grant sensor nodes to measure their residual energy. These sensor nodes possibly might have additional hardware [25]. Intuitively when more electrical circuits are added in a sensor node, the more energy consumption is required. To prolong lifetime, rather than using only electrical circuits to measure residual energy, a sensor node should be able to evaluate its remaining energy by learning, reasoning and information shared among multiple nodes. This could be achieved using cognition in wireless sensor networks [26].

B. Topology management

The key idea of topology control is to allow sensor nodes to control over certain parameters [27]. A sensor node should know distances from its neighboring nodes to itself and be able to adjust its transmission range according to distances. As a result, it can minimize the power incurred during transmission. Knowing sensor nodes' location could be succeeded by the localization technique [28].

C. Quality of Services (QoS)

The tradeoff between energy consumption and QoS in terms of reliable data delivery is active research [29]. It is challenging to optimize the network performances that minimizing energy consumption and in the meanwhile guaranteeing QoS. Data delivery of one-directional communication such as a User Datagram Protocol (UDP) is quite unreliable but requires less transmission energy whereas that of bidirectional communication such as a Transmission Control Protocol (TCP) is more reliable but desires more transmission energy. To minimize energy consumption and give reliability with one-directional communication, the possibility is that a sensor network may be exploited by energy-harvesting sensor nodes to be monitoring nodes. If monitored data are not sent or forwarded by an intermediate node, monitoring nodes will send such monitored data instead. This proposal could be done using the concept of work [30].

V. CONCLUSION

This paper presents a diversity of efficient energy consumption approaches in wireless sensor networks. Active research of efficient energy consumption is divided into data aggregation, tree construction and energy harvesting in wireless sensor networks. This paper also provides comparison, possibilities and challenges of effective energy consumption approaches in a variety of applications by wireless sensor networks. Even the active research has different ways of dealing with energy consumption but there is the same proposal that is prolonging residual energy of sensor nodes.

ACKNOWLEDGMENT

This research was supported by Faculty of Engineering, Burapha University, Thailand under the contract number 15/2555. The authors gratefully acknowledge the contributions of all committee members for their valuable suggestions.

REFERENCES

[1] S. Madden, M. J Franklin, J.M. Hellerstein and W. Hong, "TAG: A tiny aggregation service for ad-hoc sensor networks," ACM SIGOPS Operating Systems Review, 36(SI), 2002, pp.131-146.

[2] W. B. Heinzelman, A. P. Chanrakasan, and Balakrishnan, "An application-Spefific Protocol Architecture for Wireless Microsensor," IEEE trans, Wirelss Commun, Vol 1, no.4, pp. 660-670, 2002.

[3] Y. Yang, X. Wang, S. Zhu, and G. Cao, "SDAP: A Secure Hop-by- Hop Data Aggregation Protocol for Sensor Networks," the Seventh ACM International Symposium on Mobile Ad Hoc Networking and Computing(MobiHoc'06), 2006, pp. 356-367.

[4] A. Boonsongsrikul, K. suk Lhee, and M. Hong, "Securing Data Ag- gregation against False Data Injection in Wireless Sensor Networks," in Proceedings of the 12th International Conference on Advanced Communication Technology (ICACT2010), 2010, pp. 29–34.

[5] P. Andreou, A. Pamboris, D. Zeinalipour, P. Chrysanthis, and G. Sama- ras, "ETC: Energy-Driven Tree Construction in Wireless Sensor Networks," in Proceedings of Mobile Data Management: Systems, Services and Middleware (MDM '09), 2009, pp. 513–518.

[6] G. Chatzimilioudis, D. Zeinalipour-Yazt, and D. Gunopulos, "Minimum-Hot-Spot Query Trees for Wireless Sensor Networks," in Proceedings of the 9th ACM International Workshop on Data Engineering for Wireless and Mobile Access (MobiDE'10), 2010.

[7] A. Boonsongsrikul, S. K. Park, and S.H. Shin, "An Adaptively Balancing Workload Protocol on Query Trees for Maximizing Lifetime in Sensor Networks," in Proceedings of the 12th International Conference on Advanced Communication Technology (ICACT2012), 2012, pp. 195–200.

[8] P. Zhang, G. Xiao, and H.P Tan, "A Preliminary Study on Lifetime Maximization in Clustered Wireless Sensor Networks with Energy Harvesting nodes", in Proceedings of the 8th International Conference on Information, Communications & Signal Processing (ICICS), 2011, pp. 1-5.

[9] H. Yoo, M. Shim and D. Kim, "Dynamic Duty-Cycle Scheduling Schemes for Energy-Harvesting Wireless Sensor Networks, in IEEE Communication letters, Vol. 16, No. 2, 2012 pp. 202- 204.

[10] P.-H. Hsiao, A. Hwang, H. Kung, and D. Vlah, "Load-Balancing Routing for Wireless Access Networks," in Proceedings of IEEE International Conference on Computer Communications (INFOCOM'01), 2001, pp. 986–995.

[11] W. Brunette, J. Lester, A. Rea and G. Borriello, "Some sensor network elements for ubiquitous computing," Information Processing in Sensor Networks, 2005. IPSN 2005. Fourth International Symposium on. IEEE, 2005.

[12] Y. Zhang, N. Meratnia and P. Havinga, "Outlier Detection Techniques for Wireless Sensor Networks: A Survey," IEEE Communications Surveys & Tutorials, Vol. 12 No. 2, pp. 159 - 169, 2010.

[13] M. M. Gaber, Data Stream Processing in Sensor Networks. In J. Gama and M. M. Gaber, Learning from Data Streams Processing Techniques in Sensor Network, pp. 41-48. Springer Berlin Heidelberg, 2007.

[14] Chen, Dazhi, and Pramod K. Varshney. "QoS support in wireless sensor networks: A survey," International Conference on Wireless Networks. Vol. 233, 2004.

[15] A, Ian F, M. Tommaso and C., Kaushik R, "Wireless multimedia sensor networks: A survey," Journal of IEEE Wireless Communications, Vol.14, Issue 6, pp. 32 – 39, 2007.

[16] Vithya, G. and Vinayagasundaram, B. "Actuation sensor with adaptive routing and QOS aware checkpoint arrangement on Wireless Multimedia Sensor Network," Proceeding of International conference on Recent Trends in Information Technology (ICRTIT), 2011 pp. 444 - 449.

[17] M. Satyajayant, R. Martin and X. Guoliang Larry, "A survey of multimedia streaming in wireless sensor networks," IEEE Communications Surveys & Tutorials, Vol.10 , pp. 18 – 39, 2008.

[18] P. Huang, L. Xiao, S. Soltani, M. W. Mutka and N. Xi "The Evolution of MAC Protocols in Wireless Sensor Networks: A Survey" IEEE Communications Surveys & Tutorials, Vol. 15 No. 2, pp. 101 - 119, 2013.

[19] J. N. Al-Karaki and A. E. Kamal, "Routing Techniques in Wireless Sensor Networks: A Survey, "Journal of IEEE Wireless Communication" Vol. 11, Issue 6, pp. 6 – 28, 2004.

[20] P. Gyorke and B. Pataki, "Application of energy-harvesting in wireless sensor networks using predictive scheduling," Proceeding of IEEE international conference on Instrumentation and Measurement Technology Conference (I2MTC), 2012.

[21] A. GutierrezMartin, N. Dopico, C. Gonzalez, S. Zazo, J. Jimenez-Leube and J. Raos, "Cattle-powered nodes experience in a heterogeneous network for localization of herds," in press.

[22] Szewczyk R., Mainwaring A., Polastre J., Anderson J and Culler D., "An Analysis of a Large Scale Habitat Monitoring Application", In ACM SenSys, 2004.

[23] Voltree-Power-Incorporated. http://www.voltreepower.com/

[24] Lewis and Adam, "Wireless Sensor Data Acquisition Design for Engineering Applications," Proceeding of 4th IEEE International conference on Circuits and Systems for Communications, 2008. (ICCSC 2008) 26-28 May 2008, pp. 544 - 549.

[25] M. Levasseur, S. Jomphe, D. Sicard, A. Dulipovici, F. Gagnon and G. Gagnon, "A Wireless Sensor Network for Residential Electrical Energy Consumption Measurement," Proceeding of 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2012.

[26] G. Vijay, E. Bdira, M. Ibnkahla, "Cognition in Wireless Sensor Networks: A Perspective," IEEE Sensors Journal, Vol. 11, Issue 3 , pp. 582 - 592, 2011.

[27] A. Aziz, Y. Ahmet S¸ekercio˘glu, P. Fitzpatrick, and M. Ivanovich, "A Survey on Distributed Topology Control Techniques for Extending the Lifetime of Battery Powered Wireless Sensor Networks" IEEE Communications Surveys & Tutorials, Vol. 15 No. 1, pp. 121 - 144, 2013.

[28] B. Cheng, R. Du, B. Yang, W. Yu, C. Chen and X. Guan, "An Accurate GPS-Based Localization in Wireless Sensor Networks: A GM-WLS Method," Proceeding of 41th International Conference on Parallel Processing Workshops (ICPPW), 2011, pp. 33- 41.

[29] H. Al-Hamadi and I. Chen, "Energy vs. QoS Tradeoff Analysis of Multipath Routing Protocols for Intrusion Tolerance in Heterogeneous Wireless Sensor Networks" Proceeding of IEEE 10th International Symposium on Parallel and Distributed Processing with Applications (ISPA), 2012, pp. 387- 394.

[30] A. Silva, M. Martins, B. Rocha, A. Loureiro, L. Ruiz and H. Wong, "Decentralized intrusion detection in wireless sensor networks," Proceedings of the 1st ACM international workshop on Quality of service & security in wireless and mobile networks, 2005, pp. 16-23.

An optimal Algorithm based on gridding and clustering for decrease energy consumption in WSN

Alireza Firuzbakht*, Asgarali Bouyer**
Azarbaijan Shahid Madani University/ Department of Information Technology, Tabriz, Iran
* A.Firuzbakht@gmail.com
** a.bouyer@azaruniv.edu

Abstract. A wireless sensor network is a deployment of massive numbers of small, inexpensive, low energy devices that can sense, compute and communicate with other devices for the purpose of gathering local information from a physical environment. Wireless Sensor Networks are tightly constrained in energy, thus require careful resource management. We have clustering based protocols such as LEACH, which gather data from sensors of clusters and tried to decrease data transmission and energy consumption to reduce data redundancy in each cluster. But aggregation operation in clustering uses only some functions which lonely can't reduce data redundancy alone. This proposed approach emphasizes on increasing network lifetime which reduces data redundancy via clustering algorithms and gridding.

Keywords: wireless sensor networks, Energy management, grid based, cluster, sleep and wakeup, data redundancy

I. INTRODUCTION

Wireless sensor networks are wireless networks consisting of autonomous sensor nodes, communicating with each other over wireless links. Each node in a sensor network has a central processing unit, memory, a Radio Frequency transceiver, and a power source [1]. The main task of the sensor node is to collect data in regular intervals and to convert it into an electronic and emission signal to the sink node by wireless communication media. While nodes' energy is limited, sending and receiving data consume most of it. So energy efficient routing strategies are essential to prolong network lifetime. There are so many ways to reduce energy consumption in sensor networks, each emphasizing on certain points.

Some of these methods are classified, according to the designation of the protocol stack layers. For example, may be related to, protocols reduce energy consumption in MAC layer or routing protocol in networking layer. Several energy conserving mechanisms have been proposed to extend the lifetime of the network, such as Span [2], Geographic Adaptive Fidelity (GAF) [3], Sparse Topology and Energy Management (STEM) [4], Adaptive Self-Configuring sensor Networks Topologies (ASCENT) [5], Cluster -based Energy Conservation (CEC) [6], Adaptive Fidelity Energy - Conserving Algorithm (AFECA) [6], and LEACH(Low Energy Adaptive Clustering Hierarchy)[7].

Among the most important protocols are clustering based one. Essentially, there is several protocol and algorithms to improve this method. LEACH is the important algorithm that is based on other ones. A WSN using LEACH is built on several clusters, each of which has a cluster head (CH) and some non-cluster heads (Non-CHs). The CH controls all the sensor nodes within the cluster, fuses data sent by the sensor nodes, and forwards them to BS. Non-CHs, on the other hand, collect data and send them to the CH. This idea is base on other clustering based protocol such as LEACH-C [8]. LEACH-C is a clustering based algorithm where the cluster formation is performed in a centralized and base station.

LEACH-C is efficient than LEACH, and about 40% more data transfer occurs per unit energy due to the global knowledge that the base station has about position and energy level of all sensors. Additionally, LEACH-C on the contrary of LEACH forms optimal number of clusters. Actually, it is an NP-Hard problem and LEACH-C solves this problem using simulated annealing [12]. At simulating section, we compare new algorithm (LEA3C) with both of them.

Gridding in wireless sensor is an efficient tool to determine sensors, coordinate and estimate distances. There are several protocol based on griddings such as GAF [3] and GEAR [9]. They are Location-based routing protocols that can use GPS. Today using only this type of protocols is not efficient. Also gridding can use in another part of wireless sensor networks for example: a directed grid topology from the source node to the sink node is proposed in [10].

This paper reveals an algorithm that focuses on the data redundancy and tries to remove unused data transmission. This algorithm is based on a clustering protocol that can be one of the current clustering protocols. In fact, we use current clustering protocol with gridding.

The remainder of this paper is organized as follows: Section 2 defines the problem and the expected objectives in the paper. Section 3 describes our proposed algorithm while simulations and evaluation of results for new algorithm, LEACH and LEACH-C is presented in section 4. Finally, Section 5 concludes the paper.

II. PROBLEM DEFINITION

The major challenge posed by sensor networks is the wasteful usage of resources. It is important to keep the nodes active as long as possible, and it depends on their application. Most of the times, sensors in some applications sense same data that increase useless data, data redundancy and transmission in network. So we try to remove this problem with a co-working plan. The motivation of our work is to study network applications and energy consumption in sensor networks.

This paper studies the new algorithm that removes data likeness and analyzes energy consumption to prolong the network lifetime. We focus on implementing this algorithm based on a famous protocol called LEACH. We also can use other clustering protocols; but LEACH is the base of all. By survey energy consumption in all part of WSN, our strategy will be efficient in energy management.

As already mentioned, the sensor has three general parts: processing, sensors, and radio. But energy consumption in these sectors is different. Communication subsystems (radio section) have more energy consumption than the other parts. For instance, energy required to send a bit of information is more than the energy needed to execute thousands of instructions. The radio part itself has a variety of modes and these modes are also different in their consumption.

Since the radio part is important, new methods should concentrate on reducing the amount of information exchange, data redundancy, in the network. Nodes should perform the required processing before sending the data. To solve this problem, most of protocols use gathering functions such as minima, duplicate, maxima and average. But we are trying to reduce redundancy before using these functions. It means that data aggregation will be done in two levels.

So in this paper we proposed a new algorithm based on clustering and grid based protocol to remove useless data and to increase the network lifetime called LEA3C.

III. PROPOSED ALGORITHM

In WSN topology, there are numerous sensors distributed in an environment. Fig 1 shows the initial network situation. Commonly wireless sensor network has not any topology and nodes are scatter in a physical environment, but all of them can communicate with each other and the sink.

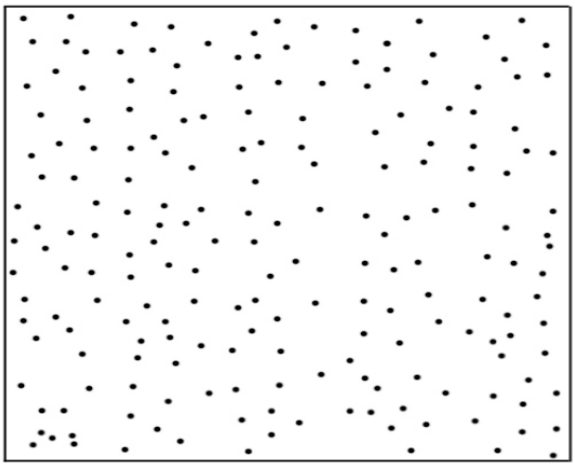


Figure 1. Initial WSN

However, each node is capable of communicating in dual mode (i.e. using low power radio and high power radio). All nodes normally communicate using high power radio, but cluster head DNs, use low power radio while communicating with nodes in an Alternate Dissemination Zone (ADZ) around it [11]. Each sensor node is aware of its own location through GPS system and coordinates *x and y*.

A. Gridding

Gridding is a best way of managing distributed node. We use a grid construction scheme where the size of each cell is determined by dual radio range of a sensor node [11]. Unlike TTDD, where source initiates grid construction, the proposed scheme starts grid construction process by keeping itself at one of crossing points (CPs) of the grid. A bi-dimensional geographical----- coordinates (x,y) of this sink thus become starting point for formation of grid of square sized cells. Let R_H and R_L be the transmission ranges of every sensor node while working in high power radio mode and low power radio mode, respectively. The cell of the grid is kept square with each side of size α set to:

$$\alpha = d / \sqrt{2} = (RH - RL) / \sqrt{2} \quad (1)$$

Where *d* is the diagonal of a cell.

All other CPs (x_i,y_i) are calculated from sink(x,y) as:

$$\{x_i = x + i \cdot \alpha,\ y_i = y + j \cdot \alpha;\ \ i,j = \pm 1, \pm 2, \pm 3, \ldots\} \quad (2)$$

With establishing a virtual grid, we access to cell of gridding as virtual square. Continually, the sink will provide the grid-ID based on the co-ordinates of the grid. But sensors that depend on a virtual square give a same ID; and value of all sensors' data of a virtual square is the same. Fig 2 shows a partitioned WSN with grid.

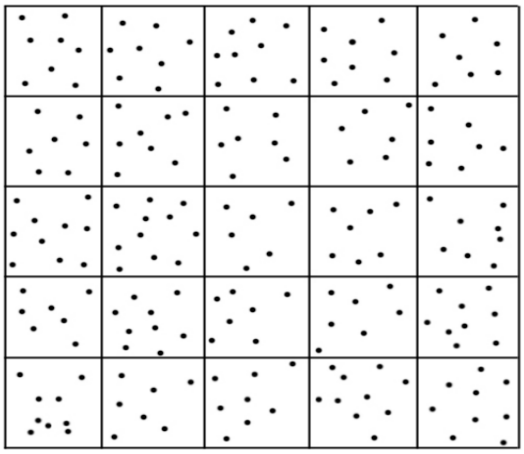


Figure 2. Partitioned WSN with grid

B. *Clustering*

Data gathering is an important phase in clustering. It means, when we depart network to the some cluster and determine a header for each cluster, data redundancy is reduced in little part of network. Our algorithm uses this feature of clustering protocols.

Different parameters have their own effect in selecting the head, but in general the following three criteria should be considered for the head:

- the maximum energy level
- close to the base station
- near to cluster center

However, the most important criterion is the energy, because it is the main criteria in selecting head with rotating selection, so its impact among the choices is greater.

C. *Final Step*

As mentioned, gridding step will primarily be done and sensors that depend on a virtual square would assign the same ID. Additionally, among them one of the nodes would become active randomly and other nodes would set to sleep mode. This process will also be done in other virtual squares. After a periodic time, the previous active nodes go to sleep mode and another node set to active mode; so sensors energy are preserved. For instance, if there are five nodes in a little square gird, energy saving is five times higher than when all of them became active. In fact, this is the first step of data gathering that removes data likeness in a network with a co-working plan. It means, sensors that are belonging to a virtual square will sense the same data respectively. As shown in Fig 3, sleep and awakening rotates among nodes of each virtual square. In each virtual square, sensors communicate with each other via R_L and active node send data to the header with R_H.

Then the header gives optimized data from nodes, in the next step it can do its data gathering by gathering functions such as minima, duplicate, maxima, and average; and subsequently sends data to the sink.

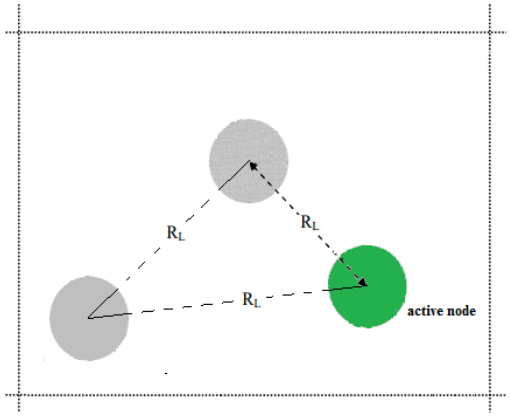


Figure 3. Sleep and wake up mode

Fig 4 shows network model for this algorithm. As shown in Fig 4, when a virtual square senses a data, it sends with a high power radio (RH) to the header of cluster and then all headers send their data to the sink with RH. So in this plan, transfers that use RH will reduce use of energy.

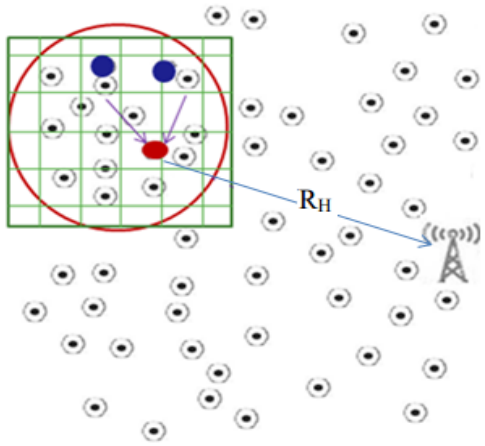


Figure 4. Network Model

IV. SIMULATION AND RESULTS

This section presents simulation results of a new algorithm, analyzing and comparing the results with LEACH and LEACH-C protocols. Simulations were carried using MATLAB software. As mentioned before, LEACH-C is efficient than LEACH, and about 40% the energy of data transfer per unit increases.

A. Parameters

Table 1 has been used for a detailed comparison of the results of the new algorithm, LEACH and LEACH-C, for a number of nodes (two environments or different scenes). Another parameter is *m*, where *m* is the number of nodes with a maximum energy level. This number can be tentatively assigned, and its value is dependent to the optimum number of clusters that we expect. In the first scene (with 100 nodes) m=80 or m=50 is assumed. As mentioned, nodes use GPS to aware their location, but once. So, we overlooked GPS energy consumption. Additionally, there are several ways to find nodes location that consume negligible energy. For example, finding the distance between neighboring nodes can be based on the input signal power. So GPS-phased energy consumption is nearly omitted and its required energy is saved. In spite of this, we can set 60 *nJ* for GPS-phased in each nodes.

TABLE 1. PARAMETERS

Parameter	Scence 1	Scence 2
N	100	400
Area	100*100	
Location of BS	(50,200)	
$d_{crossover}$	87 m	
Initial Energy	0.5 J	
E_{elec}	50 nJ/bit	
$_{fs}\varepsilon$	10 pJ/bit/m^2	
$_{mp}\varepsilon$	0.0013 pJ/bit/m^4	
EDA	5 nJ/bit/signal	
Packet Size	4000 bits	

B. Result and Analysis

This experiment is performed by using three standard criteria in sensor network routing algorithms. Where the order of death is the round number in which the death occurred:

- Time to first node death
- Since the death of half of the nodes
- Time to last node death

Time in this criterion is round in LEACH protocol that shows how many round left to nodes death.

The results are illustrated in the following tables and graphs.

TABLE II. COMPARE OF ALGORITHMS IN LIFE EXPECTANCY (FIRST SCENE)

Number of nodes =100 (Scene 1)	Algorithm	Time to first node death	Since the death of half of the nodes	Time to last node death
	LEACH	587	781	1857
	LEACH-C	614	827	1896
	New Algorithm	587	1007	1928

TABLE III. COMPARE OF ALGORITHMS IN LIFE EXPECTANCY (SECOND SCENE)

Number of nodes =400 (Scene 2)	Algorithm	Time to first node death	Since the death of half of the nodes	Time to last node death
	LEACH	713	958	2184
	LEACH-C	738	992	2208
	New Algorithm	713	1281	2216

As we see, in the above tables the superiority of the proposed algorithm (LEA3C) is clearer than LEACH and LEACH-C. But on the beginning point of this algorithm that clearness is not different a lot with others. The reason is the virtual square in which there is only one node acting as the LEACH. It means that when gridding is done, there may be only one sensor in some parts; so it consumes same energy such as LEACH. By time pass and energy maintenance in virtual squares, the advantages of this algorithm is clarified. As shown in Table 2 and 3, the needed times for death of half nodes and also all nodes significantly increase in LEACH. We essentially utilize useless redundancy of sensors and their data. So this protocol can increase the lifetime of the network and its impact in the energy consumption.

While LEACH-C is optimal than LEACH, our new algorithm (LEA3C) is better than LEACH-C. As we see in tables and graphs, the new algorithm acts like LEACH firstly. Therefore, LEACH-C is better than LEACH and our proposed algorithm at first rounds. But after few rounds, energy saves in virtual square based on our algorithm and respectively the number of death nodes is reduced.

LEACH-C is optimal than LEACH because sink has throughout knowledge about all nodes and construct optimal cluster in all rounds. Despite of this advantage, our new algorithm which has LEACH base for clustering loses less nodes rather than LEACH-C. Since this algorithm uses a two-level aggregation it helps reducing data transfer through network. So by using an optimal algorithm for clustering we would get access to expected results.

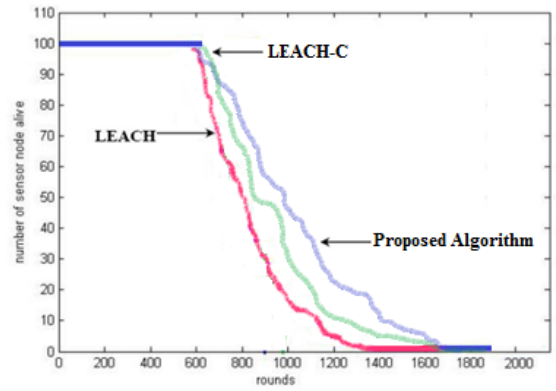


Figure 5. Compare of algorithms in terms of life expectancy in the first scene.

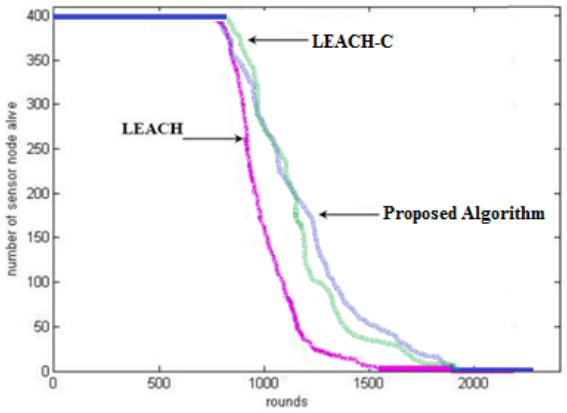


Figure 6. Compare of algorithms in terms of life expectancy in the second scene.

Another important point that is specified in the Fig 5, 6 is the uniform energy consumption throughout the network, which makes the diagram move constant. Energy consumption in new algorithm is slow. When the rounds of network increase, few sensors die and the rate of sensors death are regular. Also, LEACH and LEACH-C in middle of the rounds lose so many nodes (death nodes) that increase sensors death rate.

Experiments show that LEACH and LEACH-C consume energy not considering sensor management. It means LEACH and LEACH-C does not consider whether sensors may sense the same data or in a same physical spot. However, our new algorithm (LEA3C) identifies likeness sensors and then with a co-working plan saves energy in network.

V. CONCLUSION

The proposed algorithm was derived from clustering and the grid- based routing protocols. It follows the clustering routing protocol in data gathering and tries to optimize energy consumption with identifying sensors that sense same data besides integrating gridding and clustering protocols. For future working we can suggest to find new optimal ways to identify sensors likeness and manage them using intelligent methods.

REFERENCES

[1] Akl, R., P. Kadiyala, et al., Non-uniform grid-based routing in sensor networks. Communications (MICC), 2009 IEEE 9th Malaysia International Conference on.

[2] B. Chen, K. Jamieson, H. Balakrishnan, and R. Morris, Span: an Energy - efficient Coordination Algorithm for Topology Maintenance in Ad hoc Wireless Networks, Wireless Networks, 2002, Volume 8, pp. 481- 494.

[3] Y. Xu, J. Heidemann, and D. Estrin, Geography - informed energy conservation for ad hoc routing, Proceedings of the Seventh Annual ACM/IEEE International Conference on Mobile Computing and Networking, 2001.

[4] C. S churgers, V. Tsiatsis, and M.B.Srivastava, STEM: Topology management for energy efficient sensor networks, IEEE Aerospace Conference Proceedings, 2002, pp. 1099 – 1108.

[5] A.Cerpa and D. Estrin, ASCENT: adaptive self-configuring sensor networks topologies, IEEE Transactions on Mobile Computing, 2004, pp. 272 - 285.

[6] Y. Xu, S. Bien, Y. Mori, J. Heidemann, and D. Estrin, Topology Control Protocols to Conserve Energy in Wireless Ad hoc Networks, Center for embedded Network Sensing. Papers, 2003.

[7] Heinzelman W.R, Chandrakasan A and Balakrishnan H, LEACH: energy-efficient communication protocol for wireless microsensor network, Proceedings of the 33rd Annual Hawaii International Conference On Systems Science, 2, 2000, 3005-3014.

[8] M.J. Handy, M. Haase and D. Timmermann, LEACH-C: low energy adaptive clustering hierarchy with deterministic cluster-head selection 4th IEEE Inter-national Conference on Mobile and Wireless Communication Network, 2002.

[9] Yu Y, Estrin D, Govindan R, Geographical and Energy-Aware Routing: A Recursive Data Dissemination Protocol for Wireless Sensor Networks, In: UCLA Comp. Sci. Dept. tech. rep., UCLA-CSD TR-010023, 2001.

[10] Y. - W. Chen and C.- S. Kuo, Integrated design of grid -based routing in wireless sensor network, Advanced Information Networking and Applications, 2007, pp. 625- 631.

[11] Sharma, T. P., R. C. Joshi, et al, GBDD: Grid Based Data Dissemination in Wireless Sensor Networks, Advanced Computing and Communications, ADCOM 2008, 16th International Conference on, pp. 234- 240.

Hierarchical Clustering Using Genetic Algorithm in Wireless Sensor Networks

Mohammad M. Shurman[1], Mamoun F. Al-Mistarihi[2], Amr N. Mohammad[3], Khalid A. Darabkh[4], and Ahmad A. Ababnah[2]
[1]Jordan University of Science and Technology/Network Engineering and Security Department, Irbid, Jordan
[2]Jordan University of Science and Technology/Electrical Engineering Department, Irbid, Jordan
[3]Jordan University of Science and Technology/Computer Engineering Department, Irbid, Jordan
[4]The University of Jordan/Computer Engineering Department, Amman, Jordan
alshurman@just.edu.jo, mistarihi@just.edu.jo, ammohammad11@cit.just.edu.jo, k.darabkh@ju.edu.jo, aaababneh@just.edu.jo

***Abstract* - Wireless sensor networks consist of a large number of sensors whereas each sensor is equipped with a life-limited battery. Depletion of this energy depends on not only the long-distance communications between the sensors and sink node, but also the active state period length. Power consumption of each sensor is the most important factor that affects the network lifetime severely. In other words, decreasing the long-distance exchanges between sensors and sink node leads to prolong the network lifetime and increase the network performance. Clustering method is considered one of the best well-known approaches that efficiently helps in increasing the overall network lifetime. In this paper, we incorporate genetic algorithm (GA) with hierarchical clustering for the sake of reducing the long-distance communications. Simulation results are extremely promising and show significant improvements over heuristics and normal genetic algorithms.**

Keywords: WSN clustering hierarchy, Genetic algorithm, WSN lifetime, Power consumption.

I. INTRODUCTION

Wireless sensor network is a group of sensor devices aimed to monitor and record conditions at various locations to eventually send the sensed data to a processing center. Wireless sensor networks are widely used in many applications in which they consist of high volume of small, low cost, as well as life-limited sensor nodes. After deployment, it is not always possible to recharge or replace these sensors especially when they are deployed in inaccessible regions such as behind enemy lines, or inside a nuclear reactor [1]. Therefore, sensors energy depletion is a critical issue in wireless sensor networks. Traditional sensor network contains only one (or few) static sink(s) to gather the sensed data from sensors for further processing and decision making. Hence, all data obtained by sensors must be transmitted to the sink node in which the longer communication distance between sensors and sink, the more energy consumption is needed which in turn leads to reducing the network lifetime [2].

One of the most popular solutions to reduce the long-distance communications is clustering. In clustering, each sensor will be connected to the nearest cluster head. This cluster head is responsible for relaying the collected data to the sink. Hence, instead of sending the data directly to the sink, which requires a higher energy, the data will be relayed between the cluster heads until it reaches the sink node. Consequently, the long-distance communications will be reduced and network lifetime will be prolonged. Many algorithms can be used to perform sensor clustering in wireless sensor networks [3-5]. In our work, we used the genetic algorithm. Simulation results show that the genetic algorithm (GA) can find the solution quickly and is applicable for many network topologies (uniform or non-uniform). Fig. 1 illustrates the differences between traditional wireless sensor networks and clustered networks.

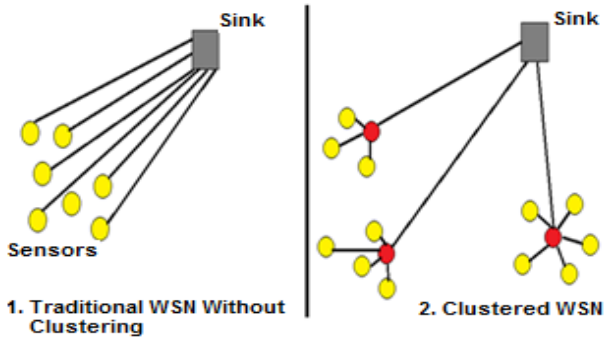


Fig. 1: Direct transmission and clustered wireless sensor networks

II. RELATED WORK

Clustering in wireless sensor network has been extensively studied by researchers. Most of them focused on the network life-time maximization by various approaches. Authors in [2] used an efficient method based on a genetic algorithm to solve sensor network optimization problem whereas each chromosome is divided into a specific number of genes according to the network size. Every gene in the chromosome has either 1 or zero, where 1 or 0 indicates that this node is a cluster head or regular node, respectively. The fitness function used in their work depends on two major factors include $Distance_i$, which refers to the summation of distances from regular nodes to cluster heads and from all cluster heads to the sink node, and H_i, which indicates the number of cluster heads available in the network. The GA should find the chromosome that has the lowest number of cluster heads and smallest $Distance_i$ as described in the following fitness function:

$$fitness = w * (D - distance_i) + (1 - w) * (N - H_i)$$

where D is the summation of distances of all nodes to the sink, N is the total number of nodes, and w is a predefined weight, ($0 \leq w \leq 1$), used to determine the algorithm

interest. For example, w=0 means the number of heads in the algorithm is only considered while w=1 means the communication distance is only considered. GA starts by randomly selecting nodes in the network to be cluster heads. Based on the fitness function, the algorithm adjusts these cluster heads. The approach proposed in [2] is an efficient and effective method to solve long-distance communications problem in which it is able to find a good solution for both uniform and non-uniform network topologies. Their results show that the number of cluster heads is about 10% of the total number of available nodes. They also achieved about 80% reduction in the long-distance communications compared with the direct transmission distance approach. This approach is considered a strong reference to this paper. In other words, we used the fitness function employed in [2] due to its efficiency in affording low computations and high improvements. It is worth mentioning that the authors in [2] did not consider the energy consumption in their simulation. In our proposed approach, we not only extend their simulations to consider the energy consumption, but also add some constrains on the genetic algorithm to obtain more necessary improvements.

Authors in [6] assumed that the network has a predetermined number of clusters. Based on the time division multiple access (TDMA), the sensor nodes transmit data to the cluster heads which in turn collect the data and transmit it to the base station. In this paper, the number of clusters is predetermined regardless of the changes made to the network. It is interesting to state that their approach is not scalable for a large number of deployed nodes.

The research idea found in [7] divides the network into equal size clusters whereas each cluster has the same number of sensor nodes. They used particle swarm optimization (PSO) approach to perform clustering. This approach provides a balance in the energy consumption for each cluster head found in the network. Unfortunately, it does not work well in the networks that have unevenly distributed sensors which results in having some clusters with a different number of nodes.

Agarwal et al. [8] used an approximation algorithm to arrange the network into a specific number of clusters (K clusters). The sensors belong to a specific cluster based on their distances to the cluster head. Heideri et al. [9] used a genetic algorithm to minimize the long-distance communications and achieve the minimum number of cluster heads for power preservation in the network. The fitness function used depends on many factors. One of them is the transfer energy (E) parameter which represents the energy consumed to transfer the aggregated message from the cluster head to the base station. The fitness function F can be represented by:

$$F = \frac{1}{E} + (TD - RCSD) + (N - TCH)$$

where TD denotes for the summation of distances of all nodes to the sink. $RCSD$ is the summation of distances of regular nodes to cluster heads and all cluster heads to the sink. N refers to the network size. TCH refers to the number of clusters. GA search for chromosome that has the largest value of F and smallest values of E and $RCSD$. The method proposed in [9] is efficient and reduces the long-distance communications which consequently reduces the energy consumption. Their proposed approach can be used in both uniform and non-uniform distribution. Furthermore, it can be used when the sensor nodes in the network are mobile.

Authors in [10] redeveloped the basic model based on LEACH protocol to calculate the energy consumptions in three phases of data transmissions from sensor nodes to the sink. It was shown through extensive simulations of this model that the overhead energy is at least 20% of the network's total energy consumed in data transmissions. The optimum value of cluster heads based on the network lifetime has been calculated taking into consideration the energy consumed in the setup phase as well.

III. PROPOSED APPROACH

Since the energy consumption is a critical problem in wireless sensor networks, we used GA to cluster nodes in WSN in order to minimize the long-distance communications between the sensors and sink which leads to reducing the energy consumption in the network. Each node in the network can be either a regular node (member in cluster) or cluster head (CH). Each member sensor belongs to just one cluster head (i.e., the nearest one).

A. Genetic algorithm

Genetic algorithm is a model of machine learning which derives its behavior from evolution processes in nature [11]. Genetic algorithm creates a number of chromosomes randomly to perform an initial population then the fittest survived chromosomes move to the next generation.

1) Chromosome

Any solution to any problem in GA called chromosome or individual. The chromosome is represented by a list of parameters also called (genes) [2]. If the gene value is 1, then the corresponding node is a cluster head and 0 otherwise. Each chromosome represents a network topology whereas its size should be equal to the network size (i.e., number of sensors). On the other hand, the number of chromosomes (topologies) is determined by the population size parameter in which increasing the population size leads to increasing the number of suggested topologies. The chromosome (network topology), which is shown in Fig. 2, indicates that the network has 10 sensors. Sensors 4 and 9 are cluster heads and the others are normal nodes.

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
0	0	0	1	0	0	0	0	1	0

Fig. 2: Network chromosome

2) GA Operations

Genetic algorithms have three basic operations [12]: selection, crossover and mutation. Selection operation is used to choose the chromosome (individuals) based on their fitness value from the population in the current generation. In this paper, we used the roulette-wheel to per-

form the selection operation in which the chromosome that has a large fitness function value has a higher probability to survive to the next generation over others. In crossover operation, the chromosomes are recombined resulting two new child chromosomes to be added to the next generation population. Commonly, the genetic algorithm has a probability of crossover between 0.6 and 1. Increasing this value leads to increasing the crossover occurrence. It is noteworthy to mention that the crossover operator has many types out of which the single point crossover operator which means that after selecting the chromosomes, GA will generate a random number (between *0* and *network size -1*) to be splitted from this point into two parts to be then recombined. In this paper, we used this type of crossover operator. Lastly, the mutation operator is applied to each bit (gene) of a chromosome. Typically, the mutation operator flips the bits of the chromosome data structure. Generally, the genetic algorithm has a probability of mutation between 0.01 and 0.001. Similar to crossover operator, increasing this probability will increase the mutation occurrence.

3) Fitness function

Measuring the fitness or performance of chromosomes is done by a specific fitness function. The fitness function establishes the basis for selecting chromosomes that will be mutated. The fitness function used in our approach is the same as the one employed in [2]. When D_i and H_i have small values, this means that the fitness function has a large value. As a part of its task, the genetic algorithm tries to find the largest fitness function value in order to get the minimum number of clusters and long-distance communications for the sake of maximizing the network lifetime.

4) GA termination

GA is terminated right after a specified number of generations (runs). After that, it checks for the best chromosome found in the population [12]. It is good to state that a larger fitness function value means a minimum number of cluster heads and communication distance. However, after a certain number of runs (according to simulation results: larger than or equal to the network size), the variations in GA results will be small. Consequently, GA will find the solution that refers to having the largest fitness function value.

B. Approach Description

After applying the genetic algorithm, the hierarchical clustering will be applied on the selected topology. Generally, the proposed approach has the following steps:

1. Deploying a certain number of sensors in the specific territory and determining the location of the base station.
2. Determining the population size (the number of chromosomes or various topologies).
3. Determining the number of runs (generations): how many GA runs will be applied on chromosomes.
4. Applying genetic algorithm on the network as follows:
   A. Genetic algorithm will randomly generate a number of chromosomes according to a predetermined value (i.e., population size).
   B. Fitness function will be calculated for each chromosome in the population.
   C. Applying genetic algorithm operations (selection, crossover, and mutation).
   D. The previous steps will be repeated until the predetermined number of runs is reached.
   E. Finally, GA chooses the fittest chromosome (solution) from the population.
5. After finishing the genetic algorithm, the hierarchical clustering will be performed by the following steps:
   A. Each cluster head calculates its distance to the sink and all cluster heads.
   B. If the distance to the sink is the minimum distance compared with distances to reach other cluster heads, then the cluster head will be connected directly to the sink, otherwise, the cluster head will be connected to the nearest cluster head.
   C. The number of cluster heads connected to each other forms a group. The nearest cluster head to the sink is selected to be a super cluster head and gathers the required data from cluster heads within its group to be then relayed to the sink node.

IV. SIMULATION RESULTS

After extensive simulations, the results indicate that the proposed approach is efficient since it decreases the long-distance communications and energy consumption. For a different number of nodes, we consider the following parameters:

- *The Area of region* is 800 × 600 m^2,
- *Crossover probability* (P_c) is 0.9,
- *Mutation probability*(P_m) is 0.1,
- *w* is 0.5,
- *Population sizes* (according to network sizes of 50,100,150 and 200) are 50,100,150 and 200 respectively.
- *Number of generations* (according to network sizes of 50,100,150 and 200) are 50,100,150 and 200 respectively,
- *Selection operator*: Roulette-Wheel,
- *Crossover type*: One point.

The radio model used in our work is the free space radio model which stated in the equations described below given E_{elec}=50 nJ/bit as the energy being dissipated to run the transmitter or receiver circuitry and ε_{amp}=10 pJ/(bit×m²) as the energy dissipation of the transmission amplifier. Transmission (E_{Tx}) and receiving costs (E_{Rx}) are calculated by using the following equations [4]:

$$E_{Tx}(k,d) = E_{elect} \times k + \varepsilon_{amp} \times k \times d^2$$
$$E_{Rx} = E_{elect} \times k$$

where E_{elect} and ε_{amp} are constants and k is the number of transmitted or received bits. We considered the transmitted data to be only one bit. Hence, the communication distance (d) is the dominant factor in affecting the power consumption, this is to say, increasing (d) increases the power consuming and vice versa.

Fig. 3 shows the energy consumption when the sensors communicate directly with the sink after employing both GA with and without hierarchical clustering. It is noticed that the long-distance communications are decreased after clustering and decreased more when the hierarchical clustering is applied with knowing that decreasing the long-distance communications is reflected positively on the energy consumption.

Table I shows the number of clusters after applying GA with hierarchical clustering on different sink locations. The number of clusters is increased when the sink is located at the center of network area for a fixed value of network size. As a matter of fact, this is expected because all sensors are located around the sink whereas the possibility of forming more clusters gets increased more and more. For example, when deploying 100 sensor nodes, the values of formed clusters are 17 and 11 when the sink is located at (0,0) and (400,300), respectively.

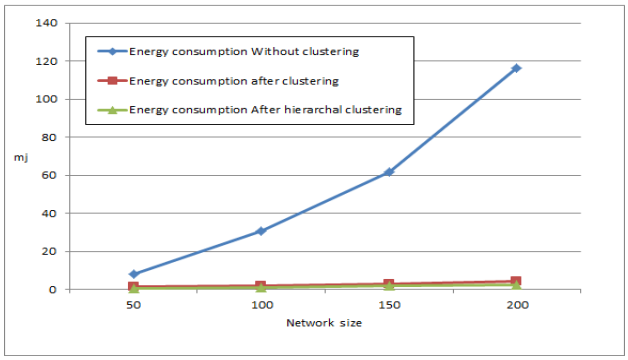


Fig. 3: Decreasing energy consumption after clustering

TABLE I
NUMBER OF CLUSTERS WITH DIFFERENT SINK LOCATIONS

Network size	Sink at (0,0)	Sink at (400,300)
50	13	7
100	17	11
150	22	14
200	34	28

When the value of w is changed to 0, this will force the genetic algorithm to consider the number of clusters in the fitness function only. Therefore, the network should have the least number of cluster heads. The optimal number is only one head. Fig. 4 verifies the effectiveness of our proposed algorithm though the topology might not be realistic.

In Table II, a reduction is shown in the long-distance communications when considering direct communication, GA with and without hierarchical clustering. The hierarchical genetic clustering reduces the long-distance communications more than the normal clustering. When the network size is increased, the efficiency of the hierarchical GA clustering is increased.

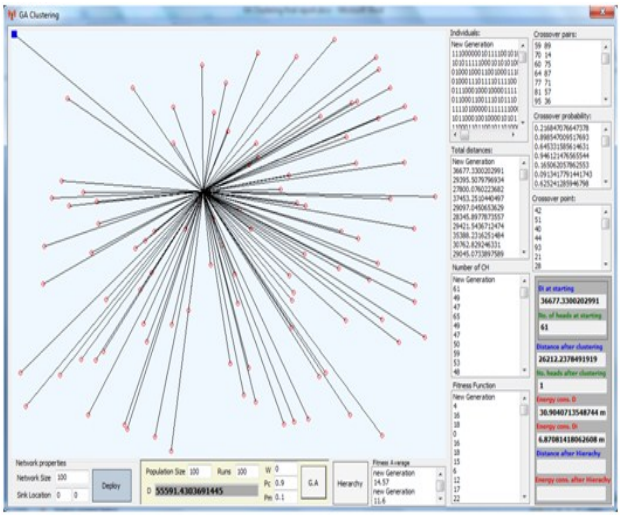


Fig. 4: One cluster head when w=0, network size=100

TABLE II
LONG-DISTANCE COMMUNICATIONS

network Size	Long-distance communications before clustering D (m)	Long-distance communications by GA clustering D_i (m)	Long-distance communications by hierarchical GA clustering (m)
50	27980	21881	20161
100	51175	42061	39166
200	108398	92485	86795

Fig. 5 shows the decreasing of long-distance communications when the number of generation increases with a network size of 100 nodes. Actually, decreasing the long-distance communications as the genetic generations increase means that our proposed approach is very efficient approach and performs correctly as expected.

We compared our work with Shiyuan et al. [2] where the authors used genetic algorithm to cluster the network nodes. Table III shows the results of this comparison. Our improvements are clearly noticed especially when the network size is increased.

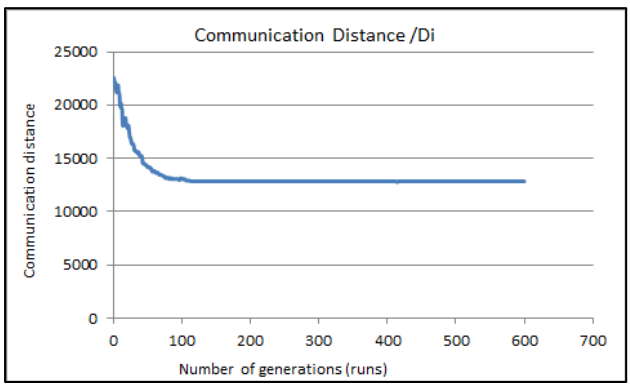


Fig. 5: Long-distance communications throughout generations/runs

TABLE III
COMPARISON BETWEEN OUR APPROACH AND GC BY SHIYUAN

Node size	Long-distance decreased by (hierarchical Clustering)	Long-distance decreased by Genetic Clustering (by Shiyuan)
100	82%	76.85%
200	86%	81.20%
400	87%	82.20%

Fig. 6 shows the average distance reduction of our algorithm when it is compared with [2]. The reduction is about 85% while in [2], it is about 80%. From $E_{Tx}(k,d)$ equation described previously, the distance parameter is the only factor that affects the energy consumption. As a result, our approach reduces the energy consumption more than what is obtained in [2].

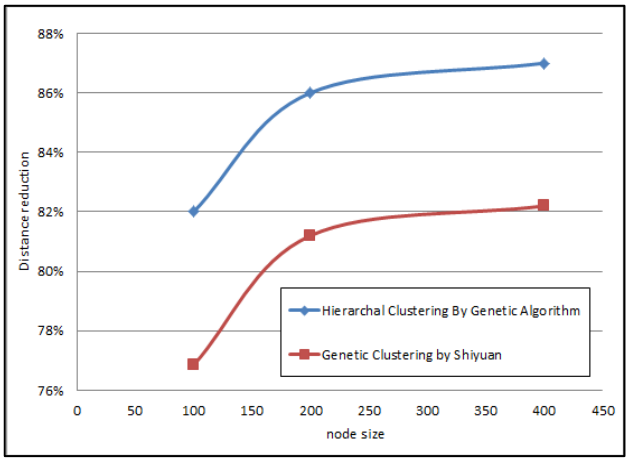


Fig. 6: Comparison between our approach and [2]

V. CONCLUSION

Clustering sensors into groups reduces the long-distance communications between sensors and the sink. Consequently, the energy consumption in the network is reduced. Due to the fact that it can find the solution very quickly without complex computations needed, clustering using genetic algorithm is an efficient method to cluster the networks. Simulation results illustrate that the hierarchical clustering using genetic algorithm decreases the long-distance exchanges very well. On the other hand, hierarchical clustering using genetic algorithm is an applicable and scalable approach where it can be implemented in large number of nodes with different base station locations and nodes deployment styles.

REFERENCES

[1] XuXu and Weifa Liang, "Placing optimal number of sinks in sensor networks for network lifetime maximization," *Proceeding of IEEE International Conference on Communications* (*ICC 2011)*, Kyoto, Japan, 5-9 June, 2011.

[2] Shiyuan Jin, Ming Zhou, and Annie S. Wu, "Sensor Network Optimization Using a Genetic Algorithm," *Proceedings of the 7th World Multiconference on Systemics, Cybernetics, and Informatics*, Orlando, FL, July 27-30, 2003.

[3] M. J. Handy, M. Haase, and D. Timmermann, "Low energy adaptive clustering hierarchy with deterministic cluster-head selection," *Proceedings of IEEE 4th International Workshop on Mobile and Wireless Communications Network*, pp. 368-372, Sweden, –9 11 Sept. 2002.

[4] Ge Ran, Huazhong Zhang, and Shulan Gong, "Improving on LEACH Protocol of Wireless Sensor Networks Using Fuzzy Logic," *Journal of Information & Computational Science*, vol.7, no. 3, pp. 767–775, 2010.

[5] Jenn-Long Liu and Chinya V. Ravishankar, "LEACH-GA: Genetic Algorithm-Based Energy-Efficient Adaptive Clustering Protocol for Wireless Sensor Networks," *International Journal of Machine Learning and Computing*, vol. 1, no. 1, pp. 79-85, 2011.

[6] Wendi RabinerHeinzelman, AnanthaChandrakasan, and HariBalakrishnan,"Energy-Efficient Communication Protocol for Wireless Microsensor Networks," *Proceedings of the 33rd Hawaii International Conference on System Sciences*, vol. 8, Maui, Hawaii, January 4-7, 2000.

[7] Raghavendra V. Kulkarniand Ganesh Kumar Venayagamoorthy, "Particle Swarm Optimization in Wireless Sensor Networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part C,* vol. 41, no. 2, pp.262-267, 2011.

[8] P. K. Agarwaland C. M. Procopiuc, "Exact and approximation algorithms for clustering," *Algorithmica*, vol. 33, pp. 201-206, 2002.

[9] Ehsan Heidari1 and Ali Movaghar, "An efficient method based on genetic algorithms to solve sensor network optimization problem," *International journal on applications of graph theory in wireless ad hoc networks and sensor networks*, vol.3, no.1, pp. 18-33, 2011.

[10] Rabia Noor Enam, Mumtazul Imam, and RehanInamQureshi, "Energy Consumption in Random Cluster Head selection Phase of WSN," *IPCSIT*, vol. 30, pp. 38-44, 2012.

[11] Luiz Antonio,Nogueira Lorena, andJo˜ao Carlos Furtado, "Constructive Genetic Algorithm for Clustering Problems,"*IEEE Transactions on CAD of Integrated Circuits and Systems,* vol. 21, pp. 969-974, 2002.

[12] Genetic Algorithm Web reference, http://en.wikipedia.org/wiki/Genetic_algorithm.

Efficient wireless signal emitting sensor distribution within an environment modified for the needs of the aging population

V.L. Prčić*, M. Đurek** and D. Šimunić**
* Infoart d.o.o., Zagreb, Croatia
** Faculty of Electrical Engineering and Computing, Zagreb, Croatia
vprcic@infoart.hr

Abstract - With the average human life span increasing significantly due to numerous achievements in medical science, the need to modify the residential environments of individuals has grown accordingly. Some of these changes involve building fast and responsive systems designed to continuously monitor the activities within the space they are situated in. Such sensors may vary from simple light switches that react to movement to complicated systems that search for specific events and alert the designated party. The distribution of such sensors greatly dictates the overall system reliability. In this paper the problem of finding the optimal sensor distribution will be elaborated with focus on a solution using a computer program designed on the idea of genetic algorithms that can solve the problem in a simple and cheap manner. The program uses rudimentary data as input, such as floor plans and regions of interest in the form of a picture. Using a hybrid genetic algorithm, while taking physical properties like signal attenuation, dissipation and resonance in account, it provides a result in a form of a sensor distribution pattern such that the overall signal strength is maximized.

I. INTRODUCTION

From the beginning of the computer era, the usage of the available computing power has been directed to solve problems that were otherwise solved by humans, but at a much slower pace. Since then, the always expanding set of possible problems for a computer so solve has grown to the extent where not all problems were deterministic in the aspect of the existence of an answer, but a fuzzy answer had to be considered as a good one. What this agreement brought with it was the notion of a "better answer" in contrast to the previously only one "right answer". In light of this new setting, a number of approaches have been devised to find ever increasingly better answers.

In overall, the idea behind all of the mentioned approaches lies is the fact that there is no way to represent the problem being solved in a computer environment that would provide direct means for finding an answer. It speculates that it is better to find many bad answers and modify them so that their quality in answering the problem increases and not relying to the fact that a single right answer will be found. And since there is no difference regarding the way those answers are generated, they may as well be built upon random choice, but one that would yield good results – a heuristic.

The heuristic method that is the foundation of this work is a method created by John R. Koza called genetic programming[1][2][3][4][5][6]. The idea behind it is that using a tree-like computer structure one can define a computer program where each node in that structure would represent a single function. This function would produce a single output based on a number of inputs gathered from other nodes that are child nodes to the first one.

Using this technique Koza created a computer system that utilizes the power of genetic programming, colloquially called the "invention machine", with which he was able to reconstruct 21 instances of previously patented inventions and to construct 2 patentable inventions[7][8] [9].

As well as genetic programming, heuristic methods are based upon nature's blueprint because of the fact that we can be sure that they work by observing the evidence that lies before us. They include some that are widely recognized like genetic algorithms, neural nets, simulated annealing or ant colony optimization and can be applied to many different areas, but also some that are not yet so well adopted like gene expression programming or cuckoo search.

II. GENETIC PROGRAMMING

Koza recognizes two types of nodes that construct the structure: the terminals that represent the initial values that will be fed into the structure, and the primitive functions that represent the nodes that do the calculations. In this manner, he defines two sets – the terminal set that contains all the terminals that are available for usage and a function set that similarly contains all the primitive functions. This way, one is able to create complex structures that behave in a certain way depending on the initial inputs that are fed into them. And the more complex the structure is the more complex is its behavior. An example of a simple structure can be seen on figure 1. In this structure one can see two terminals that are both numerical values, and two primitive functions 'f' and 'g' where the first one is a function of one variable, and the second one a function of two.

Beside the two sets, one must also introduce a number of other objects for the genetic programming algorithm to

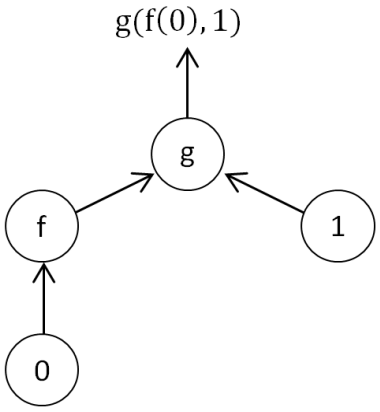


Figure 1. A representation of a tree-like structure that could be the result of a genetic programming algorithm.

be fully defined: the fitness measure, the parameters of the algorithm and the termination criterions. Within the three mentioned parameters, one could argue that the most important one is the fitness measure relying on the fact that the other two are fairly independent of the problem being solved.

The fitness measure is a function that determines a quality of a solution gained through the algorithm of genetic programming. With this function one can compare two solutions and determine which of two is better in regards to the problem being solved. This function is highly problem-dependent and in a way models the problem itself. An example of such a function can be given using the following setup. Suppose one is to define the algorithm of genetic programming that would solve a simple problem – to define any positive integer using only division and multiplication of number one. Here, the terminal set and the primitive functions set can both be determined from the task's description. Regarding the fitness measure, if we denote the wanted value with υ and the value obtained from one of the solutions with γ then the fitness measure could be defined in a number of ways with the few of them being:

- $\mathrm{FM}(\upsilon, \gamma) = -|\upsilon - \gamma|$
- $\mathrm{FM}(\upsilon, \gamma) = -(\upsilon - \gamma)^2$
- $\mathrm{FM}(\upsilon, \gamma) = \exp[-(\upsilon - \gamma)^2]$

It is easy to observe that all of the mentioned functions have the property that they get smaller in value the further the wanted value and the calculated values are apart from one another, and attain a maximum value when the two are the same. And even if they are different, all of them would be good candidates for the fitness measure.

III. Genetic programming operators

After the problem has been defined through the definition of the genetic programming algorithm, one could start to seek its answer. But before, the ways in which those answers are found have to be explained. As it was mentioned earlier, many heuristic methods are based on one or many processes that occur in nature, so it is no surprise genetic programming is as well. More specifically, it is based on the process of evolution.

The algorithm of genetic programming isn't accomplished in one single step, but rather a number of them called iterations. In the beginning, an initial population of possible answers is generated at random. The algorithm continues so that in each iteration a number of solutions are picked from the population and an operator is applied to them yielding a number of new solutions that are then added to the new population. After a number of iterations, the fitness measure is applied to all solutions in the new population and if the termination criterions aren't met, the process continues. Here the two objects that are needed to fully define the genetic programming algorithm (the algorithm parameters and the termination criterions) can be easily explained. The algorithm parameters define values like the size of the initial population or the number of iterations that are performed during the application of operators to the current population, and the termination criterions define when a suitable solution has emerged, and can be defined as a predetermined number of runs or a minimal value of the fitness measure of a single solution in a population.

As for the operators, Koza describes three nature-inspired operators and one architecture-based. The first three will be explained in detail, while the last one will be skipped since it is tied to the specific implementation of Koza's solution that doesn't allow the repeated usage of a sub-tree in a solution, so other actions had to be made to make it possible.

The far simplest of the three operators is the operator that simply copies a solution to the new population and therefor describes the survival of a biological entity. The second operator is the crossover operator that represents the procreation of two biological entities that create a new entity. This operator is the foundation of most heuristic algorithms based on evolution because it allows the creation of a composed solution with aspects of both parents that were involved in the operation. The main idea is that a solution could be created that would carry on a winning trait of both parents and thus pushing the whole algorithm one step further to a better solution. The last operator that Koza defines is mutation. This operator allows for a good solution to be, in A manner of speaking, tweaked to become an even better one, and it is as vital and as common to most evolutionary algorithms as the crossover operator.

IV. The idea behind Heuristic Algorithm Generating System

The reason why HAlGeS (Heuristic Algorithm Generating System) was created is to provide the means of problem solving with a powerful method to those that aren't involved in this kind of area through a simple but versatile interface. This means that the system ought to be robust and adoptable, but shouldn't force its user to dwell on problems that do not concern him. Such a solution implies that a number of adaptations should be made in order to make this possible.

The main role of HAlGeS is to become a system that could solve a user problem using a set of primitive functions and a set of operators through the algorithm of

genetic programming, but with the exception that the architecture of HAlGeS allows users to import and use primitive functions as well as operators that were defined by other users or by themselves. This means that given a specific domain in which a problem is to be solved, a set of already defined primitive functions and operators can be given to the user, and with HAlGeS one can find a solution to the domain-specific problem without the need of understanding the logic behind it. Alternatively, if one wished so, one could define a custom primitive function or operator so others may use it. As it is easy to notice, this is similar to the idea of "open source", only applied to genetic programming.

V. IMPLEMENTATION

HAlGeS is implemented using a Microsoft programming language C# which may seem in contradiction with the "open source" parallel described in chapter 4, but was chosen for a number of reasons, and the main one being the fact that at the time of its implementation no such system has been created using this technology. Besides that, the mentioned programing language is suitable to use for this problem because of the object-oriented paradigm it is built upon.

To implement the genetic programming algorithm, a number of definitions had to be made. The two types of nodes described by Koza were replaced with one where the terminals are defined as primitive functions with no input arguments. This merging was made so the implementation would be easier. Other modifications to Koza's definitions were made as well, but the most significant one is to the structure that defines a single solution. As Koza defines it, each solution is a tree-like structure. This implies that if we represent a sub-tree of a solution as a separate primitive function, then the output of this function can only be used as an input argument to a single primitive function in a higher hierarchy position. This approach brings with it the problem of "code redundancy", but also slows the convergence of the genetic programming algorithm because if the mentioned sub-tree structure is essential to the discovery of a better solution, then it would need to be "reinvented" for each primitive function in a solution that uses it. Because of this problem, Koza introduced subroutines into his design, and the appropriate operators that would deal with them mentioned in chapter 3.

In HAlGeS, the mentioned problem is solved by defining the solutions as directed acyclic graphs. This way, any number of primitive functions can use a specific sub-graph's output as an input argument. Furthermore, it allows the creation of solutions that have several outputs, in contrast to Koza's definition where a single solution is bound to have only one.

Another important change in the definition of the solution structure is the possibility to define the output type of a primitive function. As in many computer programming languages, there are different types of values that are needed to create a computer program, so this kind of thinking is expanded to HAlGeS through the grouping of primitive functions in subsets based on the type of the output value they produce. This regrouping is taken into consideration when a solution is being created or altered since it determines that input arguments of a primitive function can only be represented with primitive functions whose output value type matches the input argument type.

One can see that the mentioned redefinitions bring improvement either in a form of flexibility or as the reduction of redundancy, so in coherence with this approach one last change has been made in the definition of primitive functions concerning their input arguments and the property of commutativity. Since many of the primitive functions that are generally used are commutative, this property has been implemented in HAlGeS as well. This means that a primitive function can have an undetermined number of input arguments where a single input argument is an ordered set of several values or a single value. Every input argument has the same structure regarding the number of values it contains, the order they are arranged in and the value type they have.

As for the operators that are used in the algorithm of genetic programming, they too have been redefined in several ways, but to explain them fully a different change should be explained. In HAlGeS, there isn't a single group of solutions that undergo the application of operators but a type of grouping is introduced that is, once again, inspired by nature. All of the solutions are contained in a single group called the world. The world contains a number of smaller groups called habitats and within these habitats there are a number of solutions simply called algorithms. The main idea is for each of the habitats to evolve in a different direction, and then using a number of specialized operators migrate solutions from one habitat to another. This would allow the creation of different "ideas" in a form of sub-graphs that would then be able to join.

According to that, there are three types of operators or tweakers that can be defined in HAlGeS based on the reach of their influence: the world tweaker, the habitat tweaker and the algorithm tweaker. The world tweakers are mostly responsible for the creation and destruction of habitats, as well as for migration of algorithms. The habitat tweakers create new algorithms for diversification, perform crossovers and apply selection, while algorithm tweakers deal with individual algorithms through mutation and a number of other ways that will be described shortly. Beside the mentioned differentiation of tweakers, each tweaker can be defined further with two values indicating its behavior. The first one indicated whether a tweaker needs for all of the algorithms it can reach to have the correct quality assigned to them through the quality function, and the other one indicates whether it changes the quality of an algorithm so after the tweaker is applied the algorithm needs to be reevaluated. Here the topic of algorithm evaluation and quality has been brought up, but since it is of the most importance, it will be explained in detail in the following chapters. This kind of approach to tweaker differentiation enables the faster execution of the genetic programming algorithm because it removes the need of reevaluation after each tweaker

application and this is of great importance since unlike in Koza's model, in HAlGeS all of the tweakers are applied during a single run of the genetic programming algorithm.

There are currently 14 tweakers and 54 primitive functions implemented in HAlGeS. The 54 primitive functions represent some of the most common functions in use regarding double precision numbers and a number of specific primitive function that are focused on the field of image transformation and analysis. The 14 tweakers contain the ones responsible for crossover, mutation, selection, migration, diversification, elitism, habitat control and one that optimizes a certain algorithm in a way that needs additional explanation.

To further expand the possibilities of HAlGeS, another modification is introduced in the definition of primitive functions. During the development of HAlGeS, it has become obvious that some primitive functions require parameters that additionally describe their behavior. A simple example of this would be a primitive function implementing the exponential function. Here, the value of the exponent is a parameter that describes the primitive function's behavior. To make this kind of parameterization possible, two kinds of values were defined: the parameters and the descriptors. The descriptors of a primitive function are specific values that for a certain realization of a primitive function do not change during the genetic programming algorithm, and the parameters are values that do change but depend on the descriptors. As an example of this could again serve a primitive function representing the exponential function that can attain exponent values only in a specific range of values. The lower and the higher boundary would in this case be the descriptors of this primitive function, but the value of the exponent itself would be its parameter because it could change during the execution of the genetic programming algorithm and it depends on the values of the boundaries.

What the tweaker that optimizes an algorithm is meant to do is to create a genetic algorithm based on the descriptors and parameters of that solution and then optimize it by finding the one combination of its parameters that yield the best results. This approach could be used to create more tweakers that user some other kind of heuristic like simulated annealing to optimize solutions.

Similarly to descriptors and parameters that were defined for primitive functions, tweakers too have certain values that define them more completely. These values are the implementation of the parameters Koza introduced in the definition of a genetic programming algorithm. But contrary Koza's approach that implies that once the parameters are set they cannot be changed, in HAlGeS every parameter can be modified during the execution of the genetic programming algorithm. This enables the user to change certain parameters like the likelihood of a mutation or migration if he feels it would affect the end result. Also, each tweaker has the ability to store any kind of information programmatically within a specialized list contained within the world and habitat implementations. An example of where this list is useful could be the tweaker that implements the idea of elitism. Elitism is the idea of always saving the algorithm with the highest fitness measure in a certain habitat. Using the mentioned list, a tweaker can store the best algorithm in it so it may compare it to all the other algorithms in the next iteration and add it to the habitat as needed.

VI. EVALUATING ALGORITHMS

The main aspect of the genetic programming algorithm that needs to be defined is the way the algorithms are evaluated and qualified. Like mentioned before, the fitness measure that is used to compare different solutions depends on the problem being solved. For this reason, this is the only part of code the user must write on his one for HAlGeS to be able to work. This has been made easy by defining an abstract class named **evaluator** that the user's fitness measure class must extend, and the only functionality it must implement is the one that assigns a value between zero and one that represents the quality that a certain algorithm exerts in solving the user's problem.

Here, the last modification of the primitive functions comes to light. In order for the user to be able to feed some data in an algorithm, a connection between the primitive functions used in the genetic programming algorithm and the evaluator class had to be accomplished. The evaluator defines what primitive functions that require input from outside it supports, and then each primitive function is feed with this data when it is asked to evaluate.

VII. IMPLEMENTATION OF NON-HEURISTIC ASPECTS

Besides the genetic programming algorithm that is implemented in HAlGeS, a number of other functionalities had to be revised. First of all, a graphical user interface was created. Using this interface, the user can define, modify and control all of the aspect of the genetic programming algorithm. He can add and remove primitive functions, as well as tweakers and their parameters, manually create or destroy habitats, observe the evolution of the best algorithm through iterations and so on.

Also, for HAlGeS to be truly useful, one should be able to save his work for later, so this was also implemented through serialization of all algorithms and tweakers in the form of XML files or the possibility to create custom ways of exporting algorithms like in the form of a DLL file or a simple graph image. The C# classes implementing this behavior are called **exporters**, and are implemented easily using an existing abstract class that is a part of HAlGeS.

VIII. APPLICATION

To apply the built system on the problem of optimizing the total signal power of a distributed network of emitters, two primitive functions had to be defined. The first primitive function describes the emitter with the position being the variable parameter, and the second function describes a unifying function that acts as a

collector for the emitter primitive functions. The position is defined with two numerical values between 0 and 1 that can be viewed as a percentage value, one for each axis of the two-dimensional blueprint that describes the space in which the optimization process occurs. So, for the vertical axis, value 0 describes a position at the top of the blueprint, where the value 1 describes a position at the bottom. The genetic operators used for solution construction where the ones defined in the previous chapters.

Additionally, an evaluator had to be implemented that used the definition of the quality of a solution as a function of the normalized average power of the signal calculated on a mesh composed of a constant number of points. In each of those points, the percentage of the maximal power for each emitter was calculated. The average value of those percentages was considered as a measure of the emitter's usefulness.

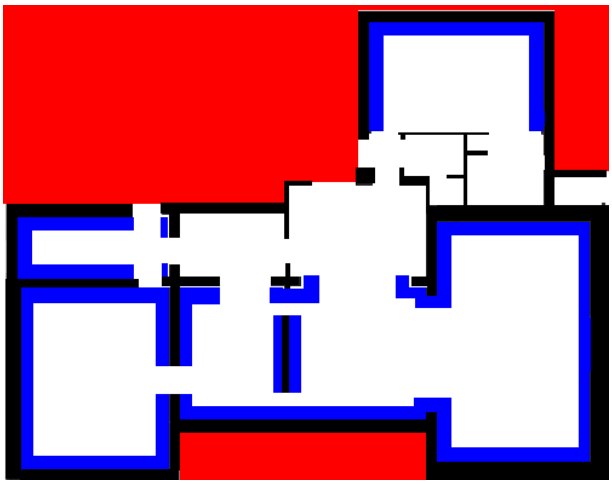


Figure 2. An example of the evaluator imput image. Black indicates an obstacle, blue a region where an emitter can be placed, white the region where the signal should be optimal and red a region of no interest.

Figure 3. An example of the result obtained from the optimization process with 12 placed emitters (blue circles). Blue dots represent possible emitter positions, and green to red dots indicate calculated power strength.

The evaluator's input is composed of an image that represents the optimization space by defining the physical space configuration (blueprint) and the regions of interest. The specified regions of interest include a region where an emitter can be placed and a region where the signal should be optimal. These regions are distinguished by usage of different colors. If a solution proposes the positioning of an emitter outside the area where it's possible to place one, a quality of zero was assigned. Also, the estimated quality of the solution was calculated only based on values calculated in points that are within the second specified region of interest.

IX. Conclusion

Using HAlGeS as a base ground for solving the specified problem proved to be successful. The system converged on a solution that covered the optimization area with satisfactory quality. Using the approach based on evolutionary algorithms implicates a process of repeated solution evaluation. Keeping this in mind, the benefits of using the specified approach become clear since it means a vast number of possible solutions is checked and tested during the running course of the algorithm. This way a broad field of possible configurations is evaluated simultaneously, as opposed to checking a single solution and moving on to the next one. Furthermore, a single configuration might develop a solution for a specific sub problem, but fail due to its overall quality, and rather than being cast away completely, the generated idea, represented as an emitter configuration, might migrate to another configuration and improve its quality.

Further work should focus on enabling parameterization so that the provided solution can be constrained to a greater level of detail with, for example, regions of interest with different interest factors. Also, it would be opportunistic to consider expanding the problem to a three dimensional area, but this would require solving the problem of usage ergonomics.

X. Application in Healthcare

The idea of a smart home specifically designed for the needs of the aging population has grown in the last years somewhat due to the increase of available technology ranging in size as well as in price, and somewhat because of the fact that the average age is getting higher. Implementing any kind of monitoring that would allow a caretaker to be involved in the daily routine of an elderly person without jeopardizing the independent lifestyle that most of the population in question wants, implicates the usage of sensors distributed within the living space connected using a wireless network. Such sensors may range from a simple device that counts the number of times someone enters a room to those that would take specific action if a deviation from the daily routine is detected. In such situations it is needed to distribute the said sensors in an optimal way, and here the work described in this paper can be applied. But this approach doesn't need to stay confined to personal living spaces since the same idea can be applied to any kind of institution that requires this kind of monitoring – from hospitals to nursing homes.

XI. LITERATURE

[1] Koza, J.R. "Genetic Programming: A Paradigm for Genetically Breeding Populations of Computer Programs to Solve Problems." Stanford University, Computer Science Department (1990)

[2] Koza, J.R. "Genetic Programming: On the Programming of Computers by Means of Natural Selection." The MIT Press, Massachusetts (1992)

[3] Koza, J.R. "Genetic Programming II: Automatic Discovery of Reusable Programs", The MIT Press, Massachusetts (1994)

[4] Koza, J.R. "Genetic Programming III: Darwinian Invention and Problem Solving", The MIT Press, Massachusetts (1999)

[5] Genetic Programming Inc. Homepage (2003) [online] "What is Genetic Programming?." Available at: http://www.genetic-programming.com/ gpanimatedtutorial.html

[6] Genetic Programming Inc. Homepage (2003) [online] "Flowchart (Executional Steps) of Genetic Programming." Available at: http://www.genetic-programming.com/gpflowchart.html

[7] Koza, J.R. "Genetic Programming IV: Routine Human-Competitive Machine Intelligence", The MIT Press, Massachusetts (2003)

[8] Genetic Programming Inc. Homepage (2003) [online] "Definition of “Human-Competitive”." Available at: http://www.genetic-programming.com/humancompetitivedefinition.html

[9] Genetic Programming Inc. Homepage (2003) [online] "1,000-Pentium Beowulf-Style Cluster Computer for Genetic Programming". Available at: http://www.genetic-programming.com/machine1000.html

TeleAssistant – Assisted Living Platform Applied to Fall Detection

Marko Čavka*, Roko Krpetić**, Darko Vasić**, Vedran Bilas**

*Hrvatski Telekom d.d.
Services and Network Systems Development Department
Ozaljska 91, 10000 Zagreb, Croatia
marko.cavka@t.ht.hr

**University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{ vedran.bilas, darko.vasic } @fer.hr, roko.krpetic@gmail.com

Summary – According to aging trends and strategic orientation for providing higher healthcare standards, assisted living is one of the key research areas in the western world. Fall is the most common cause of injury of an elderly person. Therefore, with its timely detection we can ease its consequences or even avoid death. In this paper, we present architecture developed in Hrvatski Telekom that provides functionalities of automatic alarm notification in case of fall detection and manual notification for the "on demand assistance request". Based on this architecture, the prototype has been built and consists of: wireless fall detection sensor, communication gateway, home router and server. This architecture and prototype has been described in more detail, along with its performance benchmark of fall detection ratio, wireless coverage, energy efficiency and quality of communication channel.

I. INTRODUCTION

The timely fall detection is an important component of any assisted living system that is to have a real positive effect on healthcare standard of disabled and elderly persons. The fall is one of the leading causes of death for people over the age of 65 [1]. The probability of fall increases three times for people affected by Alzheimer's disease. Nearly 40% of the fallen patients visited an emergence room and needed to be hospitalized. Nearly 3% of those falling helplessly lie without any support for more than 20 minutes. The proportion of injurious falls is significantly higher at night as opposed to daytime or evening. The walking (ambulation) is the most common activity at the time of fall and fracture [2]. The cost forecast of medical care for elderly fall related injuries is above 40 billion by 2020 [1, 3].

Thus, physical activity monitoring and automated fall detection is important for those individuals who have previously sustained a fall-related fracture and for physically active elderly people who are alone, constantly or occasionally [2].

The single and most important task of the fall detector is to detect reliably all falls with minimum number of false positive alarms, since any fall, if the person is left without help can be dangerous, and any false alarm increases the intervention costs.

There are three main approaches to fall detection [1]: vision-based using cameras, environmental using infrared and pressure sensor, and wearable using accelerometers and gyroscopes. This study is focused on the wearable approach to fall detection using accelerometers. Since it is necessary to measure the whole body movement, the sensors are typically placed on the sternum [4], lower back [5], and waist [6, 7]. Among these, the waist is most preferable, since it is located near the body center of mass, and the sensor at this location measures the acceleration representing the major human motion [8]. Also, the waist devices are easily attachable, cause less constraint in body movement and minimal discomfort, which results in greater patient's compliance. Detection of a range of activities was reported, such as walking, posture, activity transitions and falls [7].

Approaches to the problem of movement classification, including fall detection are diverse, ranging from the fixed-threshold methods to the artificial neural networks and fuzzy logic [7]. In order to minimize the use of power-hungry operations such as radio communication, the signal processing should be simple and made onboard a wearable device (i.e. detector). The threshold-based fall detection algorithms promise to meet the constraints of the limited memory and processing resources of a wearable device.

In this paper we present the architecture that provides functionalities of automatic alarm notification in case of fall detection and manual notification for the "on demand assistance request". We have built a prototype consisting of a wireless fall detection sensor, communication gateway, home router and server. The paper is organized as follows: The system architecture and overview of the main functionalities are given in Section II; Technologies used to implement the prototype system are described in Section III; The prototype's performance benchmark is provided in Section IV; Section V contains the conclusions.

Hrvatski Telekom

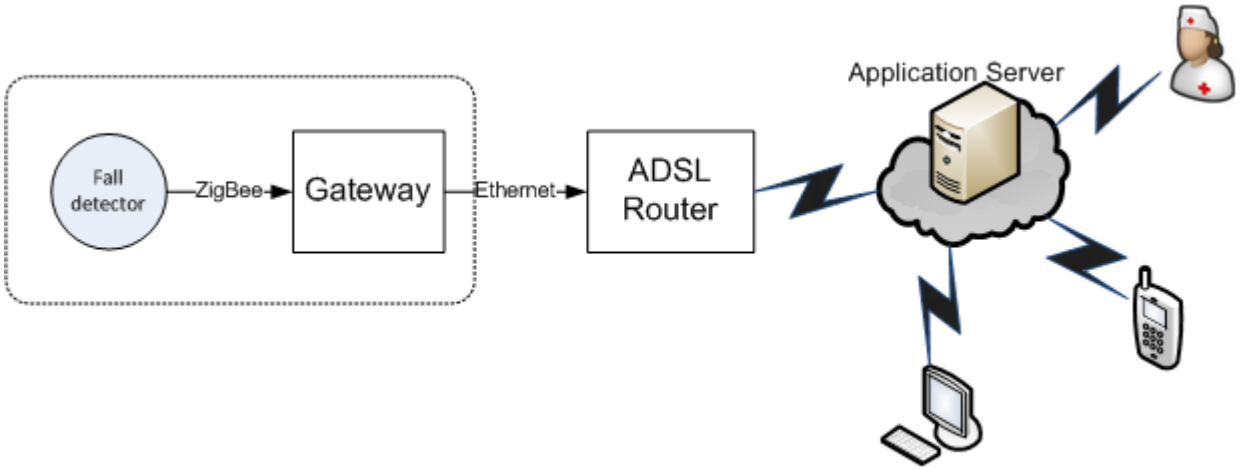


Figure 1: Architecture of the TeleAssistant system

II. System Architecture

Main functionalities of TeleAssistant system can be described with four characteristic messages that are carried over its components.

- Fall detection notification – message that indicates person's fall. This message is a signal to a caretaker for an immediate assistance.
- Assistance request notification – message that is transmitted when a person manually requests for assistance. This message is also a signal to the caretaker for assistance but it is triggered purposely.
- Low battery notification – message that is dispatched in case of low battery charge in the wireless fall detector. It implements self-checking threshold based algorithm.
- Wireless (ZigBee) network problem notification – message sent in case that wireless sensor is inactive for some period of time (1 minute in prototype).

TeleAssistant architecture consists of several components: wireless fall detector, communication gateway, home (DSL) router and application server. Figure 1 depicts the architecture of TeleAssistant.

Wireless fall detector is a small (90 x 60 x 23 mm) device carried on waist by hooking it up on a users' belt. Its hardware architecture is built from easily available, off-the-shelf components to ensure development flexibility and expandability. Energy consumption and price are two other main criteria for the selection of the components. Our wireless fall detector device is shown in Figure 2. It has pushbutton and buzzer which allows it to send assistance request notification and sound the alarm in case of a detected fall, assistance request or low battery signal.

Communication gateway together with wireless fall detector form a wireless sensor network (WSN). Communication gateway serves as an interface between ZigBee and IP network (Internet). It is also the main node (coordinator) in WSN, which is the destination for all messages sent from the wireless sensor(s). It is connected to the home (DSL) router which is standard equipment for the Internet access provided for the customers of Hrvatski Telekom using MAXadsl service. Communication gateway carries on three main functionalities:

- Reception of all messages sent from every node within the WSN.
- Sending messages to the application server.
- Checking integrity of the ZigBee network – in case wireless fall detector is inactive, alarm is sent to the application server.

Application server receives HTTPS messages from the communication gateway. Based on these messages it can perform actions such as sending a notification SMS or an e-mail. For the prototyping purposes, the application server stores and displays these messages inside the web application.

Figure 2: TeleAssistant's wireless fall detector device

III. USED TECHNOLOGIES

Root concept of this project was development of a system that collects data from wireless sensor devices and transmits it to the Internet through the digital subscriber line (DSL) which is a common service provided by Hrvatski Telekom. The wireless sensors can be used for various purposes like controlling house appliances and lightning, environmental monitoring, monitoring of electricity consumption, medical and protection purposes. In this project, a fall detector is used as a wireless sensor device. The wireless sensor and communication gateway form a wireless sensor network (WSN), which is used because of its favourable features like: low energy consumption, small device size, simple operation and acceptable hardware cost. For the purpose of TeleAssistant project WSN should implement multi-hop communication, so that its range can be easily upgraded with additional repeater nodes. Hence, we chose ZigBee based WSN. Various wireless protocols are compared in Table I [9, 10].

Wireless fall detector consists of an accelerometer sensor, microcontroller, WSN transceiver, battery, pushbutton and a buzzer. Digital accelerometer (Analog Devices) ensures low energy consumption. Atmel microcontroller controls the hardware and runs fall detection algorithms. WSN transceiver uses standard Digi XBee interface which enables an easy switch to another wireless network protocol. Rechargeable polymer lithium ion battery provides 3 to 7 days of autonomy and is recharged using miniUSB interface which lasts around 3 hours.

Threshold based fall detection algorithm is implemented as a finite state machine. The key parameter of the algorithm is total acceleration which is a vector sum of accelerations for each axis. The fall is detected when following types of movement occur in a row within defined timing: free fall, impact, inactivity and non-upright posture. The first three types of movements are detected by comparing total acceleration to defined thresholds. Non-upright posture is detected by computing a module of a vector difference between total acceleration vector and initial total acceleration vector when a person is upright and comparing this number to a defined threshold.

TABLE I: COMPARISON OF WIRELESS PROTOCOLS

	Bluetooth	**Bluetooth Low Energy**	**ZigBee**	**Wi-Fi**
IEEE standard	802.15.1	802.15.1	802.15.4	802.11
Frequency	2.4 Ghz	2.4 GHz	869/915 Mhz 2.4 GHz	2.4 GHz
Max speed	1-3 Mb/s	1 Mb/s	250kbit/s	54 Mb/s
Range	100 m	50 m	100 m	100 m
Max nodes	8	N/A	65000	2007
Complexity	High	Low	Low	High
Consumption	Medium	Very low	Very low	High

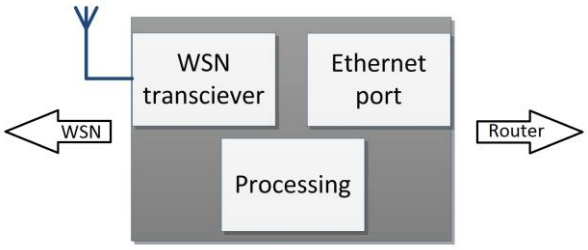


Figure 3 : Communication gateway block diagram

Power management system manages transitions between active and sleep states of the microprocessor and WSN transceiver with a goal to maximize device autonomy. When accelerometer senses no movement, the microcontroller is set to sleep mode. When a motion is sensed, microcontroller switches to an active state and starts running the fall detection algorithm. Wireless transceiver is in active mode only during wireless transmissions.

Communication gateway serves as a bridge between WSN and Internet. Its block diagram is depicted in Figure 3. On one side it is connected with the ZigBee protocol with wireless fall detector and on other side through the Ethernet with the home router.

From the communication gateway data is transmitted through the home router to the application server where it is saved for further usage. Usually the information that is transferred is personal or sensitive hence the HTTPS protocol is used for the purpose of encryption [11]. Application server is based on open-source Java EE platform which provide us runtime environment for developing and running multi-tiered, reliable and secure network application – called enterprise applications. We use well accepted Oracle's Glassfish application server [12].

IV. PERFORMANCE BENCHMARK

In order to verify system functionalities and measure its capabilities, several tests have been conducted.

A. Communication chain test

This test verifies end to end communication within the TeleAssistant. Messages are transmitted from wireless fall detector to application server and it is vital that every message reaches its destination. Reliability of communication channel is given as ratio of sent and received messages. For this test wireless fall detector was configured to send alarm every minute. Messages are being logged on communication gateway and application server, and upon test completion, these two logs are compared. Results of this test are presented in Table II.

TABLE II: END TO END COMMUNICATION TEST RESULTS

Test duration time	22h
Sending message interval	60s
Number of sent messages	1304
Number of received messages	1304
Send/receive ratio	100%
Average delay time	1.12s
Maximum delay time	5s

B. Wireless fall detector energy consumption

In order to compute fall detector autonomy estimate, device current consumption is measured in both active and sleep states, and during wireless transmissions. These measurements together with the battery capacity are inputs into the autonomy equation, as shown in Table III. Autonomy is a function of a parameter p_{active} which represents an amount of time person is active (moving) as opposed to time when person is not active (stands or sits calmly, sleeps). From an elderly person we can expect this parameter to be in a range between 20 % and 50 % in which case we estimate device autonomy between 7,5 and 3,5 days.

TABLE III: DEVICE AUTONOMY ESTIMATE

Average static current, p_{active} = 30%	3,03 mA
Battery voltage	3,3 V
Frequency of sending messages	60 min
Battery capacity	400 mAh
1h energy consumption	36 J
Autonomy	5,49 days

C. Fall detection rate

This test verifies fall detection performance in terms of sensitivity and specificity. Sensitivity indicates how many true fall events will be detected. Specificity indicates how many events detected as falls are true falls. Performance was tested under laboratory conditions according to standard methodology seen in literature. Ten test subjects were asked to act out three types of fall on the mat (forward fall, fall on the back, fall on the side) and three types of activities of daily living (ADL): sit on a chair, sit on a toilet, lying down in the bed. These activities were conducted twice, by all of the 10 test subjects. Test results are presented in Table IV.

TABLE IV: FALL DETECTION PERFORMANCE TEST RESULTS

Number of volunteers	10
Total falls	60
Total ADL	60
Sensitivity	97%
Specificity	95%

D. Wireless signal coverage – range test

This test estimates ZigBee signal coverage between the gateway and fall detector. The test was conducted using two types of Digi XBee modules (XBee and XBee Pro) in an indoor environment which layout is shown in Figure 4. The fall detector was considered to be within range if the test message was received by the gateway within 3 seconds. XBee Pro range was measured at 24 m, where for XBee we measured 18 m range. For reference, black rectangle representing an apartment of 150 m^2 is plotted in the Figure 4. Range can be extended using developed ZigBee range extender.

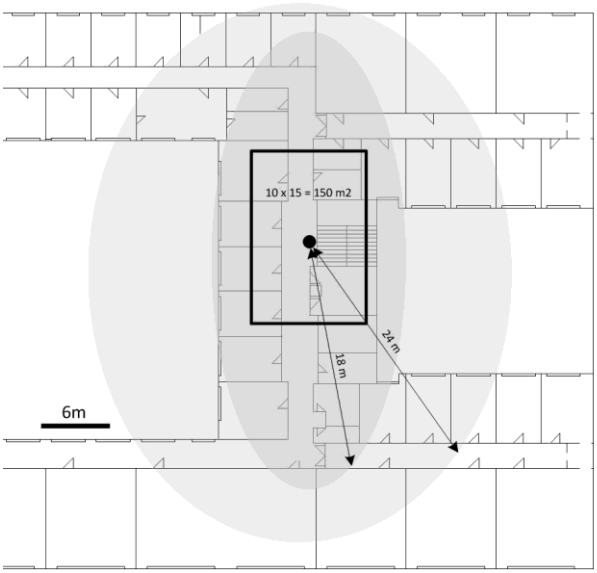


Figure 4: Range test results

V. CONCULSIONS

Main goal of TeleAssistant project was set in Assisted Living context by increasing healthcare standard for the incapable and elderly persons. This goal was fulfilled with development of the functional and verified system for the surveillance and fall detection. With this system we can monitor person's state and detect possible fall and therefore ease its consequences or even avoid death.

Within TeleAssistant project we have developed platform for demonstrating wireless technology that can be extended with additional functionalities like: managing energy efficiency, monitoring home appliances and increasing life standard.

TeleAssistant system was developed with the expertise of colleagues from FER, and it was implemented within the planned timeframe. Built prototype serves as a proof of concept and with additional enhancements can become a possible extension of Hrvatski Telekom service portfolio. Collaboration with FER provided Hrvatski Telekom with valuable knowledge of wireless sensor networks and shows us the insight for the future implementation of *Internet of Things* [13].

REFERENCES

[1] S. Abbate, M. Avvenuti, P. Corsini, J. Light, and A. Vecchio, Wireless Sensor Networks: Application-Centric Design. InTech, Dec. 2010, pp. 1–17. [Online]. Available: www.intechopen.com

[2] C. Lord and D. Colvin, "Falls in the elderly: Detection and assessment," in Proc. of the Annual Inter. Conf. of the IEEE Engineering in Medicine and Biology Society, vol. 13, Oct. 1991, pp. 1938–1939.

[3] C.-C. Wang, C.-Y. Chiang, P.-Y. Lin, Y.-C. Chou, I.-T. Kuo, C.-N. Huang, and C.-T. Chan, "Development of a fall detecting system for the elderly residents," 2nd Inter. Conf. on Bioinformatics and Biomedical Engineering, ICBBE 2008, May 2008, pp. 1359–1362.

[4] A. K. Bourke, J. V. O'Brien, and G. M. Lyons, "Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm," Gait& Posture, vol. 26, no. 2, pp. 194–199, 2007.

[5] T. Zhang, J. Wang, P. Liu, and J. Hou, "Fall detection by embedding an accelerometer in cellphone and using KFD algorithm," International Journal of Computer Science and Network Security, vol. 6, no. 10, pp. 277–284, Oct. 2006.

[6] M. J. Mathie, J. Basilakis, and B. G. Celler, "A system for monitoring posture and physical activity using accelerometers," 3rd Annual Inter. Conf. IEEE Engineering in Medicine and Biology Society, Istanbul, Oct. 2001.

[7] D. Karantonis, M. Narayanan, M. Mathie, N. Lovell, and B. Celler, "Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring," IEEE Transactions on Information Technology in Biomedicine, vol. 10, no. 1, pp. 156–167, Jan. 2006.

[8] C.-C. Yang and Y.-L. Hsu, "A review of accelerometry-based wearable motion detectors for physical activity monitoring," Sensors, vol. 10, no. 8, pp. 7772–7788, 2010.

[9] Zigbee Alliance Homepage, [Online]. Available:www.zigbee.org, 2011.

[10] J.-S. Lee, Y.-W. Su, and C.-C. Shen, "A comparative study of wireless protocols: Bluetooth, uwb, zigbee, and wi-fi," *33rd Annual Conference of the IEEE Industrial Electronics Society, 2007. IECON 2007.*, Nov. 2007, pp. 46–51.

[11] OpenSSL Cryptography and SSL/TLS Toolkit, [Online]. Available: www.openssl.org, 2011.

[12] GlassFish - Open Source Application Server, [Online]. Available: glassfish.java.net, 2011.

[13] Internet of Things – Europe, [Online]. Available: www.internet-of-things.eu, 2013.

RADIO COMMUNICATIONS SYSTEMS

Properties of Hybrid FSO/RF Link with 60 GHz RF backup link

M. Tatarko[1], Ľ. Ovseník[2] and J. Turán[3]
[1,2,3] Technical University of Košice, Department of Electronics and Multimedia Communications, Košice, Slovakia
matus.tatarko@tuke.sk[1], lubos.ovsenik@tuke.sk[2], jan.turan@tuke.sk[3]

Abstract - Free Space Optic (FSO) is an optical wireless technology which uses optical beam for communication and transmission of data. FSO communication link offers gigabit data rates and low system complexity. On the other hand, there are negative atmospheric effects which have a bad influence on availability and reliability of connection. The main bad atmospheric effect is caused by fog due to similar diameter of fog particles which correspond with used wavelength in FSO. Radiofrequency (RF) links have lower data rates, but enable transmission of data in foggy atmospheric conditions. In hybrid FSO/RF lines is able to combine advantages of both systems and achieve higher link availability. This paper deals with atmospheric effects on the FSO and 60 GHz RF link.

I. INTRODUCTION

Hybrid FSO/RF communication link consists of two main systems [1]. 1) Free Space Optics (FSO), which provides optical connection. 2) Radio Frequency (RF) link, which provides back up of primary link. Availability and reliability of this hybrid communication link is better than for separate links. Important part of FSO/RF is switch or router, which makes switching between FSO and RF link. In this paper we will deal with advantages and drawbacks of FSO communication link. Second part of this paper deals with advantages of 60 GHz RF link and impact of atmosphere to RF communication link.

II. CHARACTERISTIC OF HYBRID FSO/RF LINK

The FSO communication link we can define as telecommunication technology which use infrared optical beam to transmit information between two static points. It is a broadband communication technology which need direct visibility between transmitter and receiver, called Line of Sight (LOS). This technology uses optical modulated pulses for fibreless optical data transfer [2]. In free space optics light pulses are transmitted via atmosphere. Due to atmosphere conditions FSO technology has a number of drawbacks. Whenever the medium´s condition is unstable and unpredictable, it becomes difficult to manage the FSO transmissions. The main causes of disruptions are fog, absorption, scattering and scintillation.

Secondary negative influences on FSO link are caused by physical obstructions, buildings sway, rain and snow.

In RF communication link, data is transmitted through the air too, but by digital radio signals at a given frequency, typically in the 9 kHz to 300 GHz range [2]. In combination with FSO systems, RF systems use 60 GHz antennas. Beam width for antenna´s diameter 30,48 cm is 4,7°. This 60 GHz antenna allows using high speed bit rate and connection has high resistance to interference, high security and multiple use of frequency [3]. Factors which have negative effects on FSO are negligible for RF systems. This is a main reason, why are RF links used as backup links. For FSO systems, fog and with fog join visibility is critical factor for availability and reliability of them. On the other hands, RF links show almost negligible fog attenuation, while they usually suffer from other precipitation types like rain or wet snow. Combining these two technologies to FSO/RF hybrid network may increase overall availability, guaranteeing quality of service and broadband connectivity regardless of atmospheric conditions [1]. The RF link should be available whenever the FSO link is not i.e. it should not be influenced by fog or other weather effects reducing visibility.

III. VISIBILITY AND FOG SENSOR

As was mention above, visibility has major influence on overall availability of hybrid FSO/RF link. Thus, it is important to know value of visibility in place where the hybrid link will operate. We can compute visibility [km] from information obtain from fog sensor. It measures density of fog [g/m^3], relative humidity [%] and temperature [°C]. These three parameters are important for subsequent steady and statistical evaluation of the quality of FSO/RF communication environment [4].

Fog sensor consists of two main hardware parts, which can be divided to Outdoor and Indoor unit (Fig.1). Outdoor unit is connected with Indoor unit via 8 pins cable with RS- 422 line and Indoor unit is connected with computer via serial cable RS 232. Outdoor unit performs the measurement and it has not need external power [4].

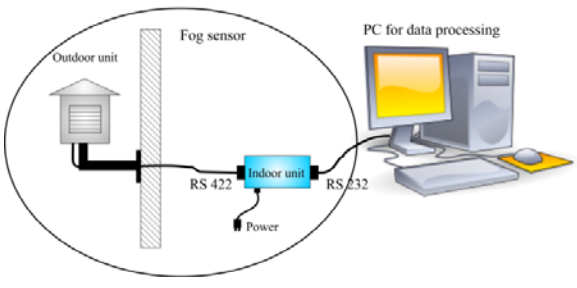


Figure 1. Hardware scheme of fog sensor

Obtained data are processed in PC in simulation program Matlab that allows us to obtain other parameter, which describes fog. This parameter is marked like *LWC* and it means liquid water content of the atmosphere. It describes fog as mass of water drops in the volume units. It is measured in [g/m^3]. From measured data of fog sensor we can calculate *LWC*, but we have to pre-process measured data of density of fog. From *LWC* we can calculate a visibility *V* [km]:

$$V = d \cdot (LWC)^{-0,65}, \quad (1)$$

where *d* is a parameter (without unit) and takes on specific values for different fog conditions as is shown in TABLE I. below [5]:

TABLE I. TYPICAL VALUE OF PARAMETER "*d*" FOR DIFFERENT TYPES OF FOG

Type of fog	Parameter "*d*"
Dense haze	0,013
Continental fog (dry and cold)	0,034
Maritime fog (wet and warm)	0,060
Dense haze and selective fog	0,017
Stable and evolving fog	0,024
Advection fog	0.02381

In this case, we have mentioned stable and evolving fog. So visibility is given by equation [5,6]:

$$V = 0,024 \cdot (LWC)^{-0,65}, \quad (2)$$

IV. AVAILABILITY OF HYBRID LINK

Availability of FSO link strongly depends on fog and its influence on communication link. In this case fog is represented by visibility. For FSO communication link, there are two possibilities how to obtain information about overall availability. First, we can capture information about fades and then percentual evaluate data during whole year. Another way how to determine availability of FSO link is capture information about visibility in place, where the link will operate. For second way, the fog sensor is a good device for measuring visibility. In this paper, two FSO systems will be used for calculating availability in campus of Technical University in Košice (TUKE). First FSO system is Lightpointe, Flightstrata 155E, second system is FSONA, Sonabeam 155 E. Edge values of visibility for 1 km FSO link is calculated by software package FSO SystSim, which was created at TUKE. This software calculates availability for specific systems by mathematic models and energy balance of FSO connections. After entering inputs parameters of individuals systems, edge values of visibility was calculated. Edge visibility for 1 km long FSO link is 515m for Lighpointe 155E and 620m for Sonabeam 155 E. From this distances is known, that Sonabeam will have worst availability than Lightpointe.

On the other hand, the main drawback of RF link is caused by rain. If we want to determine availability of RF link for 1 km, we need to know how often per year are rainy days. Measuring about fallen precipitations is performed worldwide. With these results of measurement, models of world rainy zones have been proposed (Fig.2). International Telecommunication Union has classified these rainy zones in different regions around the world. From Fig. 2 we can see that Slovakia belongs to rainy zone letter H.

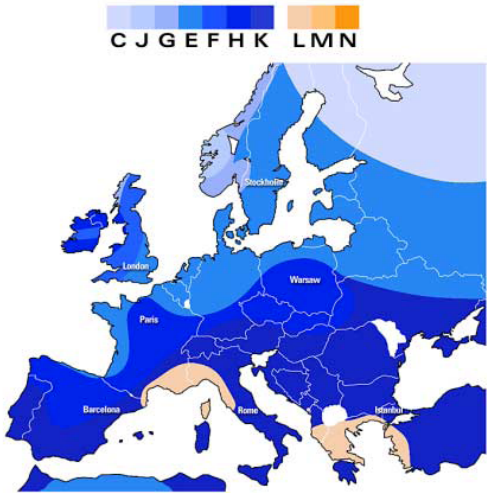
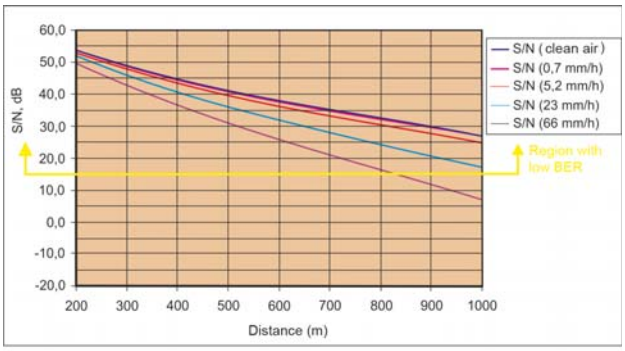


Figure 2. Rainy zones of Europe

For 60 GHz RF link 33 cm antenna is used from HXI Company. From Fig. 3 we can see an availability of 60 GHz link for 1 km. It is 99,9 %.



Availability of system	Volume of precipitations (Region H) mm/Hr	Max. distance for BER=10^{-7}
99	0,7	1240 m
99,9	5,2	1050 m
99,99	23	950 m
99,999	66	730 m

Figure 3. Availability of 60 GHz antenna for 1 km

In these days communication system is reliably when its availability is 99,999 % (only 318 seconds per year can be link down).

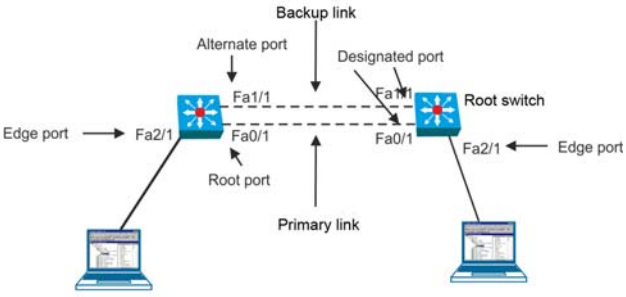


Figure 4. Hybrid FSO/RF link

When we join these two communication systems together to hybrid FSO/RF link it is possible to obtain five 9´s availability (Fig 4).

These two links are switching by CISCO switch with RSTP protocol. It is important to connect primary link to lower ports on both switches. Time required to switch between links is 6 seconds.

The sample calculation of availability of hybrid link is shown below for year 2010:

$$D_{FSO} = \left(1 - \frac{time\,of\ \ fades}{total\,time}\right) \cdot 100\,\%\,, \tag{3}$$

$$D_{FSO} = \left(1 - \frac{304}{2.8760}\right) \cdot 100 = 98{,}1566\,\%\,, \tag{4}$$

where D_{FSO} is availability of FSO Flighstrata 155E. *Total time* is given like two multiply hours per year due to half an hour measuring interval [3].

$$D_{RF} = \left(1 - \frac{time\,of\ \ fades}{total\,time}\right) \cdot 100\%\,, \tag{5}$$

$$D_{RF} = \left(1 - \frac{304{,}08466}{8760}\right) \cdot 100 = 96{,}5287\,\%\,, \tag{6}$$

where D_{RF} is availability of RF link. *Total time* is given like hours per year. *Time of fades* is given by [3]:

$$time\,of\ \ fades = NS \cdot TS + TF_{FSO}\,, \tag{7}$$

where *NS* is number of switching RF link, *TS* is total time, TF_{FSO} is time of fades of FSO.

$$time\,of\ \ fades = 51 \cdot 0{,}00166 + 304\,. \tag{8}$$

Other calculated values for FSO systems and for RF system are entered in the next two tables:

TABLE II. AVAILABILITY OF FSO LINKS

FSO system	Year	Availability
Flightstrata155E	2007	98,1621
Sonabeam 155-E		97,1095
Flightstrata155E	2008	98,3904
Sonabeam 155-E		98,1963
Flightstrata155E	2009	98,5308
Sonabeam 155-E		98,3288
Flightstrata155E	2010	98,1566
Sonabeam 155-E		97,9019
Flightstrata155E	2011	98,2368
Sonabeam 155-E		98,1287

Overall availability of hybrid FSO/RF link using FSO Flightstrata 155E is:

$$D_{FSO/RF} = \left[1 - \left(1 - D_{FSO}\right) \cdot \left(1 - D_{RF}\right)\right] \cdot 100 = \\ = 99{,}99832\,\%. \tag{9}$$

TABLE III. NUMBERS OF SWITCHING RF LINK

Year	Flightstrata 155E	Sonabeam 155-E
2007	48 times	62 times
2008	44 times	46 times
2009	48 times	49 times
2010	51 times	58 times
2011	49 times	53 times

Overall availability of hybrid FSO/RF link using FSO Sonabeam 155-E is 99,99809 %.

V. CONCLUSION

Hybrid FSO/RF link uses both links strength to overcome their weaknesses. By combining these two systems will the hybrid FSO/RF link be able to achieve higher availability of transmission lines as an independent FSO or RF line. Because of the RF hybrid FSO/RF can reduce the impact of heavy fog and FSO can reduce the impact of the rainfall. Dense fog is the most important adverse factor of FSO links and rain is the most important adverse factor of RF connections. Advantages of using FSO link in case of lost connection may not be lost due to using a 60 GHz RF system as back-line which has similar properties than FSO link. Average availability of hybrid FSO/RF system in Košice is 99, 998205 %

ACKNOWLEDGMENT

We support research activities in Slovakia / Project is co-financed from EU funds. This paper was developed within the Project "Centrum excelentnosti integrovaného výskumu a využitia progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055.

REFERENCES

[1] D. M. Forin, G. Incerti, G. M. Tosi Beleffi, B. Geiger, E. Leitgeb, F. Nadeem, „Free Space Optical Technologies“, Trends in Telecommunications Technologies, March 2010.

[2] M. Ahmad, A. Fuqaha, O. Awwad, B. Khan, “Synergies of Radio Frequency and Free Space Optics Communication: New Hybrid Solutions for Next Generation Wireless Mesh Networks”, International Journal of Computer Networks (IJCN), Vol. 4, 2012.

[3] R. Luna, D. K. Borah, H. Tapse, „Behavior of hybrid optical/RF channels...“, *Proc. IEEE GLOBECOM*, Honolulu, HA, November-December 2009.

[4] Ľ. Ovseník, J. Turán, P. Mišenčík, J. Bitó, L. Csurgai-Horváth, „Fog density measuring system“, Acta Electrotechnica et Informatica, 2012, 1-5.

[5] S. Sheikh Muhammad, M. Saleem Awan, A. Rehman. „PDF Estimation and Liquid Water Content Based Attenuation Modeling for Fog in Terrestrial FSO Links”. Radioengineering, Vol. 19, No. 2, June 2010.

[6] V. Slodičák, „Some useful structures for categorical approach for program behavior”, Journal of Information and Organizational Sciences, Vol. 35, No. 1, 2011, 99-109.

Comparison of New Technologies for Data Exchange in the Maritime HF Frequency Band

Sanjin Valčić*, Zoran Mrak * and Danko Kezić**
* Faculty of Maritime Studies, Rijeka, Croatia
** Faculty of Maritime Studies, Split, Croatia
svalcic@pfri.hr, mrak@pfri.hr, danko@pfst.hr

Abstract - Recently, the use of HF radiotelex on ships, as the only terrestrial communication system for data exchange has been kept to a minimum, and the need for alternative systems, which would use the existing HF communication channels to exchange data between ships and end users on land, arose. Satellite communication systems represent the main aid for the exchange of large amounts of data with high transmission rates, but at high costs of equipment and services. Hence, there is the need to explore new, simpler and cheaper technologies for data exchange, which could be an alternative or a substitute for the existing radiotelex system. In this paper, a detailed analysis of the two systems, which could easily replace the existing maritime NBDP (*Narrow-Band Direct Printing*) telegraphy, has been done. One of the systems is based on the OFDM (*Orthogonal Frequency Division Multiplex*) protocol, while the other, on the PACTOR-III protocol. The systems were compared on the basis of the various relevant parameters, such as the modulation method, data rate, etc. The analysis results can be used in the selection of the optimal data exchange system with respect to the specific features of maritime communications.

I. INTRODUCTION

The radiotelex system, in the maritime environment, has been originally used for routine (commercial) direct communications between ships and between ships and land end users via coastal radio stations on the MF and HF frequency band. With the introduction of the GMDSS (Global Maritime Distress and Safety System) in 1992, radiotelex becomes mandatory for ships sailing in the sea area A4 to perform all communication priorities (distress, urgency, safety and routine) and an option for sea area A3. The main reason for this introduction was to overcome the language difficulties that may arise during voice communications relating to distress, urgency and safety.

The changes in communication systems on land, resulting from the introduction of new technologies, have caused the cessation of the radiotelex usage onboard vessels for commercial communications. Information encoding in the radiotelex system is not adapted to new forms of data exchange, such as E-mail, SMS, etc. Moreover, the IMO (International Maritime Organization) has determined that the use of radiotelex in the sea area A3, in terms of security, is no longer necessary, while it is still required for the sea area A4 because of its robustness if compared to voice communications in the HF frequency band [1]. Also, a great reduction in the performance of voice communications in the maritime HF frequency band has been noticed, thus leaving a large part of HF frequency spectrum unused. The reason for this lies in the development of commercial satellite systems, which are much easier to operate and which provide a wider range of services when compared to terrestrial HF systems. However, due to the inaccessibility of the price and the lack of satellite coverage in areas above 70° N and S latitude, which are becoming more attractive for research, as well as for navigation, new competitive terrestrial systems, that would be applicable in maritime environment for the exchange of small quantities of data, are being developed.

For the purpose of an efficient utilization of frequency spectrum, the IMO and the ITU (International Telecommunication Union) encourage and propose new systems for sending text messages and data exchange, adapted to the needs of maritime environment, which could easily replace the existing maritime radiotelex system and thus become competitive to satellite systems. Two such systems, data services HF modem protocol using orthogonal frequency division multiplexing (OFDM) and Electronic mail system using PACTOR-III protocol, selected on the basis of direct applicability to the existing maritime HF frequency band, were thoroughly analyzed and compared in this paper [1][2][3].

Systems are based on the multicarrier modulation, which is less susceptible to multipath fading, has higher spectral efficiency and is easier to implement than the single carrier modes. However, they require the use of powerful DSP processors. Moreover, they are sensitive to the frequency offset, as well as to the phase noise of oscillators and they have peak-to-average problem, which reduces the efficiency of the RF power amplifiers on the transmitter. On the other hand, they retain the quality of service in the harshest propagation conditions typical for the HF frequency band.

II. THE OFDM PROTOCOL

The orthogonal frequency division multiplexing or OFDM is a modulation procedure that uses multiple separate subcarriers inside a particular communication channel. The orthogonality between subcarriers implies that the frequency shift is inversely proportional to the duration of the transmitted symbols. The number of subcarriers depends on the environment in which the system will be used. The OFDM concept has been known

since the 70-ies of the last century, but it has been impossible to implement until the development of the DSP processors, which enabled the performance of IFFT (Inverse Fast Fourier Transform) and FFT (Fast Fourier Transform) algorithms. The digital modulation method of subcarriers can be performed using QPSK (Quadrature Phase Shift Keying) or QAM (Quadrature Amplitude Modulation). Due to their robustness, the QPSK modulation is used for transmission in poor channel conditions, while the QAM for better channel quality [4][11][14].

The HF data modem services protocol using OFDM, proposed by the ITU, uses 32 subcarriers with the central frequency at 1700 Hz. Each subcarrier is modulated separately by the DQPSK modulation method, in which one modulation level carries two bits of information. The spacing between subcarriers is 83.33 Hz, which means that the duration of symbols is 12 ms. The required bandwidth of the frequency spectrum is 2666.67 Hz, which fits into the existing SSB (Single Side Band) channel of 3 kHz bandwidth in the HF frequency band.

The proposed OFDM protocol is a synchronous half-duplex ARQ (Automatic Repetition reQuest) protocol, with one cycle duration of 1.52 s. The structure of these cycles provides establishing the connection between stations at distances up to 40,000 km. One cycle includes transmitted packets (data burst) and confirmations of the receipt. In order to preserve the integrity of data and to verify the correctness of the receipt, the protocol uses a CRC (Cyclic Redundancy Check).

Furthermore, in order to achieve the required orthogonality of subcarriers within the modulation procedure, the IFFT is used, and therefore, on the receiving end within the demodulation procedure, the FFT must be performed. For the reduction of the peak-to-average problem, which depends on the number of subcarriers and occurs in all types of multicarrier modulation, the OFDM protocol-based modem uses a scrambler algorithm for the randomization of subcarrier phases. The problem caused by the sensitivity of the receiver to the frequency offset is solved by applying the frequency recovery algorithm for the central frequency in the tolerance range of ± 50 Hz.

The proposed OFDM protocol provides an overall bit rate of 4613 bps and the effective bit rate of 2877 bps at normal propagation conditions. The ITU in its recommendation did not anticipate the use of any type of online data compression, which could achieve higher effective bit rates.

In order to establish a connection between stations, the use of robust FSK (Frequency Shift Keying) modulation is predicted. Once the connection is established, the system modems automatically switch to the OFDM protocol for transmitting data [2].

III. THE PACTOR PROTOCOL

The PACTOR protocol is a protocol for exchanging data over the radio waves on the HF frequency band, was developed in Germany in 1991, with the aim to overcome the disadvantages of the radiotelex and the Packet-Radio system in poor propagation conditions. Important features of both systems are contained in the PACTOR protocol: the CRC to preserve the integrity of the data from the Packet-Radio, and the synchronous exchange of short data blocks from the radiotelex. The PACTOR protocol introduced the new feature of using online data compression (Huffman code) to achieve much higher data rates. A key enhancement that was introduced in this protocol was the use of half-duplex memory ARQ, in which the connection maintains established even in bad propagation conditions. The operational principle of the memory ARQ is as follows: defective packets or blocks of data are stored on the receiving end until enough data is collected in order to reconstruct the original packet or block [5][6].

As the system developed with further advancements in the technology, three types of PACTOR protocol were introduced: PACTOR-I, PACTOR-II and PACTOR-III.

A. The PACTOR-I protocol

PACTOR-I was the first version of the PACTOR protocol and it used a robust two tones FSK (Frequency Shift Keying) modulation with a shift of 200 Hz between tones. In this protocol, the data rate and the modulation method were adjusted to the channel bandwidth of 600 Hz. Further effective increase in the data rate was provided by the use of online Huffman compression. The protocol is suited to the 8-bit ASCII code, unlike the radiotelex system which uses a 5-bit code. Using this protocol, the maximum distance between stations on which the communication can be established can reach 20,000 km. Moreover, the protocol can be used in extremely poor propagation conditions, where the signal to noise ratio SNR can be up to -4 dB, due to the use of memory ARQ. The technical characteristics of the PACTOR-I protocol are shown in Table I [5][6].

TABLE I. TECHNICAL CHARACTERISTICS OF THE PACTOR-I PROTOCOL

Baud rate	100 or 200 bauds depending on propagation conditions
Modulation method	two tones FSK with a shift of 200 Hz
Online compression	Huffman compression
Channel bandwidth	600 Hz
Total cycle duration	1.25 s
Maximum distance at 100 bauds	20 000 km
Lowest S/N at 100 bauds	-4 dB

B. The PACTOR-II protocol

PACTOR-II protocol, just like PACTOR-I, is based on the synchronous half-duplex ARQ protocol and uses various methods of modulation based on DPSK (Differential Phase Shift Keying) and convolutional coding depending on the quality of the channel. Moreover, PACTOR-II uses two-tone DBPSK modulation in its most robust mode, which can be thought of as two single-tone DBPSK signals used together, spaced at 200 Hz, resulting in a total data rate of 200 bit/s. Depending on the quality of the channel, a more complex multi-level modulation

methods can be used (Table II.), to increase the data rate which require the application of DSP processors.

TABLE II. MODULATION METHODS OF THE PACTOR-II PROTOCOL

Modulation method	Total bit rate	Convolutional code rate	Net bit rate
DBPSK	200 bps	1/2	100 bps
DQPSK	400 bps	1/2	200 bps
8-DPSK	600 bps	2/3	400 bps
16-DPSK	800 bps	7/8	700 bps

Beside the use of the Huffman compression, like in the PACTOR-I protocol, there is also an alternative, the PMC (Pseudo-Markov Compression), which can achieve speeds up to 1200 bps. The required channel bandwidth, to implement and use the PACTOR-II protocol, is 500 Hz, which corresponds to the existing channels of the radiotelex system in the maritime HF frequency band.

The improvement in comparison to the PACTOR-I protocol, is expressed in the introduction of a long-range mode, with the total time cycle of 1.4 s, for the operation at distances up to 40 000 km and a data mode, with a total cycle time of 3.75 s, for transmitting packets of maximum size of 2208 bits. Unlike the PACTOR-I protocol, the PACTOR-II protocol uses convolutional coding, which enables its usage in extremely poor propagation conditions in the communication channel, with the SNR = -18 dB [5][6].

C. The PACTOR-III protocol

The PACTOR-III protocol is also a synchronous half-duplex ARQ protocol designed for commercial purposes in order to provide higher transmission speeds and better robustness than PACTOR-II, while using the whole SSB channel of 3 kHz bandwidth. The maximum of 18 subcarriers (frequency tones), with 120 kHz spacing can be used, and their number depends on the propagation conditions in the channel. Moreover, the PACTOR-III protocol uses 6 speed levels, with different methods of modulation and convolutional coding. In Table III., SL stands for the speed level, NT for the number of tones, MM for the modulation method, CCR for convolutional code rate, PDR for the physical data rate and NDR for net data rate.

TABLE III. SPEED LEVELS OF THE PACTOR-III PROTOCOL

SL	NT	MM	CCR	PDR	NDR
1	2	DBPSK	1/2	200 bps	76.8 bps
2	6	DBPSK	1/2	600 bps	247.5 bps
3	14	DBPSK	1/2	1400 bps	588.8 bps
4	14	DQPSK	1/2	2800 bps	1186.1 bps
5	16	DQPSK	3/4	3200 bps	2039.5 bps
6	18	DQPSK	8/9	3600 bps	2722.1 bps

The maximum channel bandwidth, when using 18 tones, is 2.2 kHz (400 - 2600 Hz). The center frequency is 1500 Hz, the lowest subcarrier is 480 Hz, and the highest subcarrier is 2520 Hz. In Figure 1., SL stands for speed level, CN for channel number, TF for tone frequency [Hz], and 'x' indicates that the tone is used in the respective speed level.

	CN	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
SL																			
1							x							x					
2					x		x		x			x		x		x			
3				x	x	x	x	x	x	x	x	x	x	x	x	x	x		
4				x	x	x	x	x	x	x	x	x	x	x	x	x	x		
5			x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
6		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	TF	480	600	720	840	960	1080	1200	1320	1440	1560	1680	1800	1920	2040	2160	2280	2400	2520

Figure 1. Number and position of used tones at different speed levels

The PACTOR-III protocol uses either the DBPSK or the DQPSK modulation method, depending on the number of subcarriers, i.e. the speed level. In order to achieve the highest data rate at speed level 6, it is necessary that the SNR of the channel is at least 14 dB depending on the modulation method. The total duration of a cycle and the transmission modes remain the same as in the PACTOR-II protocol. In PACTOR-III protocol, online Huffman and PMC compression can also be used and it is possible to achieve a maximum data rate up to 5.2 kbps under normal propagation conditions.

Using PACTOR protocols, connections between stations are established on the most robust PACTOR-I protocol, and then, depending on the conditions in the channel, the system modem automatically changes the protocol type and speed level [7].

IV. COMPARISON OF RELEVANT PARAMETERS OF OFDM AND PACTOR-III PROTOCOLS FOR POSSIBLE APPLICATIONS IN THE HF FREQUENCY BAND

Both protocols, the OFDM and the PACTOR-III, are designed based on the synchronous half-duplex ARQ protocol, which allows high reliability when exchanging data in a relatively poor propagation conditions typical for the HF frequency band, such as selective fading, multipath propagation, Doppler spread, signal delay etc. [13]. Moreover, both protocols are based on the multicarrier signal modulation method, which further reduces the impact of these factors. However, by using multiple separate subcarriers, the systems become more sensitive to the frequency offset and phase noise of oscillators [12]. In addition, the peak-to-average problem also occurs. Therefore, both protocols require the use of linear output amplifiers, which are already in use on ships to carry out communications in the HF frequency band, in order to avoid signal distortions. There are methods for improving the peak-to-average ratio in both systems. PACTOR-III provides considerably more transmitter output power than OFDM protocol, when using the same power amplifier, thereby increasing the S/N at the receiver. The peak-to-average ratio in PACTOR-III protocol depends on the speed level and ranges from 1.9 to 5.7 dB. Even at speed levels 5 and 6 this ratio is 3 dB lower than in the OFDM protocol. This means that the mean transmitted power of the PACTOR-III protocol is twice the mean transmitted power of the OFDM protocol in the same propagation conditions.

Furthermore, by the use of existing maritime HF transmitters whose maximum allowable peak emitted power is up to 1200 W, both protocols can be used successfully to perform communications up to maximum distances of respective protocols.

In order to synchronize subcarrier frequencies between the transmitter and the receiver, the PACTOR-III protocol uses two FSK connect tones at 1400 and 1600 Hz, which are symmetric with the respect to the center frequency of 1500 Hz. The OFDM protocol, which is more sensitive to frequency offset, uses the frequency recovery algorithm to continuously adjust the center frequency of 1700 Hz in steps of 1 Hz.

The PACTOR-III protocol, depending on the speed level, uses a different number of subcarriers (from 2 to 18) and modulation methods, which can be either two or four level DPSK. The OFDM protocol always uses 32 subcarriers with a four level DPSK modulation to preserve the orthogonality of subcarriers, in which the reference phase is unrequired for the demodulation at the receiver.

For a successful transmission and reception, the OFDM protocol uses the classic ARQ mode in which data blocks received with errors are simply rejected and retransmission is required until a valid receipt is achieved. The PACTOR-III protocol uses the memory ARQ mode to store data blocks with errors until there is enough collected data to reconstruct the original packets from all stored incorrect blocks. This is the reason why the PACTOR-III protocol is able to maintain the connection in extremely poor propagation conditions. Furthermore, in the PACTOR-III protocol, the transmission rate is automatically adapted to the quality of the communication channel, thus also reducing the error probability.

Effective transmission rates of both protocols, which can be compared within normal propagation conditions, are almost equal (PACTOR-III 2722 bps without compression online, OFDM 2876 bps). When compared to the maritime radiotelex system, which achieves 100 bps, the use of these protocols enables the exchange of text messages, E-mail messages, and the sending of small amounts of data over long distances at significantly higher transmission rates. Moreover, both systems are designed for use on 3 kHz channel bandwidths and so can easily be implemented in the maritime HF frequency band without redesigning existing channels.

V. CONLUSION

For purposes of data exchange in the maritime HF frequency band, the ITU does not propose the use of only one system, but, in order to encourage technological progress and competitiveness, defines the basic requirements that systems must meet. These basic requirements include the use of existing channels of 3 kHz bandwidths, interoperability at provider level during ship-to-shore and shore-to-ship communications, compatibility with existing maritime HF transceivers, and an automatic adjustment to the existing ARQ and FEC (Forward Error Correction) protocols needed for ship-to-ship communications in the GMDSS system [1].

Protocols compared in this paper are implemented in software-defined radio systems, which bring economical, technical and spectrum efficiency benefits. Moreover, it is possible to implement such radio systems without further regulatory changes in the maritime environment. These software-defined radio systems are already in use for commercial purposes of HF data exchange in the maritime environment. Each of the systems consists of a series of coastal stations, distributed in such a manner which provides global coverage, and mobile stations (ship systems). Every coastal station uses an allocated variety of channels in the HF frequency band. They are automatically selected by the ship equipment, depending on channel conditions, in order to perform best quality communications. This allows the operators to significantly facilitate the use of the HF frequency spectrum, without knowing the propagation conditions. The operational principle of both systems is based on the mediation of coastal stations in the exchange of data between ships and land end users using the IP protocol. Some of the services provided by these systems are electronic mail, SSAS (Ship Security Alert System), LRIT (Long Range Identification and Tracking), reception of maritime safety information and weather forecasts, etc. [8][9][10].

However, both systems are still being developed, in order to increase data rates, and hence the amount of data. A system using the OFDM protocol is modified in such a way to reduce the connection time, whereas a system using the PACTOR protocol introduces more speed levels and more complex methods of modulation and convolutional coding (PACTOR-IV protocol). Furthermore, with the increase in the number of users, a need for additional communication channels for data exchange also increases. For this purpose the existing 3 kHz bandwidth channels from the maritime HF frequency band could be very easily utilized without any modifications. On the other hand, the existing 500 Hz bandwidth channels could be used by systems operating on PACTOR-II protocol in poor propagation characteristics, during which it is not possible to establish contact using OFDM or PACTOR-III protocol.

Future work should focus on the creation of an integrated maritime communication system, which would include all terrestrial and satellite systems. Such system should automatically select the most suitable means to establish the communication, depending on the desired type of communication, propagation conditions in the channel, the distance between the users, the emitted power, etc.

ACKNOWLEDGMENT

The authors acknowledge the support of research project “Information-communication technologies in smart maritime systems” (112-1121722-3314) funded by the Ministry of Science, Education and Sports of the Republic of Croatia.

REFERENCES

[1] IMO, “Replacement for use of NBDP (radio telex) for maritime distress and safety communications in the maritime MF/HF bands”, COMSAR 12/9, 2007

[2] Recommendation ITU-R M.1798-1, “Characteristics of HF radio equipment for the exchange of digital data and electronic mail in the maritime mobile service”, Geneva, 2010

[3] IALA, “Maritime communications plan”, edition 1, Saint Germain en Laye, France, 2009

[4] L. Litwin and M. Pugel, “The principles of OFDM”, RF Signal Processing, 2001

[5] S. Ford, “Arrl's HF Digital Handbook”, USA, 2007

[6] R. Proesch, “Technical handbook for radio monitoring HF”, Germany, 2009

[7] SCS GmbH & Co. KG, “Manual for Proffesional Firmware (Version 3.2)”, Hanau, Germany, 2002

[8] Wavecom, “GW-OFDM Global Network”, Info Leaflet, 2012

[9] C. Mayer, “HF: An inexpensive complement to satellite communications”, 7th Annual Digital Ship, Athens, 2009

[10] NTIA, “High frequency radio automatic link establishment (ALE) application handbook”, USA, 1998

[11] Smart Antenna Research Laboratory, Geogia Tech, “OFDM Simulation Using Matlab”, Atlanta, 2000

[12] C. Stimming, “Frequency offset tracking in MCMA systems”, 2001

[13] Harris Corporation, “Radio communications in the digital age”, vol. 1, HF Technology, Edition 2, USA, 2005

[14] J. Terry and J. Heiskala, “OFDM Wireless Lans: A theoretical and practical guide”, Sams Publishing, USA, 2002

A modular simulation framework for analyzing coexistence issues in wireless radio networks in close spatial proximity

R. Neumeier*, G. Ostermayer *
* FH OÖ Forschungs & Entwicklungs GmbH/ Research Center Hagenberg, Hagenberg, Austria
roland.neumeier@fh-hagenberg.at, gerald.ostermayer@fh-hagenberg.at

Abstract - In many fields of technology more mobile transmission paths will be used in the future. In order to ensure a secure transmission of data, studies for coexistence of the used radio technologies with other currently wide used radio systems (WLAN, Bluetooth, ZigBee, etc.), are needed. Since an analytical treatment is too complex and a real test set is usually too expensive, the objective is the construction of a suitable simulation environment to investigate the performance of co-existing wireless systems in an economical way. Investigations of a variety of representative application scenarios indicated, that a universal applicability and customizability of such a simulation framework can be achieved by a high-grade modular design. To keep the simulation results as realistic as possible, the use of statistical radio channel data was largely omitted. Instead, an accurate recording of the simulated scenario is done to compute the required channel characteristics by ray-tracing/-launching algorithms. The signal processing chains of the transmitters and receivers (user- and jammer-radio) were reproduced precisely to their specifications and for easy interchangeability they are connected to the channel via a universal defined interface. Finally an application scenario is presented and the results of the simulations are shown and discussed.

I. INTRODUCTION

Nowadays, wireless data transmission paths are used for a large number of applications. For example, such applications could be in logistics management, control and monitoring of industrial plants as well as management and billing of electric charging stations etc., however, there are virtually no limits in application areas.

Since many other systems, that operate in the same frequency band (WLAN, Bluetooth, ZigBee, etc.), often are in spatial proximity, there may be mutual interference. Therefore one objective of this project was to explore strategies to ensure reliable and high quality transmission of the data in coexistence scenarios. Therefor an investigation of the effects of coexistence of different wireless and / or mobile radio systems in close spatial proximity is needed. Since an analytical treatment is too complex and a real test set is usually too expensive, the objective is the construction of a suitable simulation environment to investigate the performance of co-existing wireless systems in an economical way.

Investigations of a variety of representative application scenarios indicated, that the requirements for such a simulator are very different and partly opposite. For example, errors, which occur on the transmission of control data for an industrial robot, require different handling than errors on video or audio streaming transmissions. Send/receive chains that act as probed connection in one scenario can act as jammer in the next scenario. In order to manage these requirements within one universal applicable and customizable simulation environment, a high-grade modular design is needed.

To keep the simulation results as realistic as possible, the use of statistical radio channel data was largely omitted. Instead, an accurate recording of the simulated scenario (space, position transmitter/receiver/jammer, etc.) is done to compute the required channel characteristics by ray-tracing/-launching algorithms. This ray-tracing/-launching simulator was built as separate, independent module to allow multiple simulations with different radio parameters in the same environment. The signal processing chains of the transmitters and receivers (user- and jammer-radio) were reproduced precisely to their specifications. For easy interchangeability of the individual radio technologies these signal processing chains are connected to the channel via a universal defined interface.

In the first section this paper describes the motivation and the objectives to build such a simulation environment. In the next part the structure of the simulator and its components, which was fully implemented in MATLAB Simulink, is explained. Main focus is on the modular design of the simulator and the aspects of the selected division of the single modules. In this context an application scenario (Bluetooth connection is disrupted by WLAN transmitters) is presented and the results of the simulations are shown and discussed.

II. SIMULATION FRAMEWORK

This section describes the structure of the simulation framework. One objective of this project was not to build the wireless transmission channel out of statistical data. Instead the geometrical dimensions and surfaces in the investigated scenario should be used. To build a channel model out of the geometrical room data and the position of sender and receiver a ray-tracing simulator is needed. In

order to get a simulation environment that is independent from commercial, expensive ray-tracing suites, a simple and for now rudimentary ray-tracing simulator was built.

A. *Raytracer*

For most simulation scenarios it is necessary to run a simulator many times with different parameters on send/receive-chains using the same channel model. Generating a channel model from a ray-tracer is a computationally intensive process, so running this process each time a new parameter set is simulated would lead to exorbitant execution time extension. Also saving a generated channel model for later use or reproduction of a previous simulation campaign would be beneficial. For that reason the ray-tracing simulator was built as separate, independent module. One more benefit of this breakup is the possibility to exchange this still simple module with a commercial software suite for channel generation or to use data from channel measurements in real environments.

According to [1] and [2] the ray-tracing simulator was built in *Mathworks MATLAB*. The current implementation allows defining a rectangular or trapezoid room with one or more objects inside the room. These objects also have to be in rectangular or trapezoid shape. Sender and receiver can be positioned anywhere in the room. Each surface or only parts of surfaces can assigned with different reflection and transmission coefficients. With this features a simple room model is created which can be saved for later use. After starting the channel model generation the ray-tracer computes all possible ray-paths from the sender to the receiver. These paths can contain line of sight, reflection on objects or walls and transmission through objects. The current implementation does not support effects like refraction, diffraction and scattering. At each reflection or transmission point all relevant data (distance from recent to current point, attenuation, angle of incidence etc.) are stored for later analysis purposes. The calculated full paths can contain multiple reflections and transmissions from sender to receiver. This can lead up to an infinite number of possible paths, although not all of them are relevant for channel model generation, because signals over those paths would get too weak to be recognized by the receiver. To limit these paths all paths reaching a total attenuation level higher than a given threshold are omitted.

Fig. 1 shows the ray-tracer result of a very simple room with one object in it and transmission through objects disabled (enabling transmission would lead to a result not suitable for graphical representation in this paper). Out of the stored date while calculating the possible signal paths from the sender to the receiver a tapped delay line model as shown in Fig. 2 is calculated.

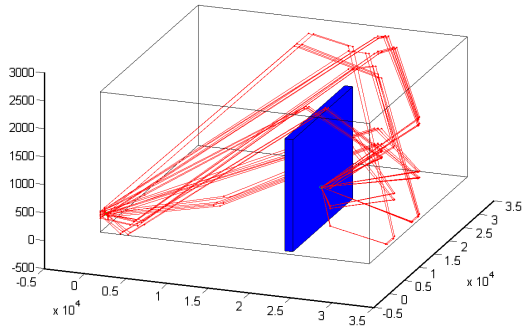


Figure 1. Result of ray-tracing simulaton: room model with signal paths.

B. *Simulator Structure*

This section explains the design of the whole simulation framework and its components as well as the aspects of the selected division of the single modules. Fig. 3 shows the block diagram of the framework. Main components of the simulator are the send/receive chains, the core simulator with signal channel management and the analysis component (throughput and bit error rate (BER) calculation).

In order to have the main engineering focus on the core simulator, the send/receive chains used in this simulator are extracted of the Simulink examples. Depending on the selected model, these examples are very accurate and with little or no modification they exactly follow the corresponding standards. In the current version Wireless LAN 802.11g [3] and Bluetooth in Version 1.1 [4] are implemented, for future research 802.15.4 [5] will also be adopted. One requirement for all of these send/receive chains is, that they have to be dividable between sender and receiver. The included channel model must be removable to be exchanged by the core simulator. Several control information between sender and receiver may remain. For interchangeability of the different wireless standards the different send/receive chains are connected to the core simulator via universal defined interface. This Interface contains the required constitution of the delivered data streams and transfer parameters (used transmission bandwidth, channel, sampling frequency etc.) needed for processing in the simulator core.

The core simulator is responsible for processing and combining the input data streams of the connected senders and forwarding them to the appropriate receivers. Minimum requirement for investigations in coexistence scenarios is one complete send/receive chain, this is the one which will be tested, and an interfering sender. This sender does not necessarily need a receiver if it works standalone. As shown in Fig. 3, multiple interferers of different technology or multiple receivers (more than one

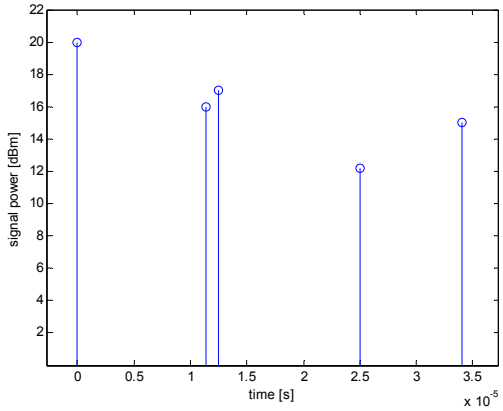


Figure 2. Raytracer result: tapped delay line model

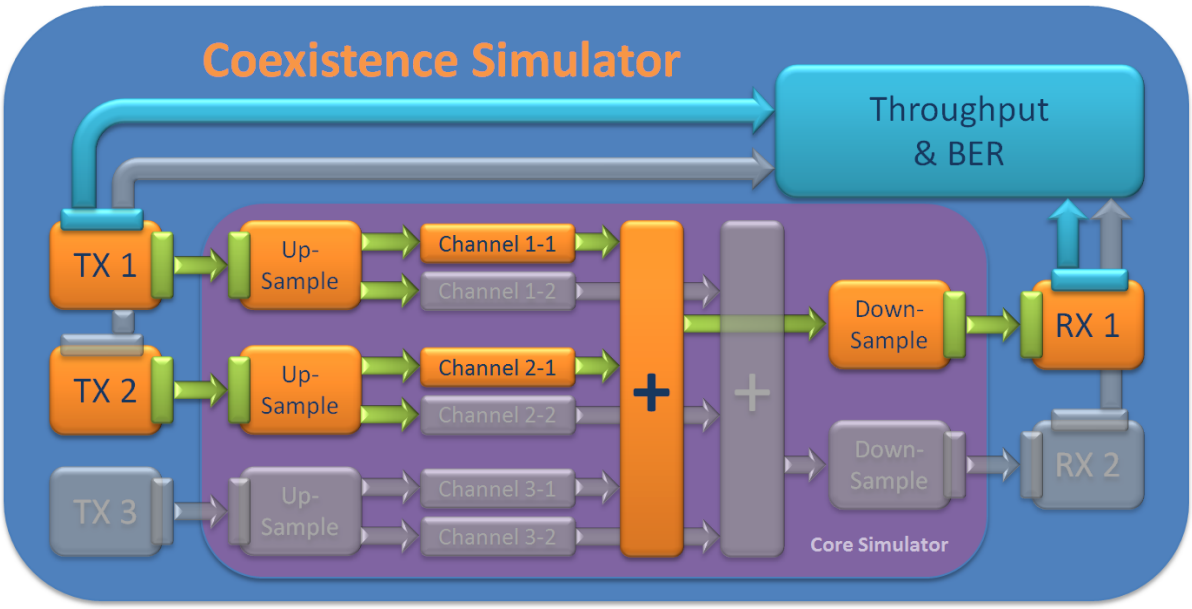


Figure 3. Structure of the Coexistence Simulator Framework

connection that will be investigated) are also possible. First, the delivered baseband signals of the connected senders are up sampled with different factors to get a common sampling rate for all given data streams. The destination sampling frequency depends on the connected wireless technologies and will be calculated from the parameters given through the sender interfaces. The channel models are represented by Rayleigh fading in combination with Additive White Gaussian Noise (AWGN) signal processing blocks. The block parameters are given by the tapped delay line (TDL) models extracted from the ray tracing simulations. The delays of these TDL models are matched to the used common sampling frequency. As shown in Fig. 3, each sender-receiver pair needs its own channel block. This is necessary because each sender and receiver is located on a different position in the simulated room, so there is also a different channel between each unit. For example, an investigated connection has one channel between sender and receiver, but the interfering sender, located elsewhere in the room, experiences different channel effects to the disturbed receiver. The number of channel blocks needed is given by the product of the number of senders and receivers used in a simulation scenario. After passing the channel blocks the signals are combined. This combination has to be performed separately for each connected receiver. After combination the signals are down sampled to the corresponding receiver sampling rate. This sampling rate is given by the interface parameters of the connected receivers.

Finally, the connected receivers try to decode the given signals correctly. To perform this operation the use of channel filtering in the receivers as described in the corresponding standards is obligate. To be able to manage different signal levels at the various sender outputs, separate attenuation blocks have been implemented. This blocks a freely adjustable and may also be used to simulate different transmitter power levels if not possible by the used send/receive chains.

The analysis component matches the data streams from the senders and the corresponding receivers. For that operation the delays of all components involved in signal transmission are needed. For that reason each sender and receiver has a small analysis interface. That interface contains the transmitted or received data and the entire delay of the component. The delay produced by the core simulator is known by the analysis component and does not have to be forwarded separately. In the current version, calculation of throughput in [bits/s] and bit error rate (BER) in [%] is implemented. For each investigated wireless connection (all connections where a receiver is present) a separate analysis component is required.

III. Example Scenario / Results

This section presents a sample scenario that was used for creating the first simulation results. After that the obtained results are presented and discussed.

A. Configuration

The sample scenario consists of a room with one wall in the middle of the room, but this wall does not reach the top and one side wall of the room. The investigated wireless connection is a Bluetooth connection with the receiver on one side of the middle wall and the receiver on the other side. The interfering technology is a Wireless LAN sender on the same side as the Bluetooth receiver. The signal power of the interfering WLAN sender is controlled with a attenuation block on the output of the sender. Table I shows the main parameters used for the simulations. The signal power of the Bluetooth connection was constant over all simulation runs, the WLAN interferer signal power increased each simulation run by 5dB using the attenuation block.

TABLE I. SIMULATOR CONFIGURATION PARAMETERS

	Parameter	*Value*
Bluetooth	Role	Investigated Connection
	Connection type	Data
	Packet type	DM1
	Data transmission	continious
	Number of hop frequencies	79
WLAN	Role	Interferer
	WLAN standard	802.11g
	Channel	6
	Data transmission	50 % of time
	Signal power attenuation	-90 – 0dB in 5 dB steps
Sim. Core	Number of taps Bluetooth	5
	Number of taps WLAN	3
	AWGN Noise Bluetooth	0 dBm

B. Simulation Results

Simulations on the sample scenario using the above parameters have been run. On each simulation cycle the signal power of the interferer has been increased in 5 dB steps. On positions where the output curves have changed dramatically the step size was reduced to 1 dB steps. The signal power of the Bluetooth connection remained the same over all simulation cycles. Fig. 4 shows the BER of transmitted data over the Bluetooth connection. From -90 to -80 dB WLAN interferer attenuation the BER remains zero, because the interfering power is too low to be recognized by the Bluetooth connection. From -80 to -20 dB the BER rises nearly constant to 7%. From here on a dramatically rise in BER occurs. In that most of the jammed Bluetooth packets could not be decoded by the receiver Starting at -5 dB the curve flattens and will not rise very much. This occurs, because the WLAN interferer only sends 50 of the time and the used WLAN channel bandwidth is narrower than the whole frequency hopping

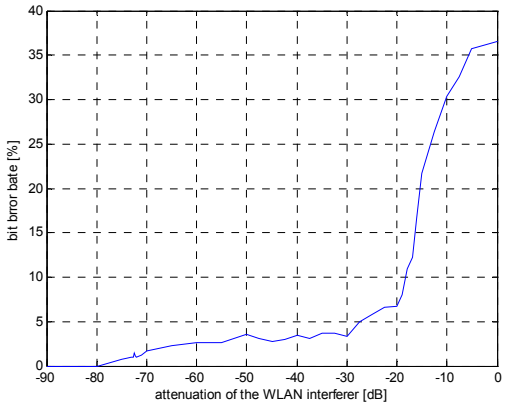


Figure 4. Simulation result: Bluetooth BER on increasing interferer signal power

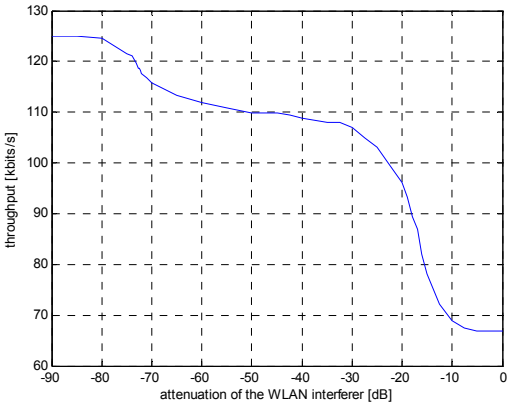


Figure 5. Simulation result: Bluetooth troughput on increasing interferer signal power

bandwidth of the Bluetooth connection, so not all Bluetooth packets are interfered. Fig. 5 shows the throughput of the Bluetooth connection over the signal power attenuation of the WLAN interferer. In general the development of the throughput is analog to the BER curve. Starting at -80 dB to -30 dB the throughput decreases moderately from 125 to 108 kbit/s. After that a big drop in data rate occurs, but then the throughput gets constant at rate of 67 kbit/s.

IV. CONLUSION

We proposed a first version of a simulation framework for investigating coexistence issues on wireless connections in close spatial proximity. We explained why such a framework is needed and presented the structure of this framework and the motivations why this structure was selected. Finally we presented the first simulation results of a defined sample scenario. Future work is the implementation of additional radio technologies and the refinement of the ray tracing and analysis components.

ACKNOWLEDGMENT

This project was supported by the program *Regionale Wettbewerbsfähigkeit OÖ 2010-2013*, which is financed by the European Regional Development Fund and the Government of Upper Austria.

REFERENCES

[1] R. Hoppe, "Effiziente Modellierung der Wellenausbreitung für die Funknetzplanung in urbanen Szenarien und innerhalb von Gebäuden," Shaker, 2002, ISBN 3-832-20749-X.

[2] O. Stäbler, "Modellierung und Simulation von Funkkanälen für Mehrantennensysteme (MIMO) mit einem strahlenoptischen 3D Wellenausbreitungsmodell," diploma thesis, Erlangen, 2010.

[3] "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications", IEEE Standard 802.11-2007, IEEE, 2006

[4] "Wireless medium access control (MAC) and physical layer (PHY) specifications for wireless personal area networks (WPANs)", IEEE Standard 802.15.1-2005 (Bluetooth), IEEE, 2005

[5] "Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (WPANs)", IEEE Standard 802.15.4-2006, IEEE, 2006

On the Performance of DoA Estimation Algorithms in Cognitive Radio Networks: -A new Approach in Spectrum Sensing

Tanuja Satish Dhope (Shendkar)*, Dina Simunic**
* Sinhgad Institute of Technology, Lonavala (India), *Faculty of Electrical Engineering and Computing University of Zagreb, Croatia
** IEEE Senior Member, Faculty of Electrical Engineering and Computing University of Zagreb,Croatia
*tanuja__dhope@yahoo.com,**Dina.Simunic@fer.hr

Abstract – Since the spectrum is a scarce natural resource, sensing a spectrum is an important part of Green Engineering. The cognitive radio has been proposed to meet the ever increasing demand of the radio spectrum by allocating the spectrum dynamically to allow unlicensed access on non-interfering basis. But, the conventional spectrum sensing methods are seeking for spectrum opportunities in either of three dimensions viz. time, frequency and space. However, other dimensions like 'Code' and 'Angle' need to be explored further for spectrum opportunity. This paper investigates the performance of Direction of Arrival (DoA) estimation algorithms like MUSIC, Capon and ESPRIT in cognitive radio networks. The adaptive thresholding has been proposed to improve the performance of DOA algorithms. The results show that the performance of the algorithms improves with increasing number of array elements, increasing number of snapshots and increasing signal to-noise ratio. This new approach of DoA estimation of licensed user, increases channel capacity and improves frequency reuse capability by multiplexing licensed and unlicensed users into the same channel at the same time in the same geographical area by forming the beam of unlicensed user in the direction other than the licensed users' DoA direction.

Key words: Spectrum Opportunity, Cognitive radio, Direction of Arrival, Adaptive thresholding.

I. INTRODUCTION

The complexity of the wireless systems requires a careful design, especially related to bandwidth and energy efficiency. Energy efficiency and Bandwidth efficiency are main aspects of 'Green Engineering' as the energy efficiency is getting more and more on importance, due to increasing penetration of various wireless systems in different battery-oriented applications, as well as due to the more conscious global view on the need for "greening the Earth" and also the bandwidth efficiency is very important parameter, because it relates to frequency spectrum, which is naturally limited resource [1]. As a result, innovative techniques that can offer new ways of exploiting the available spectrum are needed. Recently, there have been growing interests in cognitive radio(CR)[2], where secondary opportunistic radio exploits opportunistically spectrum left-overs, or so-called "White Spaces", by means of knowledge of the environment and cognition capability, and adapts their radio parameters accordingly [3]-[6]. In CR, sensing a spectrum is crucial task and it is important aspects of Green Engineering. In [6]-[10] spectrum opportunities have been studied well for the upcoming IEEE 802.11af standard (known as "super-Wi-Fi" or "Power over Ethernet") which is based on cognitive radio (CR) utilizing the bandwidth within TV broadcast stations. CR makes opportunistic use of the spectrum by allowing unlicensed or secondary user (SU) to reuse the spectrum whenever the licensed or primary user (PU) is inactive.

The conventional definition of the spectrum opportunity, which is often defined as "*a band of frequencies that are not being used by the PU of that band at a particular time in a particular geographic area*", only exploits time, frequency, and space spectrum dimensions [11]. But, 'Angle' and 'Code' dimensions are needed to be explored further for spectrum opportunity .The angle dimensions are not explored as it is assumed that the PUs and/or the SUs transmit in all the directions. But with the recent advances in multi-antenna technologies through beamforming, multiple users (PU and/or SU) can be multiplexed into the same channel at the same time in the same geographical area. Here angle dimension is different than geographical space dimension. By virtue of angle dimension, a PU and a SU can be accommodated in the same geographical area and share the same channel. Geographical space dimension refers to physical separation of radios in distance. This new dimension also creates new opportunities for spectral estimation where not only the frequency spectrum but also the angle of arrivals (AoAs) / Direction of Arrival (DoAs) needs to be estimated.

The goal of this paper is to present an analysis DoA estimation algorithms in CR context for dynamic spectrum access. This paper focuses on analyzing Capon, MUSIC and ESPRIT algorithms under AWGN channel and Rayleigh fading channel. The adaptive thresholding has been proposed to improve the performance of DOA algorithms.The organisation of paper is as follows: section II describes various DoA estimatiom algorithms.Section III presents the simulation results based on different DoA estimation algorithms along with adaptive thersholding technique which is followed by conclusions in section IV.

II. ANALYSIS OF DOA ESTIMATION ALGORITHMS

The aim of DoA estimation algorithm is to use data received by the array to estimate the DoA of PU signal. As shown in Fig. 2, the received signals are first processed through a pre-filter of bandwidth B, matched to the bandwidth of PU signal. In parameter estimation block, each antenna system performs DoA estimation to find PU signal by calculating time delays between antenna elements [12]. Once the DoA of PU signals are estimated, then using adaptive algorithm which uses cost function for calculating the optimum filter weights that maximize beam for SU toward the intended direction ;a direction other than PU signal direction while nulling beam pattern in the direction of PU signal.This approach improves frequency reuse capability and increases the channel capacity by multiplexing multiple PUs and SUs into the same channel at the same time in the same geographical area by forming beam pattern of SU in a direction that does not coincide with PU direction.

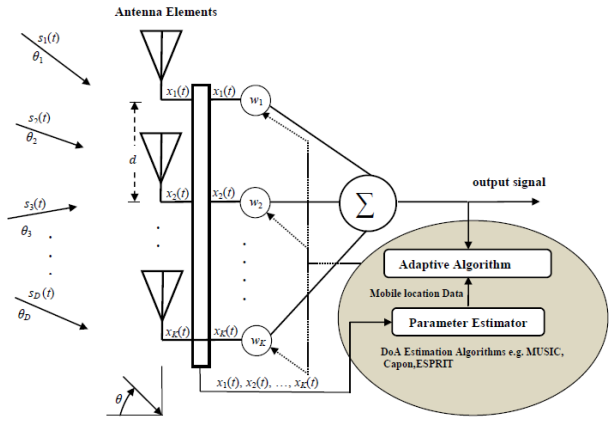


Figure.2: Beamforming at SU side with D arriving PU signals on K antenna elements (θ_1 *to* θ_D are DoAs of transmitted signals s_1 to s_J; x_1 *to* x_K are antenna - specific received signals; d =0.5λ

A. System Model

An Uniform Linear Array (ULA) is considered, with D number of PU signals of frequncy f_0 arriving at K number of array elements of SU which are equally spaced at distance d between the elements.The channel noise for all channels is mutually non-coherent narrowband at f_0.

The steering vector of dimensions Kx1 corresponding to DoA at some angle θ is given by column vector :

$$\mathbf{p}(\theta) = [\ e^{(-j(m-1)2\pi d \sin(\theta)/\lambda)}]^{T} \quad m = 1, 2, \ldots K \qquad (1)$$

Where λ is wavelength, $f_0 = c / \lambda$ and c is the velocity of light.The columnwise combination of all D steering vectors is called array manifold matrix $\mathbf{P}$ of dimensions KxD given by $\mathbf{P}(\theta) = [\mathbf{p}(\theta_1) : \mathbf{p}(\theta_2) : \ldots\ldots\ldots : \mathbf{p}(\theta_D)]$
The spatial correlation (covariance) matrix for the N number of snapshots is given by:

$$\mathbf{S_x} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{x}(t)\mathbf{x}(t)^{H} \qquad (2)$$

where H denotes the Hermitian operator and $\mathbf{x}$ denotes denotes a vector of dimensions Kx1 consisting of received signals x_K. Substitution of (1) into (2) results in

$$\mathbf{S_x} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{P}(\theta)\mathbf{s}(t)\mathbf{s}(t)^{H}\ \mathbf{P}(\theta)^{H} + \mathbf{n}(t)\mathbf{n}(t)^{H} \qquad (3)$$

$$\mathbf{S_x} = \mathbf{P}(\theta)\ \mathbf{S_s}\ \mathbf{P}(\theta)^{H} + \sigma_w^2\ \mathbf{I} \qquad (4)$$

Where σ_w^2 is noise variance, $\mathbf{I}$ is an identity matrix of size KxK and $\boldsymbol{S_s}$ is received signal power matrix.

B. Beamforming Techniques

The principle behind beamforming technique is to "steer" the array in one direction at a time and measure output power. The steering locations that give maximum power yield AoA estimates. A number of sources will correspond to a number of peaks[12]. The array response is steered by forming a linear combination of the sensor outputs[12][13].
The array output is :

$$y = \sum_{i=1}^{K} \mathbf{w}_i^{*}\ x_i\ = \mathbf{w}^{H}\ \mathbf{x}(t) \qquad (5)$$

where $\mathbf{w} = [\ w_1,\ w_1, \ldots\ldots, w_K\ ]^{T}$ is a weighting vector, which determines the radiation pattern. Given samples $y(1), y(2), \ldots\ldots, y(N)$ and hence output power is:

$$\mathbf{Po}(\mathbf{w}) = \frac{1}{N}\sum_{t=1}^{N} |y(t)|^2\ = \frac{1}{N}\sum_{t=1}^{N} \mathbf{w}^{H}\ \mathbf{x}(t)\mathbf{x}(t)^{H}\mathbf{w}$$
$$= \mathbf{w}^{H}\ \mathbf{S_x}\ \mathbf{w} \qquad (6)$$

Capon's method is based on Beamforming technique.

C. Capon's method

The key is to minimize power contributed by noise and any signals coming from other direction than desired (PU signal direction) [13].

$$\min_{w} (\mathbf{w}^{H}\ \mathbf{S_x}\ \mathbf{w}) \quad \text{Subject to } |\mathbf{w}^{H}\mathbf{v}(\theta)| = 1 \qquad (7)$$

The Capon's weight vector can be found to be:

$$\mathbf{w}_{capon}\ = \frac{\mathbf{S_x}^{-1}\ \mathbf{p}(\theta)}{\mathbf{p}(\theta)^{H}\ \mathbf{S_x}^{-1}\ \mathbf{p}(\theta)} \qquad (8)$$

Thus Capon's output spectrum is:

$$P_{capon}\ = \frac{1}{\mathbf{p}(\theta)^{H}\ \mathbf{S_x}^{-1}\ \mathbf{p}(\theta)} \qquad (9)$$

D. Subspace Based Methods

Subspace based method deals with the decomposition of the observed covariance matrix into two orthogonal spaces: signal subspace and noise subspace. The DoA estimation is calculated from any one of the subspaces. The subspace based DoA estimation algorithm MUSIC and ESPRIT provide high resolution, they are more accurate and not limited to physical size of array aperture [12] - [15].

E. MUSIC algorithm

MUSIC stands for **MU**iltiple **SI**gnal **C**lassification, one of the high resolution subspace DoA algorithms, which gives an estimate of a number of arrived signals, hence their direction/angle of arrival [12]-[15]. Estimation of DoA is performed from one of subspaces either signal or noise, assuming that noise in each channel is highly uncorrelated. This makes the covariance matrix diagonal. Writing the spatial covariance matrix in terms of eigenvalues and eigenvectors [11][13][14] gives

$$\mathbf{S_x} = \sum_{i=1}^{M} M_i \ \boldsymbol{\varphi}_i \ \boldsymbol{\varphi}_i^{\mathrm{H}} \tag{10}$$

$$\mathbf{S_x} = \boldsymbol{\varphi}_i \alpha \ \boldsymbol{\varphi}_i^{\mathrm{H}} \tag{11}$$

$$\alpha = \mathrm{diag}\left[M_1, M_2, \ldots \ldots, M_K\right] \tag{12}$$

The noise subspace eigenvalues and eigenvectors are

$$M_i \ , \ i = D+1, D+2, \ldots, K \tag{13}$$

$$\boldsymbol{\varphi}_i \ , \ i = D+1, D+2, \ldots, K \tag{14}$$

The noise subspaces can be written in the form of K x $(K-D)$ matrix:

$$\boldsymbol{\vartheta}_N = [\boldsymbol{\varphi}_{D+1}, \boldsymbol{\varphi}_{D+2}, \ldots \ldots, \boldsymbol{\varphi}_K] \tag{15}$$

Equation (15) indicates that the desired value DoA of θ_1, θ_2,....., θ_D can be found out by finding a set of vectors that span $\boldsymbol{\vartheta}_N$ and projecting $\mathbf{p}(\theta)$ onto $\boldsymbol{\vartheta}_N$ for all values of θ and evaluating the D values of θ, where the projection is zero:

$$\left\|\mathbf{v}_i^{\mathrm{H}} \boldsymbol{\vartheta}_N\right\|^2 = 0 \qquad i = 1,2, \ldots, D \tag{16}$$

Thus, MUSIC Pseudospectrum is given as:

$$P_{\mathrm{music}}(\theta) = \frac{1}{\mathrm{abs}\left[\mathbf{p}(\theta)^{\mathrm{H}} \boldsymbol{\vartheta}_N \boldsymbol{\vartheta}_N^{\mathrm{H}} \mathbf{p}(\theta)\right]} \tag{17}$$

F. ESPRIT

Its acronym stands for **Estimation of Signal Paramter via Rotational Invariance Technique.** This algorithm is more robust with respect to array imperfections than MUSIC [12],[16]. Computation complexity and storage requirements are lower than MUSIC as it does not involve extensive search throughout all possible steering vectors. But, it explores the rotational invariance property in the signal subspace created by two subarrays derived from original array with a translation invariance structure. It is based on the array elements placed in identical displacement forming matched pairs, with K array elements, resulting in m=K/2 array pairs called "doublets".

Computation of signal subspace for the two subarrays, P_1 and P_2, results in two vectors $V_1$and V_2, such that Range [S] = Range[B]. Also, there should exist a non-singular matrix T of DxD such that $\boldsymbol{V_s} = \bar{B}\,T$, where V_s can be decomposed into $V_1$and V_2 :

$$\boldsymbol{V}_1 = BT, \ \boldsymbol{V}_2 = B\boldsymbol{\varphi}\,T \tag{18}$$

$$\boldsymbol{\varphi} = diag\left[e^{jkdsin\,(\theta_1)}, e^{jkdsin\,(\theta_1)}, ., e^{jkdsin\,(\theta_D)}\right] \tag{19}$$

After various mathematical transformation as shown in [12] ,DoA can be calculated as follows

$$\boldsymbol{a}_i = e^{jkdsin\,(\theta_i)} \ \ i = 1,2, \ldots, D \tag{20}$$

DoA can be calculated by:

$$\boldsymbol{\theta}_i = sin^{-1}\left[\frac{\arg(\boldsymbol{a_i})}{kd}\right] \tag{21}$$

III. SIMULATION RESULTS

In order to evalute the performance of DoA estimation algorithms for PUs in the selected frequncy band, and to have reliable comparison between algorithms, 150 trials were run for each case and their results were averaged before the comparison. Also, Mean Squared Error (MSE) (in degrees) of DoA estimation is used for presenting accuracy. Higher the error – the higher the unreliability of the algorithm for given conditions. The performance of DoA estimations for MUSIC, Capon and ESPRIT were simulated using MATLAB.

The simulations were run for angular separation W_{AS} = {10^o, 20^o, 30^o} .Furthermore, accuracy in the case of DoA at SNR=-8dB is checked and proposal for adaptive thresholding for MUSIC and Capon is done. For analysing the performance of these algorithms, regarding impact of number of array elements, number of snapshots and SNR, simulation parameters were set as follows:

1. **For spectrum plotting :**
   - Impact of number of array elements: at SNR of -4dB, 500 snapshots for environment W_{AS} was considered.
   - Impact of number of snapshots: with array of 14 elements at SNR of -4dB for environment W_{AS} was considered.
   - Impact of SNR: with array of 14 elements with 500 snapshots for environment W_{AS} was considered.

2. **MSE for W_{AS} environment**
   - Impact of number of array elements: at SNR of -8dB, 500 snapshots

3. **Adaptive thresholding at SNR=-8dB**
   - Impact of number of array elements: at 500 snapshots for environment W_E was considered.

4. **Rayleigh Fading channel:**

MSE for MUSIC as a function of varying number array elements with SNR=-8dB, Snapshots =500**,** W_{AS} environment with Doppler frequency shift =0Hz and 30Hz

Fig. 2 to 7 clearly indicate that ,as the number of array elements varied from 5 to 12 , number of snapshots varied from 100 to1500 and SNR varied from -18 dB to 10dB,the resolution capability of MUSIC and Capon increases giving sharper peaks in the spectrum. In this situation the resolution capability of MUSIC is better than Capon even though the SNR is as low as -14dB.

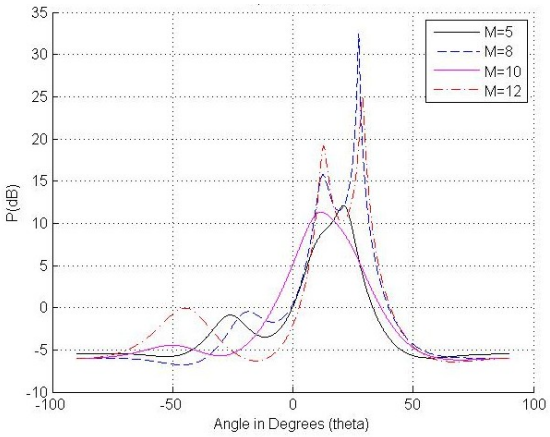


Figure 2: MUSIC for Varying number of array elements,SNR=-4dB,snapshots=500 , $W_{AS} = [10^0,20^0,30^0]$

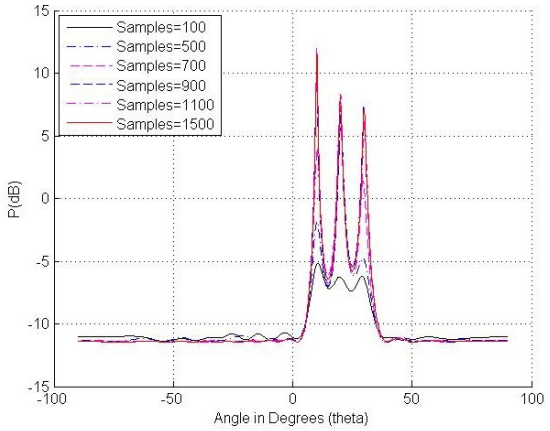


Figure 3: MUSIC for Varying number of snapshots with SNR=-4dB, array elements=14, $W_{AS} = [10^0,20^0,30^0]$

Table 1, clearly indicates that as number of array elements increased from 5 to 20, the MSE in DoA estimation for $W_{AS}=[10^0,20^0,30^0]$ also decreases for three algorithms. The MUSIC provides less MSE compared to Capon and ESPRIT from array elements of 12 onwards.

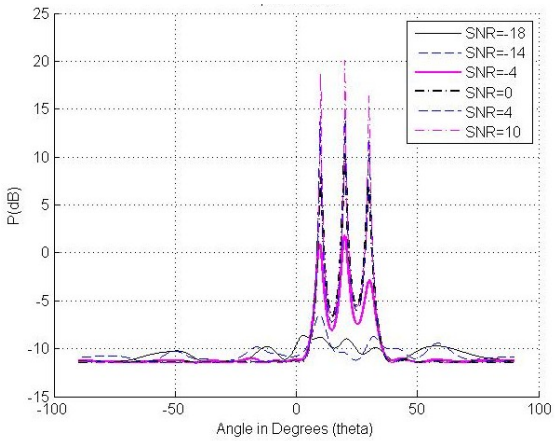


Figure 4: MUSIC for Varying number of SNR with array elements =14, snapshots=500, $W_{AS} = [10^0,20^0,30^0]$

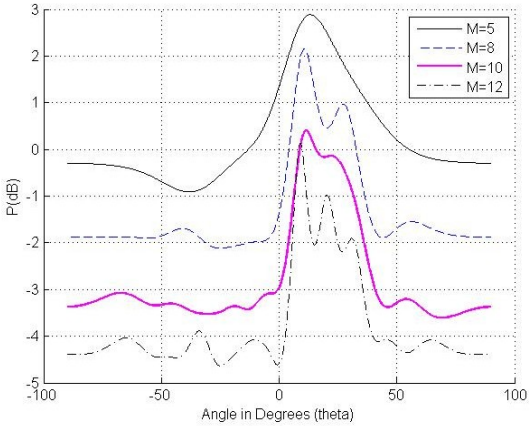


Figure 5: Capon for Varying number of array elements with SNR=-4dB, snapshots =500, $W_{AS} = [10^0,20^0,30^0]$

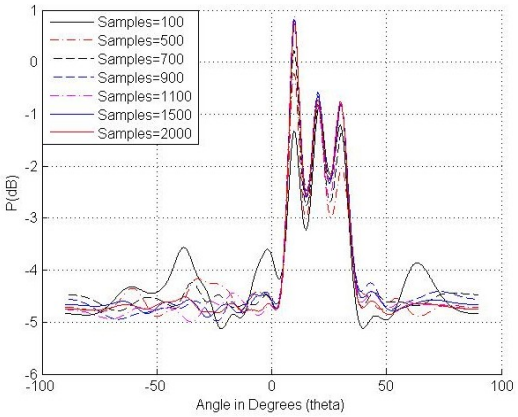


Figure 6: Capon for Varying number of snapshots with array elements=14, SNR=-4dB, $W_{AS} = [10^0,20^0,30^0]$

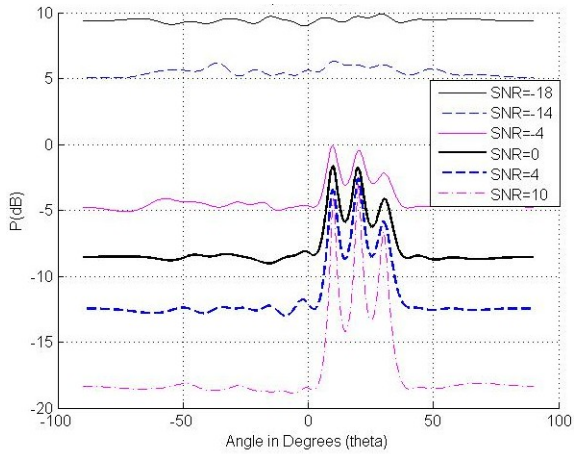


Figure 7: Capon for varying number of SNR with array elements =14, snapshots =500, W_{AS} = [10^0,20^0,30^0]

Table 1: MSE for MUSIC, Capon and ESPRIT as a function of array elements with SNR=-4dB, Snapshots=500, W_{AS} =[10^0,20^0,30^0].
{Note: the first column in each DoA algorithm is MSE for 10^0, second column is MSE for 20^0 and last column is MSE for 30^0 }

MSE for DoA (in deg.) for 150 runs for[10,20,30]									
Array Elements	MUSIC			Capon			ESPRIT		
5	0.0196	0.087025	1	0.2116	1	1	1	0.000141	1
6	0.25	1	1	0.0289	1	1	1	0.00506	1
7	0.0016	1	0.005878	0.0025	1	0.0004	0.137033	0.020733	1
8	1	0.0144	1	0.0049	1	0.006944	0.103465	0.044836	1
9	1	0.0576	1	0.0441	1	0.0025	1	0.00012	1
10	1.00E-04	2.50E-05	1	0.0009	1	0.0169	0.009633	0.001888	0.017038
11	0.0004	0.0009	1	0.0064	0.0121	0.0064	0.009984	0.002476	1
12	0.0016	0.000625	0.000711	0.0036	1	1	0.002018	0.002522	1
13	0.0064	0.000225	0.006944	0.0064	1	1	0.001481	0.00071	0.026484
14	0.0004	0.000625	0	0.0036	0.000225	1	1	0.000231	1
15	0.0004	0.000625	0.0001	0.0009	1.00E-04	1	1	0.000972	1
16	0.0009	1.00E-04	0.0016	1.00E-04	1.00E-04	1	0.218117	0.0004	1
17	0.0009	2.50E-05	1	0.0009	1.00E-04	1	0.019508	0.000412	2.66E-05
18	0.0016	0	0.0001	1.00E-04	0.0004	1	0.00632	2.55E-07	1
19	1.00E-04	0.000225	4.44E-05	0	2.50E-05	1	0.083961	0.001213	1
20	0	2.50E-05	4.44E-05	1.00E-04	0	1.11E-05	0.022765	0.000561	0.004102

In low SNR conditions, noise is always comparable with signal; the detection of desired DoA of the signal is tough task. The Fig.8 to Fig.11 show the angle detected as a function of array elements for MUSIC and Capon without and with adaptive thresholding. The adaptive thresholding sets the limit on the number of peaks detected by the algorithms so that the unwanted/ undesired peaks which corresponds to undesired DoA of the signals can be filtered out. In this case the threshold is set from 0.5 to 2 dB.

MUSIC and Capon without adaptive thresholding for snapshots of 500 at SNR of -8dB for varying number of array elements 5 to 20, shows more number of estimated DoA of PU signal. MUSIC fails to detect 30^0 angle where the angular separation is only 10^0 for any value of array elements 5 to 20.But with adaptive thresholding from array elements of 12 onwards; MUSIC clearly identifies DoA of 30^0 except at array elements of 17. Capon gives the accurate detection of all three angles for array elements of 20.

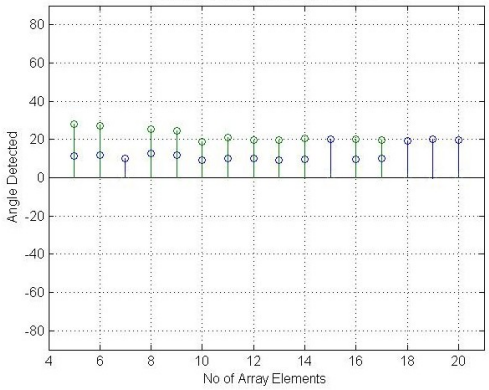


Figure 8: MUSIC for varying number of array elements with SNR=-8 dB, snapshots =500, W_{AS} = [10^0,20^0,30^0] without adaptive thresholding

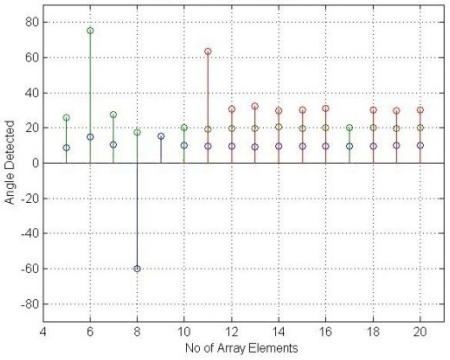


Figure 9: MUSIC for varying number of array elements with SNR=-8dB, snapshots =500. W_{AS} = [10^0,20^0,30^0]with adaptive thresholding

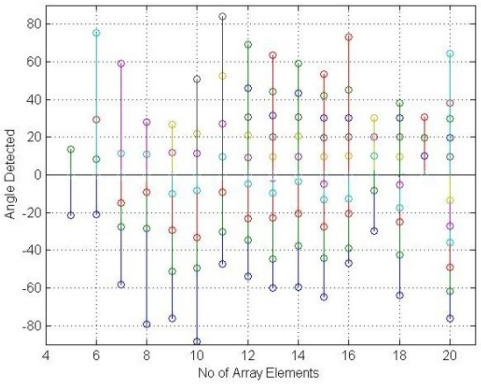


Figure 10: Capon for varying number of array elements with SNR=-8 dB, snapshots =500, W_{AS} = [10^0,20^0,30^0] without adaptive thresholding

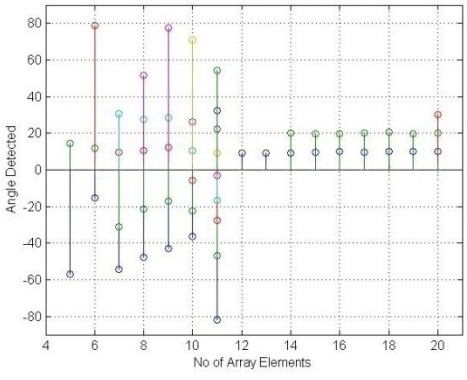


Figure 11: Capon for varying number of array elements with SNR=-8 dB, snapshots=500, W_{AS} = [10^0,20^0,30^0]with adaptive thresholding

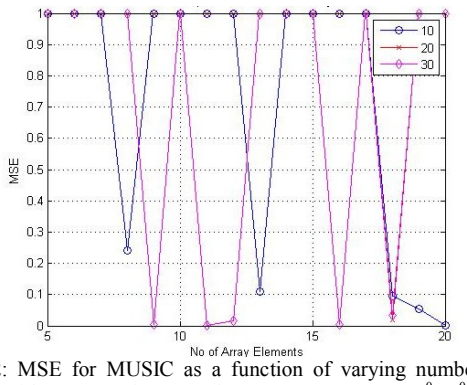


Figure 12: MSE for MUSIC as a function of varying number array elements with SNR=-8dB, snapshots =500, W_{AS} = [10^0,20^0,30^0] in Rayleigh Channel with Doppler shift =0Hz

In order to observe the effect of Rayleigh fading on the estimation on DoA for MUSIC algorithm at the SU side, flat fading Rayleigh channel is considered in Fig.12 and a time varying Rayleigh channel with Doppler frequency shift of 30Hz is considered in Fig.13 . The time variant channel is constructed using Jakes' model as shown in Fig.13. By comparing Fig.12 with Fig.13, it clearly reveals that there is degradation in estimation of DoA by MUSIC in time variant Rayleigh channel giving more MSE

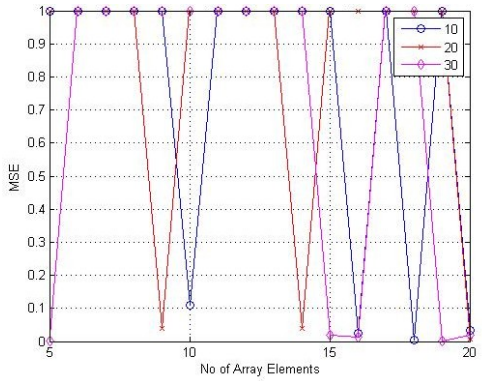


Figure 13: MSE for MUSIC as a function of varying number array elements with SNR=-8dB, snapshots =500, W_{AS} = [10^0,20^0,30^0] in Rayleigh Channel with Doppler shift =30Hz

IV. CONCLUSIONS

In this paper, the Direction of Arrival of primary user signal has been checked for any spectrum opportunity in 'Angle' dimensions in cognitive radio context. By sensing the Direction of Arrival of primary user signal at the secondary user side, the secondary user's beam can be directed to the direction that is not coinciding with the direction of primary user. The simulation results clearly indicates that as number of snapshots ,number of array elements and SNR increases the performance of MUSIC, Capon and ESPRIT improves. MUSIC outperforms the Capon and ESPRIT algorithms in all respects. The proposed adaptive thresholding technique provides better detection of Direction of Arrival in low SNR conditions. The MUSIC algorithm outperforms Capon in low SNR conditions when compared the simulation results using adaptive thresholding for varying number of array elements. The degradation in direction of arrival for primary user signal has been observed in time varying fading channel.This new approach of DoA estimation of licensed user, increases channel capacity and improves frequency reuse capability by multiplexing licensed and unlicensed users into the same channel at the same time in the same geographical area by forming the beam of unlicensed user in the direction other than the licensed users' DoA direction.

REFERENCES

[1] FCC, "Spectrum policy task force report," in Proceedings of the Federal Communications Commission (FCC '02), Washington, DC, USA, November 2002.

[2] J. Mitola and G. Q. Maguire, "Cognitive radio: Making software radiosmore personal," *IEEE Pers. Commun.*, vol. 6, no. 4, pp. 13–18, August 1999.

[3] T. Dhope, D. Simunic and R.Prasad, "TVWS opportunities and regulation:-empowering rural India,"14th International Symposium on WPMC '11, Brest, France, pp.201-205, October 2011.

[4] T. Dhope, D. Simunic and R.Prasad, "TVWS opportunities and regulatory aspects in India," 14th International Symposium on WPMC '11, Brest, France, pp. 566-570, October 2011.

[5] T. Dhope, D. Simunic and R.Prasad, "Hybrid Detection Method for Cognitive Radio,"13th International conferenace on SoftCOM SoftCOM2011,Split-Hvar-Dubrovnik,SS1-78741-1609,September 2011.

[6] Spectrum Sensing Algorithms for IEEE 802.11 af Standard in Cognitive Radio Network*, Studies in informatics and Control, vol.21 , Issue 1 , pp.93-100 , Mar. 2012

[7] T. Dhope, D. Simunic and R.Prasad, "TVWS Radio Spectrum Utilisation: Use Case of India-Looking Forward ", Journal of Green Engineering, Vol. 3, Issue 1 , pp.91-112.Oct.2012. ISSN: 1904-4720

[8] T. Dhope, D. Simunic ,"Spectrum Sensing Algorithm for Cognitive Radio Networks for Dynamic Spectrum Access for IEEE 802.11 af standard",IJRRWSN,Vol.2.,No.1,Mar. 2012(in publication)

[9] T. Dhope, D. Simunic, "Performance Analysis of Covariance Based Detection in Cognitive Radio,"35th Jubilee International Convention MIPRO 2012 ,Opatija, Croatia, May 2012 (accepted)

[10] T. Dhope, D. Simunic, "Hybrid Detection Method for Spectrum Sensing in Cognitive Radio," 35th Jubilee International Convention MIPRO 2012. Opatija, Croatia, May 2012(accepted)

[11] Tevfik Y¨ucek and H¨useyin Arslan, "A Survey of Spectrum Sensing Algorithms for Cognitive Radio Applications,"IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 11, NO. 1, FIRST QUARTER 2009,pp-116-130,2009.

[12] T. Dhope, D. Simunic and M. Djurek, "Application of DOA Estimation Alg*orithms in Smart Ant*enna Systems," Studies in informatics and Control, vol.19, Issue 4, pp.445-452, December 2010.

[13] R.J. Weber and Y. Huang,"Analysis for Capon and MUSIC DOA Estimation Algorithms," Antennas and Propagation Society International Symposium, 2009. APSURSI '09,pp. 1–4, June 2009.

[14] M. Jalali, M. Moghaddasi and A. Habibzadeh, "Comparing accuracy for ML, MUSIC, ROOT-MUSIC and spatially smoothed algorithms for 2 users," Microwave Symposium (MMS), 2009 Mediterrannean IEEE conference, pp. 1–5, Nov.2009.

[15] F.A Leon De and S.J.J Marciano, "Application of MUSIC, ESPRIT and SAGE Algorithms for Narrowband Signal Detection and Localization, "TENCON, IEEE Region 10 Conference, pp. 1-4, Nov. 2006.

[16] Z.I. Khan, R.A. Awang A.A Sulaiman, M.H. Jusoh, N.H Baba ,M.M.D. Kamal, and N.I. Khan," Performance analysis for Estimation of signal Parameters via Rotational Invariance Technique (ESPRIT) in estimating Direction of Arrival for linear array antenna," IEEE International Conference on RF and Microwave Conference, 2008, pp. 530–533, Dec. 2008.

Design and Implementation of SDR Based QPSK Modulator on FPGA

Tarik Kazaz, Merima Kulin, Mesud Hadzialic
Faculty of Electrical Engineering
University of Sarajevo
Sarajevo, Bosnia and Herzegovina
{ tarik.kazaz, mk15172, mhadzialic }@etf.unsa.ba

***Abstract*—Software defined radio (SDR) technology enables implementation of wireless devices that support multiple air-interfaces and modulation formats, which is very important if consider proliferation of wireless standards. To enable such functionality SDR is using reconfigurable hardware platform such as Field Programmable Gate Array (FPGA). In this paper, we present design procedure and implementation result of SDR based QPSK modulator on Altera Cyclone IV FPGA. For design and implementation of QPSK modulator we used Altera DSP Builder Tool combined with Matlab/Simulink, Modelsim and Quartus II design tools. As reconfigurable hardware platform we used Altera DE2-115 development and education board with AD/DA daughter card. Software and Hardware-in-the-loop (HIL) simulation was conducted before hardware implementation and verification of designed system. This method of design makes implementation of SDR based modulators simpler ad faster.**

***Index Terms*—SDR, FPGA, QPSK, DSP Builder, NCO, RRC**

I. Introduction

The term software radio, which became later the term software defined radio, was first introduced by prof. Joseph Mitola [1], as an approach for implementing a re-programmable and re-configurable radio transceiver. Benefits and advantages of SDR transceiver systems became apparent after the development of a large number of standards for wireless communication networks. The terminal equipment of operators was required to support specifications of more standards simultaneously. This had required the need to implement a transceiver system that is able to support multiple encryptions, modulation and signal processing techniques. Transceiver systems implemented in accordance with the concept of SDRs, represent a solution to these problems and challenges. One possible approach for the realization of SDR transceiver systems is the use of FPGA hardware components. FPGA platform is attractive because of its good performance, low power consumption and configurability. Digital signal processing based on the use of FPGA chips has found application in many areas such as: Mobile Systems (3G and 4G), VoIP, multimedia, radar and satellite systems.

On the other hand, Phase-Shift Keying (PSK) modulation is widely used in existing wireless technologies. Current and important applications of Quadrature PSK (QPSK) modulation are in standards: LTE & LTE-ADVANCE, IEEE 802.11b-1999, IEEE 802.11g-2003 and IEEE 802.15.4. In QPSK, two successive bits in the data sequence are grouped together into a symbol. At the input, the QPSK modulator performs a serial to parallel conversion of the input bits, and then the conversion from unipolar to bipolar signal. After these operations, it is necessary to perform signal filtering to avoid signal distortion called intersymbol interference (ISI) [2]. After filtering, the modulating signal is ready to modulate the carrier. At the end the modulated signals, generated in two separate branches of the modulator (I and Q branch), are summarized and translated into an analog signal by a digital-to-analog converter. The signal at the output of QPSK modulator is mathematically represented as defined by the following equations:

$$s_{QPSK}(t) = d_I(t)\cos(2\pi f_0 t) - d_Q(t)\sin(2\pi f_0 t) \quad (1)$$

$$d_I(t) = \sqrt{\frac{2E}{T}}\cos\left[(2i-1)\frac{\pi}{4}\right] \quad (2)$$

$$d_Q(t) = \sqrt{\frac{2E}{T}}\sin\left[(2i-1)\frac{\pi}{4}\right] \quad (3)$$

Where E is the energy per symbol, $i = 1, 2, 3, 4$, $0 \leq t \leq T$.

The differences between digital implementation of QPSK modulator on FPGA chip, and its conventional implementation in analog electronics are significant. The conventional approach would require the decomposition of the presented system into subsystems, and the need for more hardware components. For example, the serial to parallel conversion of a series of input bits would require the use of a combination of D and T flip-flops [3]. Carrier generation would require the use of voltage controlled oscillator (VCO) and a phase shifter of 90 degrees. Multiplying the carrier and the modulating signal would require two mixers. In addition, for signal shaping it is necessary to use at least two filters. With the use of FPGA integrated circuits, the previously described QPSK modulator could be implemented almost entirely on a single chip along with an indispensable addition of an analog-to-digital converter. There is an obvious advantage and convenience of system implementation on FPGA chips. To realize the QPSK modulator it is necessary to implement the blocks that will perform functions equivalent to the functions that are performed by circuits with the realization in conventional approach. The functions of voltage controlled oscillator in

the digital domain are performed by numerically controlled oscillator (NCO) [4]. NCO is used for synthesizing a discrete-time, discrete-valued representation of a sinusoidal waveform. Signal shaping filters are finite impulse response (FIR) filters whose transfer function has the shape of the raised cosine.

The goal of this paper is to use the discussed advantages and presents the design and implementation procedure of SDR based QPSK modulator using FPGA hardware platform. At the end, obtained results are discussed.

The remainder of the paper is organized as follows. Advantages of FPGA are mentioned in Section 2. Section 3 introduces the model-based design method used for creating FPGA prototype. In section 4 hardware implementation and simulation model analysis are considered. Obtained results are discussed in section 5. Finally, conclusion and discussion on future work are in section 6.

II. FPGA IN THE WORLD OF SDR

Implementation of SDR implies more specific design and analysis procedures than the implementation of conventional transceiver systems. Selection of hardware components for transceiver implementation, that follows the SDR concept, is the first and crucial step necessary for its implementation. All selected hardware components together form a hardware platform for SDR creation. During the process of forming a hardware platform it is necessary to achieve a compromise between desired flexibility, modularity, scalability and performance of the SDR system [5][6]. Flexibility is the ability of a system to handle variety of air-interfaces and protocols, even if they have yet to be defined [5]. Modularity of a system allows easy replacement or upgrading of subsystems to take advantage of new technologies [5]. Scalability is related to modularity, and it allows the system to be enhanced to improve capability such as increasing number of channels that a base station could handle [5]. Overall system performance is closely tied to other three characteristics. Performance may be quantified by power consumption, relative cost, and computational capability metrics [5]. Common hardware components used for forming digital part of hardware platform are: ASICs, FPGAs, and DSPs. Each of these components provides a certain degree of re-programmability, which is a fundamental precondition for the realization of SDR. DSPs are based on microprocessor architecture and they can be programmed using high-level programming languages such as C. This contributes to the high flexibility when using DSPs. However, the advantage of DSP flexibility is paid by their effectiveness. For example, in the case of implementing an element of system that requires performing large number of operations at the same time. When realization of this element is done on DSP, operations will almost certainly be executed sequentially. One way to circumvent this problem is to use more than one DSP chip. Other solutions include the use of hardware components such as ASICs or FPGAs. ASICs enable optimized implementation of the desired system in the fixed silicon with respect to speed and power consumption. On the other hand, FPGAs provide a higher level of hardware reconfigurability in comparison to ASICs, but less reconfigurability than DSPs [5][7]. FPGAs provide a higher degree of parallelism with regard to DSPs. It is possible to form a larger number of executable objects that will be simultaneously executed. Also during the implementation on FPGAs there are no limitations with respect to the precision of number representation. In recent years, software tools that allow programming of FPGAs have been significantly developed in order to allow programming of FPGAs from the level of model based design. This allows a faster and easier implementation of digital systems on FPGAs, compared to the classical approach of programming, which implies advanced skills in hardware description languages such as VHDL and Verilog HDL. When choosing between three previously mentioned hardware components it is necessary to make compromise between the flexibility, processing speed and energy consumption. Recently, on the market appeared hybrid components that represent a combination of the previously mentioned three components. The main goal of creation hybrid components is to achieve the highest possible energy efficiency while preserving performance requirements placed on SDR [8]. The PicoChip company has developed a multiprocessor IC that satisfies the need for real-time processing and requirements placed on SDR. According to [9] it is more optimal solution compared to ASICs, FPGAs or DSPs. In foregoing discussion the energy efficiency of hardware components is not analyzed. Consideration of energy efficiency is particularly important with respect to mobile devices. SDR systems are less energy efficient than conventional radio systems. The stated is completely understandable after taking into account system reconfigurability and the need to facilitate the implementation of various standards from a wide range of frequencies. Figure 1 represents comparison of hardware components related to programmability versus energy efficiency. By analyzing Fig. 1 it is clear that FPGAs provide improved performances over traditional microprocessor systems in relation to energy efficiency. Best performances in terms of energy efficiency have ASICs. However, considering the relationship between energy efficiency and programmability of components, it is obvious that their inverse relationship. Generally, the programmability results in a dissipation of energy, and thus reducing energy efficiency. When optimizing the energy efficiency of the entire SDR system it is important to observe the energy efficiency of analog-to-digital and digital-to-analog converters. SDR as a concept seeks to total elimination of RF part of transceiver chain, wich aims to implementation of the system where the analog-to-digital converter is directly connected to the antenna system [10]. In order to provide this, ADCs and DACs with high sample rates must be used. Current trends in the development of ADCs are determined by the need of software defined and cognitive radio [11]. Since 1995, the need for a resolution of ADC is satisfied, and there is no need for ADCs development with a resolution greater than 24 bits [11]. However, sampling rates of ADCs continue to grow with an exponential progression.

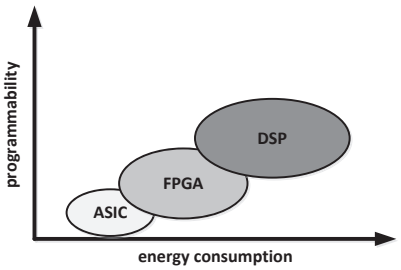


Fig. 1. Comparison of SDR hardware components

III. MODEL-BASED FPGA PROGRAMMING

The need for rapid system design and implementation on FPGA integrated circuits has initiated the development of software tools that support the design and programming of the FPGA chips from the level of model-based design. The two of the largest manufacturers of FPGA chips, Xilinx and Altera, have developed their own software tools that can be built on MATLAB/Simulink tool. When installing these tools, new libraries for blocks will be created in Simulink, which enable system design, and also its realization on FPGA chip from the model-based level without the need for substantial Hardware Description Language (HDL) coding.

A. Comparison of tools for model-based programming FPGA

In addition to Altera's and Xilinx's tools there are tools from other manufacturers with the same purpose. Among them, the most common in practice are following tools: Altera DSP Builder- Altera; Xilinx System Generator- Xilinx; Synphony Model Compiler- Synopsys; Simulink HDL Coder- MathWorks. Altera DSP Builder and Xilinx System Generator are solutions that allow programming of chips that produce exclusively Altera and Xilinx, respectively. Simulink HDL Coder is an attempt of MathWorks Company to develop a tool that allows programming of FPGA chips of any manufacturer with support for all the blocks from Simulink that support fixed-point arithmetic. However, it has not yet proven to be successful. According to many, the best solution is Synphony Model Compiler, which is an independent solution and provides support for work with FPGA chips of almost all manufacturers. Also, Synphony developed many generalized models of subsystems which are parts of system for digital signal processing. These models are available for use with the software above and are legally protected as intellectual property of Synphony. Besides the Synphony Model Compiler, Synopsys also developed Synphony C Compiler, which enables the generation of HDL code based on C and C + + code. In this paper the design and implementation of QPSK transmitter is based on Altera DSP Builder tool. Accordingly, the next subsection is focused on its working principle and the basic steps for successful design and implementation of a system on a chip.

B. Altera DSP Builder

Altera's DSP Builder allows the hardware to be abstracted to a higher level so that the FPGA and system waveform developers can operate in a common environment [12]. It is designed for digital signal processing and allows the use of Simulink, the MathWorks's model-based project environment, for FPGA design. DSP Builder provides Simulink graphics conversion to VHDL. After installing, DSP Builder contains a separate toolbox added in the standard Simulink library. The advantage of using DSP Builder is that there is no need to prior extensive experience with FPGA integrated circuit design and the design methodology of the Register Transfer Level (RTL). In this way, DSP Builder accelerates the design and implementation of systems for digital signal processing. From the standpoint of digital signal processing and telecommunications the availability of generalized blocks for: coding, performing FFT operations and filtering signals, is a great advantage. These blocks are included in the IP Megacore library and are intellectual property of Altera. Complete process of design and simulation in DSP Builder tool is shown Fig. 2.

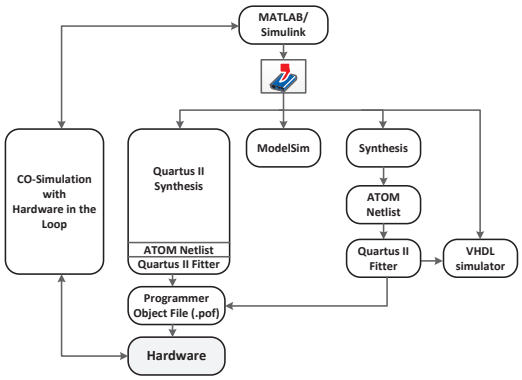


Fig. 2. Design and simulation in the DSP Builder tool

Basically, the process of system design and implementation on a FPGA chip using DSP Builder tool involves the following steps:

- Create a model using Simulink and DSP Builder blocks. Separate the DSP Builder library blocks from other blocks of source Simulink libraries by using the Input/Output blocks from DSP Builder IO and Bus library.
- Include a Clock block in the model from the DSP Builder AltLab library to specify the base clock, which must have a period in the range from 1ps to 2.1ms.
- Adjust a discrete solver in Simulink. For this purpose it is necessary to select the Configuration Parameters in the Simulation menu, and then select the Solver.
- For simulation results following, include some of Simulink blocks such as Scope or Spectrum Scope.
- If the simulation results are as expected, run Signal Compiler. As a result files are generated for RTL simulation and synthesis.

- Compile the design using Quartus II software and download to the FPGA chip. And, finally results testing.

IV. PRACTICAL IMPLEMENTATION OF QPSK TRANSMITTER

A. Hardware implementation

Hardware platform that is used for the implementation of the SDR consists of ALTERA DE2-115 development board, and HSMC data conversion card. The basic component of development board is Cyclone IV FPGA chip, whose characteristics are given in Table I. Besides the FPGA chip on development board there are many peripheral components, which are not used for implementation of SDR. Among all peripheral components it is important to note there are several oscillators- one of them with frequency of 50MHz is used as source of reference clock signal. For the adjustment of clock signal rate, PLLs are used to increase or decrease signal rate depending on the need.

HSMC data conversion card is used to convert digital signals to analog and vice versa. Therefore, development board together with data conversion card form platform for implementation of SDR transmitter. To establish functionality of platform it is necessary to provide communication between FPGA chip and data conversion card. Data conversion card is connected to the development board through HSMC (High Speed Mezzanine Card) interface. On data conversion card there are two AD and one 2-channel DA convertors. Characteristics of AD and DA converters are listed in Table II.

Logic elements	114.480
Embedded memory	3.888 Kbit
Embedded multipliers (18x18)	266
PLL	4
I/O pins	528

TABLE I
CHARACTERISTICS OF CYCLONE IV 2C70 FPGA

Number od AD converters	2
Number of DA converters (2 channels)	1
Resolution & sample rate for AD converters	14 bits & 150 Msps
Resolution & sample rate for DA converters	14 bits & 250 Msps

TABLE II
CHARACTERISTICS OF AD AND DA CONVERTERS

The first step towards ensuring the functionality of data conversion is to understand functionality of HSMC interface. HSMC interface consists of three pin banks with a total number of 173 pins. Each AD converter allocates 14 pins from HSMC interface for data. These two AD converters provide us two channels for analog-to-digital conversion. DA converter allocates 28 pins for data, 14 pins for each of the DA channels. Functionalities of other pins connected to AD and DA converters will not be discussed in this paper. It is important to emphasize that it is necessary to write Tcl (Tool Command Language) script with pin mappings to provide communication between the FPGA chip, the HSMC interface and AD/DA converters.

B. Simulation model analysis

The implemented QPSK modulator is shown on Fig. 3. Additionally, we used: Signal Compiler, PLL and Clock block. For testing purposes, the output of a pseudo-random number (PN) generator is fed to modulator input. The clock of PN sequence generator was set to 200 ns, so that the modulator could support a data rate of 5 Mbps. We generated 200 ns clock using PLL and clock signal from a 50 MHz oscillator.

According to Fig 3., the pseudorandom sequence is the input data of the parallel to serial (S/P) converter. The Signal Compiler block generated the HDL code based on the created model. Before HDL code generation it is necessary to specify the destination FPGA chip. The Clock block specified the relationship between the simulation and real clock. The PLL block enabled multiple clock generation (we created from the basic clock rate 50 MHz, 5 additional clocks). The Parallel to Serial subsystem consisted of: S/P, Multi-Rate DFF and Bus Splitter block, all from Altera's DSP Builder Blockset.

After S/P conversion each channel (I and Q) has unipolar signals. In accordance to QPSK modulation, the unipolar to bipolar conversion follows. Basis of the subsystem for unipolar to bipolar conversion is a Lookup table, which for input 0 gives output -1, and for input 1 the output 1. The I-channel Unipolar-to-Bipolar subsystem has additionally a delay unit for delay compensation between I and Q channels, incurred by S/P conversion. For filtering and carrier modulation purposes, signal upsampling is necessary. The Upsampling subsystem consisted of: Multi-Rate DFF and Upsampling block. The resulting signal frequency at the output is 25 Msps.

After upsampling, signal was shaped through Root-raised cosine (RRC) filter to avoid intersymbol interference. The RRC filter subsystem consisted of two filter blocks that performed signal scaling. The coefficients of FIR filter are usually much smaller than 1, so it is necessary to perform their scaling and converting the values in fixed-point representation. Consequently the coefficient values become much larger than 1, which leads to the need for signal scaling at the output of RRC filter. At the RRC filter output, signals have their representation in signed bit fractional format. Here one bit represented sign, while the remaining 13 bits are used for representation of decimal values. On Fig. 3 this is signed as SBF1_13. To provide the AD/DA functionality, input signals must be from proper pins that were previously mapped within a TCL script. The AD and DA convertor clocks were set to 100 MHz, which was also the sampling frequency. Generation of this clock signal was achieved by a counter from 0 to 1. After filtering the signal was shaped with a carrier signal. The carrier was generated by a numerical controlled oscillator (NCO). The NCO generated a sine and cosine signal with frequency 12.5 MHz, in integer format. An additional block realized the conversion from integer to signed fractional format. This block isolated the MSB bit from the other bits and used it for sign determination in signed bit fractional format. The other

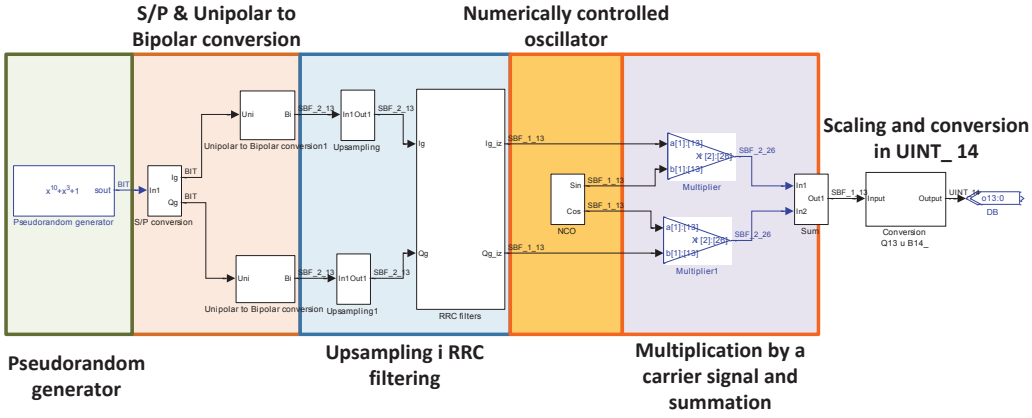


Fig. 3. QPSK simulation model

bits were used for representation of decimal values. This way ensured the desired conversion. At the end, I and Q modulator branch signals were summed. Since signal addition was made, it is necessary to re-execute scaling of the signal in order to allow its representation in SBF1_13 format. Then, the signal was returned to the integer format with 14 bits, and as such it could be forwarded to the DA converter.

V. Results

In the process of SDR transmitter design and implementation, it is important to perform simulation of designed model before its implementation. In this way it is possible to make a comparison of simulation results and results obtained after model implementation on FPGA chip. The results obtained in the simulation process are usually desired results, while real results obtained after implementation could deviate from them. Therefore, below are presented the results obtained in the simulation process, and the actual results obtained by an oscilloscope and spectrum analyzer. Fig 4. shows the signals at the modulator input, after S/P conversion and signals after unipolar to bipolar conversion.

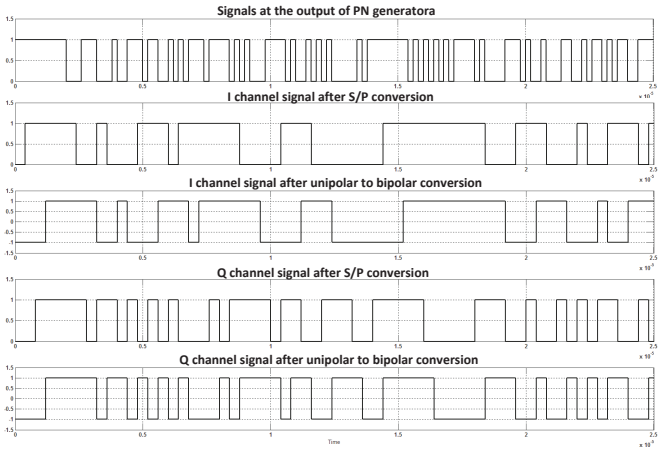


Fig. 4. Results of simulation (signals after S/P and Uniploar to Bipolar conversion)

It should be noted that the duration of QPSK signal after S/P conversion in I and Q channels is twice longer than its duration prior to S/P conversion. Also, when analyzing the figures that represent simulation results it is important to take into account delays due to signal propagation. After previously mentioned signal conversion, signal is upsampled and RRC filtered. Fig. 5 represents signals after filtering and multiplication with the carrier. Results for I and Q channels of modulator are identical. Consequently only signals recorded for I channel on oscilloscope and spectrum analyzer will be presented.

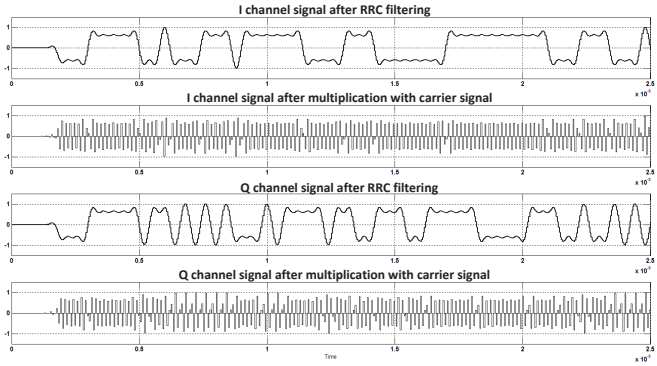


Fig. 5. Results of simulation (signals after RRC filtering and multiplication with carrier signal)

Fig. 6 represents the waveform, spectrum and eye diagram of I channel signal, respectively. It is clear that the waveform of the signal recorded on the oscilloscope is identical to waveform of signal obtained in the simulation process. Eye diagram is often used during observation of the signal characteristics with respect to noise and jitter. Vertical opening of the eye diagram defines noise margin, while the horizontal opening defines jitter margin of signal. In relation to the above, it is clear that the signal has good performance with regard to noise and jitter.

Fig. 7 shows spectrum of signal at the output of NCO and spectrum of I channel signal after multiplication with carrier

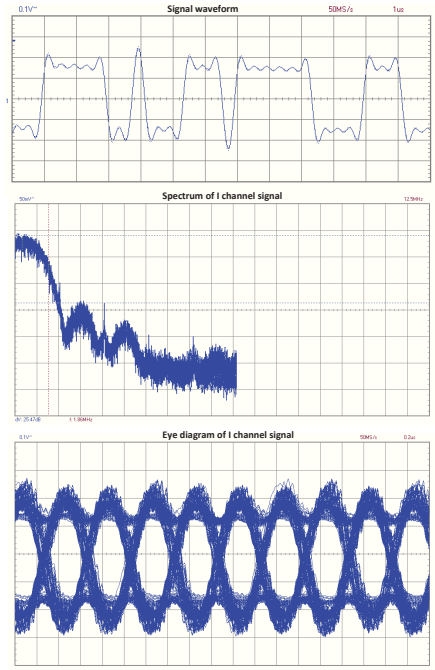


Fig. 6. Waveform, Spectrum and Eye diagram of I channel signal

signal, respectively. This NCO output signal is multiplied with signals at the output of RRC filters, and on this way digital up-conversion was achieved. Comparing the second figure with Fig. 6 it can be noted that the spectrum is now translated to frequency 12.5 MHz.

VI. Conclusion and Future Work

In different areas of Electrical Engineering, engineers try to implement the concepts of adaptability and intelligence in their systems. SDR technology is the basis for the implementation of reconfigurable, adaptive and cognitive radio systems. However SDR puts demands on the platform on which radio systems will be implemented. Generally, it is desirable to perform signal processing on the platforms for digital signal processing that have certain level of re-configurability such as DSPs and FPGAs. Signal processing in the analog domain, should be avoided as much as possible. These can be achieved by implementation of Digital Up and Down Converters and use of AD and DA converters with a high sampling frequency. In the process of selection of components that will form hardware platform it is necessary to take into account energy efficiency of components, especially in the case of implementation of radio which should provide a certain degree of mobility. The advent of FPGAs and some other reconfigurable platforms like Pico Arrays has extended the capabilities of SDR based system implementations.

In the process of design and implementation, the model based approach is used. Design and implementation of SDR system from higher level of abstraction has its advantages and disadvantages. Advantages are related to simplicity of

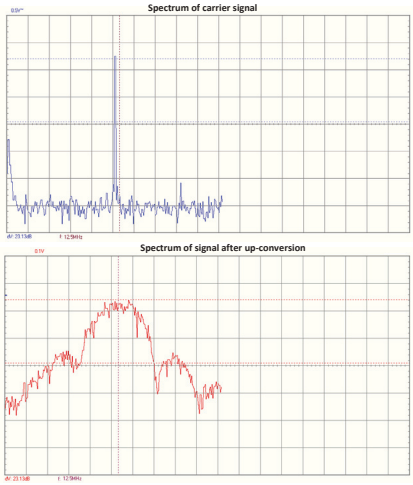


Fig. 7. Spectrum of Carrier and Signal after Up-conversion

implementation and troubleshooting based on results of simulation, while disadvantages are related to restriction to existing blocks.

Following the trends in the development of new standards for wireless communication it is clear that SDR and Cognitive radio are the future of wireless communications. SDR concept certainly extends the capabilities of conventional radio systems.

In the future we plan to implement more advanced modulation and demodulation techniques such as Orthogonal Frequency-division Multiplexing (OFDM). Our goal is to develop LTE FPGA based transceiver. For that we plan to use more specific software tools for design and implementation of radio system such as Gnu Radio and OSSIE.

References

[1] J. Mitola, "The software radio," in *IEEE National Telesystems Conference*, 1992.

[2] B. P. Lathi, *Modern Digital and Analog Communication Systems 4th edition*. Oxford University Press, 2010.

[3] M. I. Kostic, *Digitalni Telekomunikacioni Sistemi*. Naucna knjiga Beograd, 1994.

[4] A. Corporation, "Qpsk modem reference design," 2003.

[5] J. H. Reed, *Software Radio: A Modern Approach to Radio Engineering*. Prentice Hall PTR, 2002.

[6] J. Mitola, *An Integrated Agent Architecture for Software Defined Radio*. Doctoral disertation, KTH, 2000.

[7] M. Cummings and S. Haruyama, "Fpga in the software radio," *IEEE Communications Magazine*, vol. 37, pp. 108–112, 1999.

[8] M. S. Safadi and D. L. Ndzi, "Digital hardware choices for software radio (sdr) baseband implementation," in *Information and Communication Technologies ICTTA*, vol. 2, 2006, pp. 2623 – 2628.

[9] R. Baines and D. Pulley, "The picoarray and reconfigurable baseband processing for wireless basestations," *IEEE Colloquium on DSP enabled Radio*, pp. 1 – 11, 2003.

[10] Z. Hongyi, L. W. Jhon, G. C. Adam, and T. Alecander, "Airfpga: A software defined radio platform based on netfpga," *NetFPGA Developers Workshop*, 2009.

[11] B. Le, T. W. Rondeau, R. H. Jeffrey, and C. W. Bostain, "Analog-to-digital converters," *IEEE Signal Processing Magazine*, vol. 22, pp. 69 – 77, 2005.

[12] S. W. Cox and J. A. Seely, "Rapid fpga modem design techniques for sdrs using altera dsp builder," in *Global signal processing conference and expos for the industry*, 2004.

Mogućnost korištenja bežičnih komunikacija za uspostavu „mesh“ podatkovne mreže sa svim objektima pomorske signalizacije na Jadranu

I. Karin*, J. Vrlić*, Z. Belak* i I. Ružić**

* Plovput d.o.o., Split, Hrvatska

** Sveučilišni odjel za stručne studije, Split, Hrvatska

Sažetak - Na području Jadrana postoji veliki broj OPS-ova (objekata pomorske signalizacije) sa sve većom količinom navigacijske i pripadajuće opreme za napajanje koju je potrebno daljinski nadgledati u svrhu kvalitetnog održavanja. Kako bi se moglo daljinski nadzirati oprema na OPS-ovima, potrebno je sa svakim OPS-om imati kvalitetnu komunikacijsku vezu. U ovom raduje opisana mogućnost korištenja bežičnih komunikacija kako bi se kreirala isprepletena (mesh) podatkovna mreža sa svim objektima pomorske signalizacije koristeći RF module tvrtke DigiInternationalInc., uzimajući u obzir ograničenje u količini dostupne električne energije.

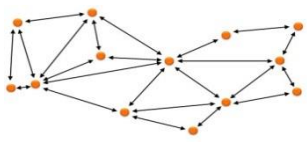
I. UVOD

Na Jadranu postoji veliki broj objekata pomorske signalizacije. Svaki objekt opremljen je određenom količinom navigacijske i druge elektroničke opreme te se taj broj svakodnevno povećava. Sva ugrađena oprema zahtjeva stalan nadzor rada kako bi se osigurala što veća razina sigurnosti plovidbe u Jadranskom moru. Iz tog razloga potrebno je uvesti nadzor svih važnih parametara na objektima pomorske signalizacije što će u nastavku ovog članka biti razrađeno. S obzirom da se spomenuti objekti nalaze na dislociranim lokacijama, postoje tehnička ograničenja za izradu ovakvog sustava. To je smanjena količina električne energije, velika udaljenost samog objekta od svih mrežnih infrastruktura, nepokrivenost resursima postojećih bežičnih sustava koji se koriste za prijenos podataka te mnogi drugi. Iz navedenih i objektivnih razloga postoji potreba za izgradnjom vlastite bežične komunikacijske mreže koja će zadovoljiti sve postavljene uvjete. U tu svrhu može biti upotrijebljen DigiMesh mrežni protokol zajedno sa komunikacijskim modulom istog proizvođača, pod imenom "XBee".

II. BEŽIČNO UMREŽAVANJE

A. DigiMesh mrežni protokol

DigiMesh je mrežni protokol koji posreduje u"peer-to-peer" umrežavanju za upotrebu u rješenjima bežičnog povezivanja. Tehničke karakteristike i arhitektura "peer-to-peer" mrežnog protokola omogućuju da ovaj protokol bude jednostavan za korištenje te opremljen s naprednim opcijama koje se upotrebljavaju kod umrežavanja. Jedna od važnijih naprednih opcija su mogućnost postavljanja usmjerivača u tzv. "sleep mode", mogućnost izrade i korištenja guste mreže te mnoge druge. Osim toga, DigiMesh operacije izvršava na bazi stalne povezanosti s protokolom te je optimiziran za što veće performanse same mreže. Na taj način mnogo manje vremena se gubi na adresiranje i ostale operacije vezane za samu mrežu, dok je izvršavanje korisničkih zahtjeva ili zahtjeva aplikacija znatno brže. U osnovi, DigiMesh je prijenosni protokol te može biti raspoređen u više različitih bežičnih sekcija širokog polja primjene, uključujući različite frekvencije kao i različite hardverske platforme. U ponudi postoji nekoliko gotovih RF modula koji su spremni za implementaciju u sustave sa različitim zahtjevima, a kao prilagodba za određeni sustav u ponudi su različiti adapteri i dodaci za upotrebu u različitim mrežnim izvedbama.



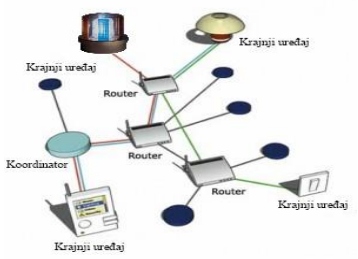
Slika 1. DigiMesh mrežni protokol

Osnovna prednost ovog mrežnog protokola je ta što je isti razvijen za rješavanje problema bežičnog povezivanja krajnjih točaka za široki spektar različitih zahtjeva i potreba te je idealno rješenje za korisnike koji trebaju: kvalitetnu i robusnu mrežu, snažan optimizirani protokol

sa „sleep mode“ opcijom kako bi se zadovoljili zahtjevi korisnika sa ograničenom količinom električne energije, mogućnost implementacije za frekvencije 900 MHz i 2.4 GHz itd. DigiMesh je dostupan u velikom broju proizvoda za različite namjene, a neki od zanimljivijih za problematiku koja je se u ovom članku obrađuje su XBee i XBee Pro.

B. ZigBee komunikacijski protokol

ZigBee je naziv za komunikacijski protokol koji koristi radio uređaj male snage baziran na standardu IEEE 802. Standard IEEE 802 je standard za izradu mreža. Najčešće se koristi u mrežama za prijenos podataka na veće udaljenosti, a to omogućuje da se mreža formira "ad-hoc", bez centralizirane kontrole te s mogućnošću da bez predajnika i prijemnika jake snage dođe do svih uređaja u mreži. U tom slučaju bilo koji ZigBee uređaj ima mogućnost vođenja mreže. Osim toga, ovaj komunikacijski protokol je usmjeren na aplikacije koje zahtijevaju nisku brzinu prijenosa podataka, dugo trajanje baterije te visoku razinu sigurnosti mreže. Upravo ove karakteristike odgovaraju zahtjevima za umrežavanja objekata pomorske signalizacije. Brzina mu je definirana na 250 kbit/s što najbolje odgovara aplikacijama koje zahtijevaju povremene prijenose podataka ili prijenose podataka sa prekidima. Iz toga proizlazi da je tehnologija koju ZigBee koristi jednostavnija i jeftinija od drugih wireless tehnologija kao što je Bluetooth ili Wi-Fi. Osim niske cijene, velika prednost mu je mala potrošnja energije pa je ova tehnologija umrežavanja široko rasprostranjena.



Slika 2. ZigBee komunikacijski protokol

ZigBee se proizvodi sa flash memorijom koja seže od 60 do 256 KB, ovisno o potrebama korisnika pa se tako koristi u industriji, medicini, znanstveno-istraživačkim djelatnostima itd. Radne frekvencije su 868 MHz u Europi, 915 MHz u Sjedinjenim Američkim Državama i Australiji te 2.4 GHz u ostalim državama diljem svijeta. Brzine prijenosa podataka sežu od 20 kb/s do 250 kb/s. Definiran je sa tri različite vrste uređaja: koordinator, usmjerivač i krajnji uređaj.

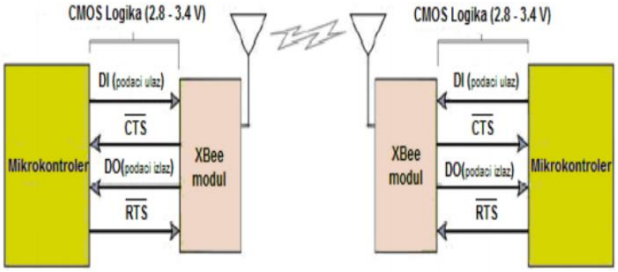
Koordinator – Odgovoran je za izbor kanala i PAN (Personal Area Network) ID. Koordinator započinje kreiranje nove mreže tj. kreira novi PAN. Nakon što je kreiran PAN, koordinator može dozvoliti usmjerivačima i krajnjim uređajima da se priključe na mrežu. On može slati i primati RF podatke i može asistirati u usmjeravanju podataka kroz međusobno povezane, isprepletene (mesh) mreže. Nije predviđen za baterijsko napajanje, odnosno nije baterijski napajan uređaj.
Usmjerivač – Usmjerivač se prvo mora priključiti u ZigBee PAN da bi mogao obavljati svoju funkciju. Nakon što se priključi on može dozvoliti drugim usmjerivačima i krajnjim uređajima da se priključe u PAN. Usmjerivač također može primati i slati RF podatke i može usmjeravati pakete podataka kroz mrežu. S obzirom da omogućava umrežavanje ostalih usmjerivača i krajnjih uređaja, te da sudjeluje u procesu usmjeravanja, usmjerivač ne može "spavati" i zato ne treba biti baterijski napajan.
Krajnji uređaj – Krajnji uređaj se mora priključiti u mrežu kao i usmjerivač, ali krajnji uređaj ne omogućava drugim uređajima da se priključe na PAN i ne sudjeluje u usmjeravanju podataka kroz mrežu. Ovi uređaji mogu slati i primati RF podatke. Najčešće su baterijski napajani uređaji. S obzirom da mogu biti u stanju spavanja (sleep), usmjerivač ili koordinator koji je dozvolio ovom uređaju da se priključi u mrežu mora sve pakete podataka namijenjenih ovom krajnjem uređaju buferirati i čuvati sve dok se ovaj ne probudi i ne bude u stanju primiti podatke koji su mu namijenjeni.
Usmjerivač ili koordinator koji dozvoljava krajnjem uređaju da se veže u mrežu i koji upravlja RF podacima namijenjenim tom krajnjem uređaju, naziva se roditelj krajnjeg uređaja (end device's parent), a krajnji uređaj je dijete (child). Mreža se kreira kada koordinator izabere kanal i PAN ID. Nakon što koordinator starta PAN, usmjerivači i krajnji uređaji se mogu priključiti u PAN. PAN ID je označio koordinator kada je startao PAN, a usmjerivači i krajnji uređaji mogu postati dio tog PAN-a i naslijediti PAN ID koordinatora kada se priključe u mrežu. ZigBee podržava usmjeravanje u mesh mreži, dozvoljavajući paketima podataka da prelaze višestruke čvorove, omogućavajući višestruke skokove (multiple hoops), kako bi stigli do odredišnog čvora. Ovo omogućava da ZigBee čvorovi budu rasprostranjeni na velikim prostorima, a da se pri tome omogući komunikacija među uređajima u mreži. Kada se uključe u PAN mrežu uređaji dobivaju mrežne 16-bitne adrese. Koordinator po defaultu ima mrežnu adresu"0".

C. XBee komunikacijski modul

XBee i XBee-PRO RF moduli razvijeni su da zadovolje zahtjeve IEEE 802.15.4 standarda, te kroz minimalne zahtjeve za energijom omoguće pouzdan prijenos podataka između uređaja i time podrže koncept bežične mreže. Moduli rade u opsegu radio-frekventnog područja bez posebnih dozvola, za potrebe industrije, naučnih istraživanja, medicine itd., u tzv. ISM 2.4GHz frekventnom opsegu i „pin-to-pin“ su kompatibilni. Generalno neke verzije modula imaju iznimno malu potrošnju energije (low-power). Postoje i jače verzije modula (kao što je i ova) koje obično imaju dodatno pojačalo za podršku mnogo većih udaljenosti. Naravno, time je potrošnja energije znatno veća, no za naše uvjete još uvijek prihvatljiva.

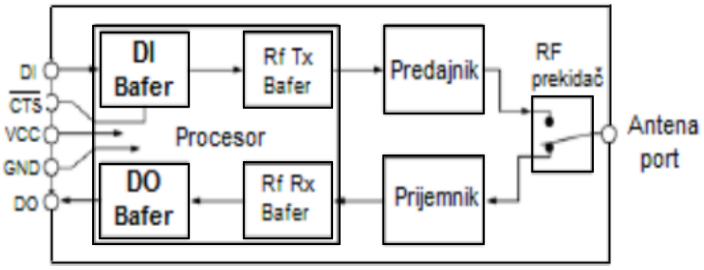
Slika 3. Xbee i Xbee PRO

Za komunikaciju sa host uređajem Xbee RF Modul koristi asinkroni serijski interfejs. Koristeći svoj serijski port modul može komunicirati sa bilo kojim kompatibilnim UART-om (Universal asynchronous receiver/transmitter). Uređaj koji ima UART interfejs može se spajati direktno na odgovarajuće pinove RF modula, kao što je prikazano na slici 4.



Slika 4. Data flow dijagram

Serijska komunikacija dakle ovisi o dva UART-a i mora biti konfigurirana sa odgovarajućim kompatibilnim vrijednostima (brzina prijenosa, paritet, start bit, stop bit, data bit). Podaci se unose u UART RF modula, koristeći DI pin (pin 3) modula, odnosno podaci pristigli serijskim putem u RF modul koristeći DI pin, smještaju se u DI Bufer. Hardverska kontrola toka podataka omogućena je CTS signalom. Naime, kada DI Buferu preostane 17 bajtova da postane pun, po defaultu se signalizira hostu (CTS signal u neaktivnom stanju - 1) da prestane sa slanjem podataka. Ponovno slanje podataka je moguće kada je dostupno 34 bajta memorije DI Bufera (CTS signal aktivno stanje - 0). Da bi se izbjegle ovakve situacije, potrebno je ili slati poruke manje od veličine DI Bufera (202 bajta) ili smanjiti brzinu prijenosa definiranu BD (Interface Data Rate) parametrom.

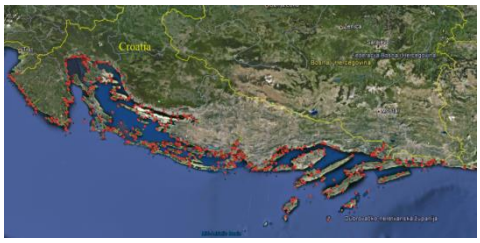


Slika 5. Unutarnji tok podataka

RF podaci, podaci pristigli RF prijenosom se smještaju u DO Bufer (slika 5.). Ukoliko DO bufer popuni svoj kapacitet, svi naknadno pristigli podaci će biti izgubljeni. Hardverska kontrola toka podataka omogućena je korištenjem RTS signala. Ako je RTS signal omogućen za kontrolu toka, parametar D6=1, podaci se neće slati u DO bufer sve dok se RTS signal ne deaktivira. Moguća situacija kada može doći do toga da DO bufer postane pun i da se desi prekoračenje kapaciteta (over-flow) je kada je brzina RF prijenosa veća od brzine prijenosa interfejs modula, tj. kada modul prima podatke od modula koji vrši prijenos većom brzinom od one kojom modul šalje podatke na host uređaj. Da bi XBee moduli mogli ispravno komunicirati, potrebno ih je na ispravan način konfigurirati. Moduli se konfiguriraju preko USB/serijskog interfejsa korištenjem odgovarajućih AT naredbi. Najjednostavniji način konfiguriranja je koristeći DIGI X-CTU program.

III. UMREŽAVANJE OBJEKATA SIGURNOSTI PLOVIDBE

Na hrvatskoj strani Jadrana postoji veliki broj objekata sigurnosti plovidbe, njih ukupno 589, koji su od ključne važnosti za odvijanje pomorskog prometa u hrvatskom teritorijalnom moru. Slika 6. prikazuje pozicije svih navedenih objekata.



Slika 6. Objekti pomorske signalizacije na Jadranu

Objekti su smješteni duž cijele hrvatske obale Jadrana, od Savudrije na krajnjem sjeveru do Prevlake na samom jugu

Republike Hrvatske. Zbog iznimno velikog broja objekata važnih za sigurnost plovidbe navedeni teritorij je podijeljen u sedam plovnih područja: Pula, Rijeka, Zadar, Šibenik, Split, Korčula i Dubrovnik. Podjela plovnih područja prikazana je slikom 7.

Slika 7. Podjela objekata po plovnim područjima

Svaki od navedenih objekata sastavljen je od osnovnih i dodatnih uređaja i opreme. Pod osnovne uređaje ubraja se primarna oprema poput elektronike svjetla, akumulatorskih baterija, elektroničkog bljeskača, solarnih panela, solarnog regulatora itd., dok dodatne uređaje mogu činiti vjetrenjače, sirene za maglu, detektori magle, rakon uređaji, AIS predajnici itd. Svu tu opremu potrebno je nadgledati te u svakom trenutku imati uvid u njeno stanje i ispravnost, a osnovni preduvjet za to je izgradnja komunikacijske mreže između svih objekata. Nadzor opreme bi mogao biti izveden umrežavanjem svih objekata DigiMesh mrežnim protokolom i to upotrebom Xbee modula. S obzirom da se u slučaju umrežavanja objekata sigurnosti plovidbe radi o velikom broju objekata koji su ujedno jedan od drugoga udaljeni i do nekoliko desetaka nautičkih milja potrebno je upotrijebiti modul iz serije Xbee Pro koji može zadovoljiti potrebe da bi signal došao do svih objekta. Za objekte koji se nalaze na međusobno manjoj udaljenosti moguće koristiti i osnovnu verziju Xbee modula. Osim udaljenosti potrebno je zadovoljiti i uvjet optičke vidljivosti dvaju modula koji međusobno komuniciraju pa i na to treba obratiti pozornost. Za objekte koji su dislocirani te se ne mogu uklopiti u ovaj način umrežavanja koristiti će se alternativno komunikacijsko rješenje koje je na toj lokaciji dostupno (GPRS, 3G, UHF linkovi, mikrovalni linkovi...). Objekata sigurnosti plovidbe ima više vrsta; svjetionici s posadom, svjetionici bez posade, obalna svjetla, svjetleće oznake itd. Idejno rješenje ovog načina umrežavanja je da se područje Jadrana, odnosno plovna područja podijele u mikroregije. Svaka mikroregija ima jedan ili više centara mikroregije koji su povezani sa svim objektima sigurnosti plovidbe u svojoj blizini kao i međusobno. Centri mikroregija bi bili isključivo svjetionici sa posadom iz praktičnih razloga, ukoliko je to tehnički izvedivo. Slika 8. prikazuje pozicije centara unutar svake od mikroregija (označen bijelom oznakom „O“).

Slika 8. Centri unutar mikroregija

Kao što je već navedeno, udaljenost između pojedinih objekata je i do nekoliko desetaka nautičkih milja, te je iz tog razloga potrebno koristiti uređaj Xbee oznake "Pro 900HP Programmable" koji ima domet do 28 NM (nautička milja) uz upotrebu adekvatne "high-gain" antene. Osim velikog dometa, uređaj podržava brzinu prijenosa podataka do 200 Kbps te izlaznu snagu od 250W uz potrošnju od maksimalno 230 mA za vrijeme predaje ili prijema. Upotrebom ovog uređaja zadovoljeni su svi osnovni preduvjeti za izgradnju mesh podatkovne mreže uz minimalnu potrošnju električne energije, što je jedan od važnijih uvjeta. S obzirom da je maksimalni domet Xbee Pro 900HP 28 NM u idealnim uvjetima navedeni podatak uzeti ćemo s rezervom, pa ćemo radi pouzdanijeg rada sustava pokušati maksimalnu udaljenost između pojedinih točaka sveti na maksimalnih 20 NM. Na taj način sustav će biti pouzdaniji te se neće koristiti njegovi maksimalni resursi što će povoljno utjecati na još neke od zadanih uvjeta.

Prvo plovno područje koje treba biti pokriveno je plovno područje Pula. Nalazi se na samom sjeveru hrvatskog dijela Jadrana te obuhvaća ukupno 66 objekata pomorske signalizacije, od čega su tri objekti sa posadom. Smješteni su na rubnim dijelovima i u sredini mikroregije te sa takvim geografskim položajem odgovaraju stvaranju podatkovne mreže. Slika 9. prikazuje povezanost mikroregije Pula.

Slika 9. Povezanost mikroregije Pula

Mikroregija Pula podijeljena je na tri centra, a to su Savudrija, Sv. Ivan na pučini i Porer. Centri mikroregije označeni su bijelim oznakama na slici 9. Crvene kružnice označavaju područja koja su pokrivena signalom Xbee uređaja. Promjer kružnice svojim rubnim dijelovima označava na karti udaljenost od 20 NM te je iz slike 9. vidljivo na koji način je mreža u ovoj mikroregiji formirana. Svi objekti obuhvaćeni trima kružnicama u

optičkoj su vidljivosti te mogu komunicirati jedan sa drugim, a pojedini čak i sa više njih te nije potrebno uvođenje alternativnih načina povezivanja.

Plovno područje Rijeka je područje sa nešto većim brojem objekata sigurnosti plovidbe. Ono ima čak 120 objekata koji trebaju biti pokriveni podatkovnom mrežom, od čega je samo svjetionik Mlaka nastanjen posadom. Smješten je na samom sjeveru plovnog područja i označen bijelom (slika 10.) bojom te kao takav jedan od pet centara mikroregije Rijeka. Ostali centri odabrani su zbog svog geografskog položaja ali i zbog mogućnosti da u budućnosti navedeni objekti postanu svjetionici sa posadom i na taj način kao centar mikroregije mogu nadgledati ostale objekte u svojoj blizini.

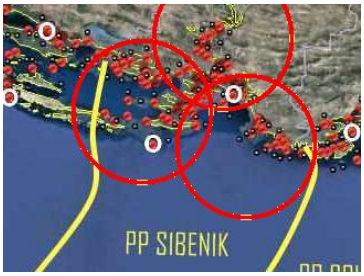
Slika 10. Povezanost mikroregije Rijeka

Svi objekti nalaze se u optičkoj vidljivosti sa bar po jednim susjednim objektom te na taj način zadovoljavaju uvjete za uspostavu mesh mreže.

Plovno područje Zadar također je veliko plovno područje sa sveukupno 163 objekata pomorske signalizacije. Od tog broja samo četiri objekta su svjetionici sa posadom te su zamišljeni kao centri mikroregije. To su redom od sjevera prema jugu Grujica, Veli Rat, Oštri Rat i Babac (slika 11.). Ravnomjerno su raspoređeni po plovnom području odnosno mikroregiji te tako pružaju idealne preduvjete za stvaranje podatkovne mreže. Osim toga, objekti u mikroregiji Zadar iznimno su blizu jedan drugom te zahvaljujući velikom dometu XBee Pro modula moguće je okolnim putovima povezati pojedine objekte van optičke vidljivosti.

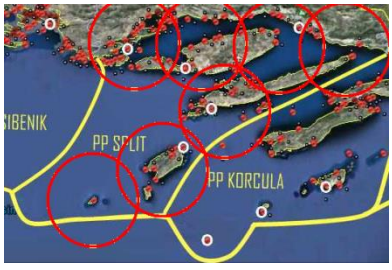
Slika 11. Povezanost mikroregije Zadar

Površinom najmanje plovno područje je plovno područje Šibenik. Pokriveno je sa ukupno 86 objekata pomorske signalizacije, od čega su samo dva pomorski svjetionici, na karti označeni bijelim oznakama. To je pomorski svjetionik Jadrija te svjetionik Blitvenica. Svjetionik Blitvenica je ujedno i najisturenija točka prema otvorenom moru u plovnom području i mikroregiji Šibenik te je kao takva zbog svoje važnosti za pomorski promet nastanjena sa svjetioničarskom posadom. S obzirom na stalnu svjetioničarsku posadu te činjenicu da je veoma blizu i u optičkoj vidljivosti sa ostalim objektima u ovom području povoljno je da bude jedan od dva centra mikroregije Šibenik. Mikroregija i povezanost unutar iste prikazani su slikom 12.



Slika 12. Povezanost mikroregije Šibenik

Plovno područje Split geografski je razvijeno i veoma veliko plovno područje. Obuhvaća ukupno 87 objekata pomorske signalizacije od čega je šest pomorskih svjetionika, a samo tri sa svjetioničarskom posadom. Pošto se radi o poprilično velikom plovnom području, za stvaranje funkcionalne mikroregije potrebno je oformiti veći broj centara mikroregije. Iako je u ostalim plovnim područjima praksa bila da svaki centar mikroregije ima svjetioničarsku posadu, u ovom plovnom području situacija bi bila nešto drugačija. Naime, mikroregija Split ima samo tri svjetionika sa posadom dok geografski oblik zahtjeva minimalno pet centara. Iz tog razloga potrebno je dodati još dva centra koji bi bili smješteni u pomorskim svjetionicima bez posade, ali pomno odabranim uzimajući u obzir da u perspektivi postoji mogućnost da se na iste uključi posada. Na slici 13. su označeni bijelim oznakama; Murvica, Ražanj, Sv. Petar, Stončica i Pokonji Dol.



Slika 13. Povezanost mikroregije Split

Plovno područje Korčula ima mali broj objekata pomorske signalizacije, njih svega 33 od čega su tri nastanjena sa svjetioničarskom posadom. Svojim položajem i oblikom zahtjeva tri centra mikroregije i oni su odabrani i označeni bijelim krugovima na slici 14. To su od sjevera prema jugu Sušac, Struga i Glavat. Razlika ove mikroregije od svih do sada opisanih je ta što je njome obuhvaćen i otok Palagruža. Otok Palagruža je najjužnija točka na kojoj se nalazi oprema za pomorsku signalizaciju na hrvatskom dijelu Jadrana. Udaljena je preko 30 NM od najbližeg sličnog objekta, otoka Sušca. Zbog navedene udaljenosti, koja je prilično velika, vidljivo je da ne postoji Xbee modul koji može vršiti komunikaciju i prijenos podataka između Palagruže i najbližeg joj otoka. Radi toga, potrebno je razmotriti druge načine prijenosa podataka koji su dostupni na toj lokaciji te je na taj način povezati sa ostalim objektima u mikroregiji. Uzimajući u obzir sve parametre koji su vezani uz otok Palagružu, analizirajući infrastrukturu koja je tamo dostupna kao i sve ostale faktore, donesen je zaključak kako je jedino pouzdano rješenje prijenos podataka alternativnim načinima komunikacije koji su u ovom radu navedeni. GSM mreža navedeni otok pokriva dovoljno dobro da bi prijenos podataka mogao biti ostvaren u svakom potrebnom trenutku. Na taj način podaci bi se prikupljali lokalno na objektu te u zadanim intervalima odašiljali najbližem centru mikroregije, a to je u ovom slučaju otok Sušac. Učestalost odašiljanja podataka prema centrima mikroregije može biti određena tek nakon detaljnog ispitivanja tehničkih uvjeta i potreba na zadanoj lokaciji. Na slici 14. prikazana je povezanost mikroregije Korčula, kao i lokacija otoka Palagruže koja je označena bijelim kvadratom.

Slika 14. Povezanost mikroregije Korčula

Plovno područje Dubrovnik najmanje je plovno podruje sa ukupno 34 objekta pomorske signalizacije. Od tog broja samo je jedan objekt sa svjetioničarskom posadom, otok Sv. Andrija koji je ujedno i centar ove mikroregije. Na slici 15. označen je bijelim krugom. Svi objekti unutar mikroregije, osim onog krajnjeg,Rt Oštra, su raspoređeni tako da komunikacijski moduli bez smetnji mogu prenositi signal jedan između drugoga. Sa jedinim izoliranim objektom koji je spomenut veza ne može biti uspostavljena upotrebom Xbee modula zbog slične situacije kao i na otoku Palagruži. Rt Oštra nalazi se na krajnjem jugu Hrvatske, na samoj granici sa Crnom Gorom te je komunikacija moguća jedino upotrebom alternativnih načina komunikacije.

Slika 15. Povezanost mikroregije Dubrovnik

IV. ZAKLJUČAK

Cilj ovo rada bio je iznaći idejno rješenje za povezivanje svih objekata pomorske signalizacije na Jadranu. Svi objekti koji su obrađeni opremljeni su elektroničkom opremom koju je potrebno nadzirati u svrhu veće razine sigurnosti plovidbe te se ta količina opreme konstantno povećava. Zbog stalnog dodavanja nove opreme bitno je da sustav za nadzor bude nadogradiv bez većih preinaka, odnosno da se bilo koja nova oprema prilikom ugradnje na objekt bez problema uključi u sustav za nadzor objekata. U tu svrhu, kao idejno rješenje odabran je DigiMesh mrežni protokol koji upotrebom Xbee modula posreduje u umrežavanju bežičnim putem. DigiMesh mrežni protokol je pogodan za ovakvu vrstu umrežavanja zbog niza naprednih mogućnosti koje pruža, kao što je veoma niska potrošnja energije, malene dimenzije, veliki broj mogućnosti spajanja dodatnih uređaja te jednostavno konfiguriranje. Naime, DigiMesh mrežni protokol vrši automatsko adresiranje svih uređaja u mreži, a s obzirom da broj objekata pomorske signalizacije prelazi pet stotina ovakav način konfiguracije je od iznimne važnosti. Zaključno, u radu je razrađeno idejno rješenje za povezivanje svih objekata pomorske signalizacije na Jadranu. Ono je najpogodnije rješenje zbog razvedenosti Jadranske obale, pouzdanosti, jednostavnosti održavanja, male potrošnje energije te mnogih drugih dobrih karakteristika. Navedene su mane i nedostatci pojedinih objekata i njihovih lokacija ali i pronađena alternativna rješenja za umrežavanje istih. Time bi svi objekti bili objedinjeni u mesh mrežu te bi bio uspostavljen daljinski nadzor istih.

V. LITERATURA

[1] ZigBee Alliance, "ZigBee specification",January 17, 2008

[2] X-CTU Configuration & Test Utility Software – documentation (http://ftp1.digi.com/support/ documentation/90001003_A.pdf)

[3] Product Manual v1.xEx - 802.15.4 Protocol, "XBee®/XBee-PRO® RFModules

ENVIRONMENTAL PROTECTION

Analiza rezultata mjerenja elektromagnetskih polja u mobilnim mrežama Hrvatskog Telekoma

Saša Gros*, Ružica Gajić**
* Končar – Institut za elektrotehniku, Zagreb, Hrvatska, sgross@koncar-institut.hr
** Hrvatski Telekom, Zagreb, Hrvatska, ruzica.gajic@t.ht.hr

Elektromagnetska polja emitirana od strane visokofrekvencijskih izvora u mobilnim mrežama su stalni predmet interesa javnosti i stručne zajednice. U Republici Hrvatskoj su dozvoljene razine referentnih veličina, definiranih u Pravilniku o zaštiti od elektromagnetskih polja, strože u odnosu na preporuku Međunarodne komisije za zaštitu od neionizirajućeg zračenja. Cilj rada je utvrditi zadovoljavaju li izmjerene referentne veličine strože granične razine propisane u Republici Hrvatskoj, u slučaju rada jednog izvora te u slučaju istodobnog rada više izvora elektromagnetskih polja. U ovom radu su analizirane izmjerene vrijednosti referentnih veličina u mobilnim mrežama Hrvatskog Telekoma. Razmatrane su bazne postaje GSM, DCS i UMTS tehnologije. Izmjerene vrijednosti u okolini izvora elektromagnetskih polja su uspoređene sa zakonom propisanim graničnim razinama. Analiza je pokazala da svi izvori elektromagnetskih polja u mobilnim mrežama Hrvatskog Telekoma zadovoljavaju referentne granične veličine određene Pravilnikom o zaštiti od elektromagnetskih polja u slučajevima samostalnog rada izvora, kao i u slučajevima istodobnog rada više izvora.

I. UVOD

Broj korisnika mobilnih komunikacija je krajem 2012. godine iznosio više od 6 milijardi [1]. Eksponencijalni razvoj mobilnih komunikacija, u smislu povećanja broja korisnika ali i u smislu povećanja broja novih usluga i načina korištenja, otvorili su mnoga pitanja o sigurnosti uporabe i utjecaju na ljudsko zdravlje.

U Republici Hrvatskoj na snazi je Pravilnik o zaštiti od elektromagnetskih polja, izdan od strane Ministarstva zdravlja [2]. U odnosu na preporuke Međunarodne komisije za zaštitu od neionizirajućeg zračenja (ICNIRP – International Commission on Non-ionizing Radiation Protection) [3], Pravilnik o zaštiti od elektromagnetskih polja propisuje znatno strože granične razine. Kako bi utvrdio jesu li i u strožim regulatornim okvirima zadovoljene granične razine, Hrvatski Telekom je proveo niz mjerenja u okolini izvora elektromagnetskih polja u svom vlasništvu. Mjerenja su obavljale isključivo tvrtke koje posjeduju ovlaštenje Ministarstva zdravlja, sukladno odredbama Pravilnika o zaštiti od elektromagnetskih polja. Rezultati mjerenja su analizirani u odnosu na propisane granične razine, u slučaju rada jednog izvora, kao i u slučaju istodobnog rada više izvora elektromagnetskog polja.

II. REGULATIVA U VEZI ZAŠTITE OD ELEKTROMAGNETSKIH POLJA

Na temelju znanstvenih istraživanja utvrđene su razine elektromagnetskih polja koje se smatraju sigurnima i nemaju štetnog učinka na ljudsko zdravlje. Osnovne smjernice je izdala Međunarodna komisija za zaštitu od neionizirajućeg zračenja.

Krovni zakon u Republici Hrvatskoj u području elektromagnetskih polja emitiranih u mobilnim mrežama je Zakon o zaštiti od neionizirajućeg zračenja [4]. Prema definiciji iz navedenog zakona, neionizirajuće zračenje jesu elektromagnetska polja i elektromagnetski valovi frekvencije niže od 3.000.000 GHz ili ultrazvuk frekvencije niže od 500 MHz koji u međudjelovanju s tvarima ne stvaraju ione.

Pravilnik o zaštiti od elektromagnetskih polja podrobnije opisuje načela i provedbu zaštite od djelovanja elektromagnetskih polja te donosi temeljna i referentna ograničenja.

Vezano uz izloženost elektromagnetskim poljima, Pravilnik o zaštiti od elektromagnetskih polja definira područja povećane osjetljivosti i područja profesionalne izloženosti. Područja povećane osjetljivosti su zgrade namijenjene boravku ljudi, čestice na kojima su izgrađene zgrade stambene namjene, škole, vrtići, rodilišta, bolnice, smještajni turistički objekti, dječja igrališta te površine neizgrađenih čestica koje su prema urbanističkom planu namijenjene navedenoj uporabi. Područja profesionalne izloženosti jesu područja radnih mjesta koja nisu u području povećane osjetljivosti i na kojima se pojedinci mogu zadržavati do 8 sati dnevno, pri čemu je kontrolirana njihova izloženost elektromagnetskim poljima. U ovom radu se veća pažnja posvećuje područjima povećane osjetljivosti, jer su u tom području ograničenja stroža.

U praktičnoj primjeni su značajna referentna ograničenja jer se ona odnose na mjerljive veličine čijim nadzorom se osigurava zadovoljenje temeljnih ograničenja. U frekvencijskom području elektromagnetskih izvora koji se koriste u mobilnim komunikacijama definirane su sljedeće referentne veličine: jakost električnog polja, jakost magnetskog polja, gustoća magnetskog toka, gustoća snage (ekvivalent ravnog vala).

Prema članku 2. stavak 3 Pravilnika o zaštiti od elektromagnetskih polja, zbog mogućih znanstveno još nepotvrđenih štetnih zdravstvenih utjecaja elektromagnetskih polja, primjenom načela predostrožnosti, granične razine su smanjene u odnosu na pripadna temeljna ograničenja.

Frekvencijski spektar koji se koristi u mobilnim mrežama Hrvatskog Telekoma obuhvaća područja od 400 – 2000 MHz i područje od 2 - 10 GHz, a referentna ograničenja za ta frekvencijska područja su definirana u Prilogu 2 Pravilnika o zaštiti od elektromagnetskih polja.

U odnosu na preporuke Međunarodne komisije za zaštitu od neionizirajućeg zračenja, na referentne veličine u navedenim frekvencijskim područjima primijenjen je faktor umanjenja koji za područje povećane osjetljivosti iznosi 2,5 – 6,25, ovisno o referentnoj veličini (Tablica 1).

TABLICA I. USPOREDBA REFERENTNIH VELIČINA ZA PODRUČJE POVEĆANE OSJETLJIVOSTI PREMA PRAVILNIKU O ZAŠTITI OD ELEKTROMAGNETSKIH POLJA I PREPORUCI ICNIRP-A

Referentna veličina	**Područje povećane osjetljivosti**		
	Pravilnik o zaštiti od elektromagnetskih polja	***ICNIRP preporuka***	***Faktor umanjenja u odnosu na ICNIRP preporuku***
Jakost električnog polja E (V/m)	400-2000 MHz: **0,55 $f^{1/2}$** 2-300 GHz: **24,4**	400-2000 MHz: 1,375 $f^{1/2}$ 2-300 GHz: 61	**2,5**
Jakost magnetskog polja H (A/m)	400-2000 MHz: **0,00148 $f^{1/2}$** 2-300 GHz: **0,064**	400-2000 MHz: **0,0037 $f^{1/2}$** 2-300 GHz: **0,16**	**2,5**
Gustoća magnetskog toka B (μT)	400-2000 MHz: **0,0018k $f^{1/2}$** 2-300 GHz: **0,08**	400-2000 MHz: **0,0046 $f^{1/2}$** 2-300 GHz: **0,20**	**2,5**
Gustoća snage (ekvivalentnog ravnog vala) S_{ekv} (W/m^2)	400-2000 MHz: **f/1250** 2-300 GHz: **1,6**	400-2000 MHz: **f/200** 2-300 GHz: **10**	**6,25**

Iz tablice I. vidljivo je da primijenjeni faktor umanjenja donosi znatno strože granične referentne razine u Republici Hrvatskoj, u području frekvencija koje upotrebljavaju mobilne komunikacije.

Dodatno, Pravilnik o zaštiti od elektromagnetskih polja definira i pojam značajnog izvora. Sukladno članku 2. Pravilnika o zaštiti od elektromagnetskih polja, značajan izvor elektromagnetskog polja pojedine frekvencije jest onaj stacionarni izvor čije elektromagnetsko polje u području povećane osjetljivosti, ili u području profesionalne izloženosti, doseže barem 10% iznosa granične razine zadane za tu frekvenciju.

Pregledom dostupnih standarda prikupljenih u projektu Svjetske zdravstvene organizacije (*EMF Project)* [5], autori nisu našli praksu postavljanja granice od 10 % kao granice za značajni izvor, niti su pronašli raširenu praksu definiranja značajnog izvora. Stav Svjetske zdravstvene organizacije i Međunarodne komisije za zaštitu od neionizirajućih zračenja je da su sve razine elektromagnetskih polja koje se nalaze unutar dozvoljenih vrijednosti temeljnih i referentnih veličina sigurne za zdravlje ljudi.

Pravilnik o zaštiti od elektromagnetskih polja definira i uvjete koji moraju biti zadovoljeni u slučaju istodobnog djelovanja izvora elektromagnetskih polja više frekvencija. Na takvim lokacijama svaki od izvora elektromagnetskog polja mora udovoljavati graničnim razinama, a dodatno treba provjeriti i njihov zajednički učinak prema zahtjevima članka 13. Pravilnika o zaštiti od elektromagnetskih polja, prema izrazu:

$$\sum_{>1MHz}^{300GHz}\left(\frac{E_f}{E_{g,f}}\right)^2 \le 1 \qquad (1)$$

gdje je :

E_f = efektivna vrijednost jakosti električnog polja u V/m na frekvenciji f

$E_{g,f}$ = efektivna vrijednost jakosti granične razine električnog polja u V/m na frekvenciji f

III. NAČIN PROVOĐENJA MJERENJA, MJERNA OPREMA I MJERNA NESIGURNOST

A Način provođenja mjerenja

Mjerenja jakosti eleketromagnetskih polja imaju za cilj utvrditi razinu neionizirajućih zračenja na mjestima povećane osjetljivosti i profesionalne izloženosti te izmjerene vrijednosti usporediti s dozvoljenim razinama prema Pravilniku o zaštiti od elektromagnetskih polja. Općenito, sustav za mjerenje jakosti elektromagnetskih polja sastoji se od mjerne antene, kabela za spajanje i mjernog instrumenta. Ovisno o korištenoj anteni i mjernom instrumentu, mjerenje razina elektromagnetskog polja može se provoditi u širokom i uskom pojasu frekvencija. Pri širokopojasnom mjerenju mjeri se ukupna razina svih elektromagnetskih polja koja ulaze u izabrano područje frekvencija. Kod uskopojasnih se mjerenja mjeri iznos polja na svakoj pojedinoj frekvenciji. Sva mjerenja obavljena su u uskom pojasu frekvencija kako bi se isključivo vrednovao utjecaj frekvencijskog područja baznih postaja Hrvatskog Telekoma.

Također, osim elektromagnetskog vala razmatra se i gustoća snage elektromagnetskog vala. U većini se slučajeva pod gustoćom snage za proizvoljni elektromagnetski val uzima ekvivalentna gustoća snage planarnog vala. U uvjetima mjerenja u slobodnom prostoru, u dalekoj zoni izvora, gustoća snage iznosi:

$$S = E \cdot H \qquad (2)$$

Ovdje su E i H efektivne vrijednosti električnog i magnetskog polja.

U slobodnom prostoru, u dalekoj zoni izvora, iznosi električnog i magnetskog polja povezani su izrazom:

$$\eta = \frac{E}{H} \qquad (3)$$

gdje je η valna impedancija slobodnog prostora i iznosi

377 Ω.

Iz navedena dva izraza, gustoća snage u dalekom polju može se izraziti i kao:

$$S = \frac{E^2}{\eta} = H^2 \cdot \eta \qquad (4)$$

Parametri elektromagnetskog polja mogu se mjeriti u bliskoj i dalekoj zoni izvora. U praktičnoj se primjeni uobičajeno uzimaju granice bliske odnosno daleke zone oko izvora. Daleka zona izvora zauzima područje za koje udaljenost od izvora *R* zadovoljava izraz:

$$R \geq \frac{2D^2}{\lambda} \qquad (5)$$

gdje je *D* najveća dimenzija izvora, a λ valna duljina elektromagnetskog vala u slobodnom prostoru.
Granice bliske zone dane su kao:

$$0{,}62\sqrt{\frac{D^3}{\lambda}} \leq R \leq \frac{2D^2}{\lambda} \qquad (6)$$

Električno i magnetsko polje u bliskoj zoni izvora nisu jednostavno matematički povezani preko impedancije slobodnog prostora. Zato u bliskoj zoni treba odvojeno mjeriti električno i magnetsko polje.

Mjerenje elektromagnetskog polja u visokofrekvencijskom području provodi se u skladu s Pravilnikom o zaštiti od elektromagnetskih polja te normom HRN EN 61156:2001 [6]. Pravilnik o zaštiti od elektromagnetskih polja predviđa da se mjerenja u okolini izvora elektromagnetskih polja moraju obaviti u tri točke u području profesionalne izloženosti te u tri točke u području povećane osjetljivosti.

Ovisno o lokaciji bazne postaje nemoguće je očekivati da se na svakoj lokaciji pronađu tri točke u području povećane osjetljivosti i tri točke u području profesionalne izloženosti (primjerice, za bazne postaje postavljene u tunelima nije moguće naći područje povećane osjetljivosti). U takvim slučajevima mjeritelji su se rukovodili principom da svakako izmjere minimalno šest mjernih točaka u okolini svake bazne postaje. Ponekad se radi o kombinaciji oba područja izloženosti, a u većini slučajeva je svih šest točaka u jednom od navedenih područja. Najjače se zračenje očekuje u glavnom snopu dijagrama zračenja antene, pa se mjerne točke pronalaze u glavnom snopu svakog sektora, u područjima gdje se očekuje nazočnost ljudi. Konačni izbor optimalnog položaja mjerne točke ovisi o konkretnoj situaciji na terenu.

Prilikom mjerenja posebnu pozornost treba posvetiti provjeri homogenosti polja u okolini mjernih točaka koja uključuje poštivanje dovoljnog razmaka mjeritelja od mjernog sustava, te zadovoljavanju kriterija daleke zone od izvora zračenja, koji ovisi o dimenzijama antene bazne postaje i frekvencijskom području mjerenja.

B Mjerna oprema

Mjerni se sustav u općem slučaju sastoji od mjerne antene, spojnog kabela i mjernog instrumenta. Mjerna oprema koja je korištena za mjerenja sastoji se od mjernog sustava FIELD NOSE proizvođača Seibersdorf Research te sustava TS-EMF proizvođača Rohde & Schwarz. Mjerni sustavi su u cijelosti automatizirani i upravljani računalima. Također, u predviđenim vremenskim periodima oprema se umjerava u akreditiranim inozemnim laboratorijima čime se osigurava sljedivost rezultata mjerenja. Mjerni sustavi sastoje se od slijedećih instrumenata:

- Antena PCD 8250 za mjerenja u području od 80 MHz do 2,5 GHz,
- Antena TS-EMF za mjerenja u frekvencijskom području od 30 MHz do 2,5 GHz,
- Spektralni analizator Anritsu MS2711B za mjerenja u frekvencijskom području od 100kHz do 3 GHz,
- Spektralni analizator Rhode & Schwarz FSH6 za mjerenja u frekvencijskom području od 100 kHz do 6 GHz.

C Mjerna nesigurnost

Važan podatak za kvantificiranje mjernih rezultata je mjerna nesigurnost. Veličine koje možemo kvantificirati su:

- nesigurnost pokazivanja amplitude na analizatoru spektra,
- nesigurnost antenskog faktora mjerne antene,
- nesigurnost slabljenja kabela.

Veličine koje ne možemo kvantificirati su:

- utjecaj okoline (temperatura, vlažnost zraka i sl.),
- utjecaj mjerne antene na raspodjelu polja na lokaciji mjerenja,
- utjecaj mjeritelja.

Za korištenu mjernu metodu, podatke o mjernim nesigurnostima pojedinih veličina čije utjecaje možemo kvantificirati dobivamo iz umjernica antena, spojnih kabela i mjernih instrumenata. Podaci o pojedinim mjernim nesigurnostima i izvorima iz kojih se određuju, odgovarajućim koeficijentima osjetljivosti te njihovim doprinosima ukupnoj mjernoj nesigurnosti navedeni su u tablici II. Na osnovi podataka iz tablice 2 računa se i proširena mjerna nesigurnost. Za vjerojatnost 95% faktor sigurnosti je $K = 2$, pa je ukupna mjerna nesigurnost mjerne metode ± 3,24 dB (45,2%).

TABLICA II. PRIMJER ISKAZIVANJA MJERNE NESIGURNOSTI ZA SUSTAV FIELD NOSE

Mjerna veličina	Izvor	Veličina x_i	Distribucija vjerojatnosti	Standardna nesigurnost $u(x_i)$	Koeficijent osjetljivosti	Doprinos nesigurnosti $u_E(x_i)$
amplituda na analizatoru spektra	umjernica	± 2 dB	pravokutna	1,16 dB	1	1,16 dB
MS 2711 B		(± 25,9%)	f = 1,73			
antenski faktor antene	umjernica	± 1 dB	normalna	0,50 dB	1	0,50 dB
PCD 8250		(± 12,2%)	f = 2			
gušenje kabela duljine 0,84 m	umjernica	± 0,08 dB	normalna	0,04 dB	1	0,04 dB
		(± 0,9%)	f = 2			
gušenje kabela duljine 5 m	umjernica	± 0,08 dB	normalna	0,04 dB	1	0,04 dB
		(± 0,9%)	f = 2			
doprinos utjecaja koje nije moguće kvantificirati	procjena	± 1,7 dB	normalna	0,98 dB	1	0,98 dB
		(± 21,6%)	f = 1,73			
$u_c(E)=\sqrt{\sum(u_E(x_i))^2}$ =						1,60 dB
						-20,20%

IV. REZULTATI ANALIZE MJERNIH IZVJEŠĆA

Analizirana su mjerna izvješća u okolini 1069 izvora GSM sustava, 565 izvora UMTS sustava i 144 izvora DCS sustava. Potrebno je naglasiti da su na određenom broju izvora obavljena periodička mjerenja nekoliko puta.

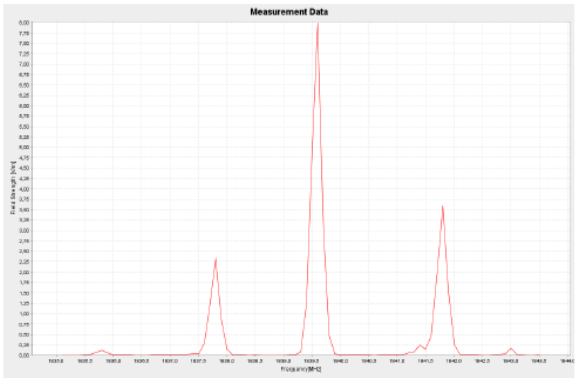
Izvješća su obrađena na način da su pregledom izvješća za pojedinu lokaciju (baznu postaju) izdvojeni rezultati najvećih izmjerenih vrijednosti jakosti elektromagnetskog polja u pojedinim točkama u području povećane osjetljivosti odnosno profesionalne izloženosti. Za potrebe obrade, razmatrani su rezultati s uračunatim mjernim nesigurnostima pojedinog izvođača mjerenja na kojeg se rezultati odnose. Također, obrada rezultata obavlja se na način da se analizira dobivena vrijednost jakosti elektromagnetskog polja za svaku pojedinačnu frekvenciju u području pojedinog sustava (GSM, DCS, UMTS) usporedbom s dozvoljenim vrijednostima za danu frekvenciju prema Pravilniku o zaštiti od elektromagnetskih polja. Također, ukoliko se mjerenja obavljaju u dalekoj zoni, što je u većini slučajeva ispunjeno, računa se gustoća toka snage ekvivalentnog ravnog vala i dobivena vrijednost također uspoređuje s dozvoljenom za dano frekvencijsko područje.

U tablici III. dan je prikaz izmjerenih vrijednosti jakosti polja pojedinih sustava Hrvatskog Telekoma. Vidljivo je da se izmjerene razine kreću u velikom rasponu, što je i razumljivo s obzirom da su obavljena mjerenja sustava različitih snaga, na različitim visinama i udaljenostima. Važno je istaknuti da su u svim slučajevima izmjerene vrijednosti jakosti polja unutar dozvoljenih vrijednosti.

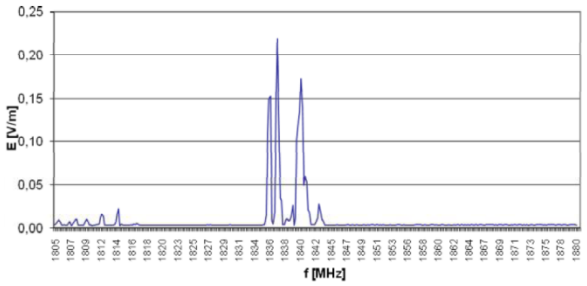
TABLICA III. PRIMJER IZMJERENIH JAKOSTI ELEKTROMAGNETSKIH POLJA GSM, DCS I UMTS SUSTAVA

Sustav	Područje izloženosti	Min. jakost polja (V/m)	Maks. jakost polja (V/m)	Dozvoljena vrijednost (V/m)	Postotak dozvoljene vrijednosti za maks. jakost polja (%)
GSM sustav	Povećana osjetljivost	0,001	11,02	16,9	65,2
	Prof. izloženost	0,005	17,20	41,25	41,7
DCS sustav	Povećana osjetljivost	0,002	7,98	23,6	33,8
	Prof. izloženost	0,041	29,98	59,0	50,8
UMTS sustav	Povećana osjetljivost	0,001	3,89	24,4	15,9
	Prof. izloženost	0,002	7,70	61,0	12,6

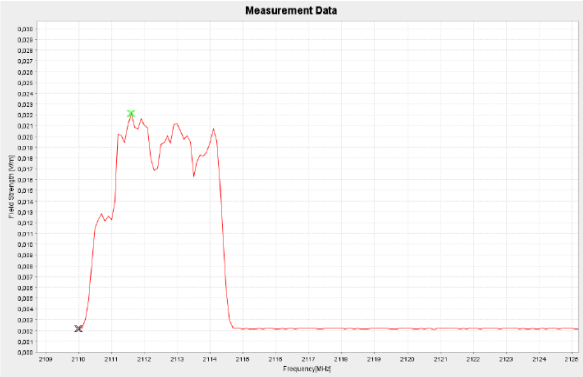
Na slikama 1. do 3. dan je primjer grafičkog prikaza rezultata mjerenja jakosti elektromagnetskih polja GSM, DCS i UMTS sustava.



Slika 1. Grafički prikaz mjerenja jakosti elektromagnetskih polja GSM sustava

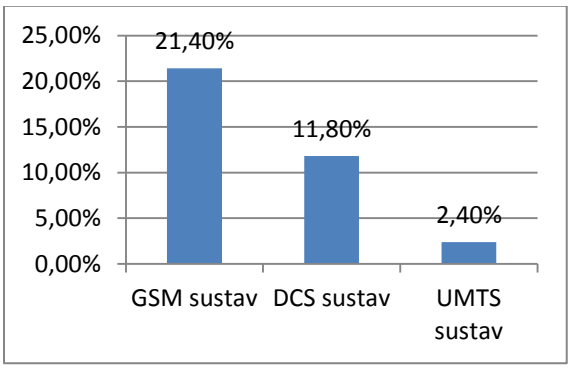


Slika 2. Grafički prikaz mjerenja jakosti elektromagnetskih polja DCS sustava



Slika 3. Grafički prikaz mjerenja jakosti elektromagnetskih polja UMTS sustava

Nakon obavljene analize rezultata mjerenja identificirani su značajni izvori elektromagnetskih polja po pojedinim sustavima te je na slici 4. dan postotak svih značajnih izvora po sustavima, ovisno o području povećane osjetljivosti ili području profesionalne izloženosti.



Slika 4. Postotak značajnih izvora po sustavima

Iz slike 4. vidljivo je da postoji određeni broj izvora gdje izmjerene vrijednosti jakosti elektromagnetskih polja prelaze iznos od 10% dozvoljene vrijednosti, ali također i znatno veći broj gdje su izmjerene vrijednosti manje od 10% dozvoljenih.

Na temelju analize je vidljivo da je postotak značajnih izvora mali, pogotovo u UMTS i DCS tehnologiji, čak i uz strogo postavljenu granicu za značajne izvore.

Obavljena je i analiza rezultata mjerenja koja je imala za cilj kategorizirati izvore sukladno njihovim tehničkim značajkama. Prema veličini područja koje sustavi Hrvatskog Telekoma pokrivaju, razlikuje se nekoliko vrsta ćelija: makroćelije, mikroćelije, pikoćelije. Za makroćelije je karakteristično da raspon emitirane izlazne snage iznosi od nekoliko desetaka vata do nekoliko stotina vata, te da pokrivaju područja od nekoliko stotina metara do par desetaka kilometara. Mikroćelijama se nastoji ostvariti pokrivenost signalom u gušće naseljenim urbanim područjima na udaljenostima do par stotina metara, s izlaznom snagom od nekoliko vata. Pikoćelije osiguravaju pokrivenost unutar poslovnih i trgovačkih centara te industrijskih postrojenja gdje je slabija pokrivenost signalom, također sa snagom od nekoliko vata. U tablicama IV., V., VI. dana je analiza udjela izmjerenih visokofrekvencijskih izvora po sustavima prema opsegu izračene snage izvora u vatima. Postotci u drugoj koloni u navedenim tablicama pokazuju koliko je baznih postaja izmjereno od ukupnog broja baznih postaja u danom opsegu izračene snage (provjera relevantnosti uzorka). Ova analiza se, s obzirom na izračenu snagu, može poistovjetiti s analizom koja bi se obavila prema vrsti ćelije jer je definirano da određena vrsta ćelije pokriva određeni raspon snage zračenja.

TABLICA IV. ANALIZA UDJELA IZMJERENIH BAZNIH POSTAJA I ZNAČAJNIH IZVORA GSM SUSTAVA U UKUPNOM BROJU IZVORA GSM SUSTAVA

Opseg izračene snage - ERP [W]	Udio izmjerenih baznih postaja sustava GSM u ukupnom broju GSM postaja	Udio značajnih izvora baznih postaja GSM sustava u ukupnom broju izmjerenih postaja
< 10	65,4%	38,2%
10 - 100	67,0%	54,1%
100 - 350	43,4%	24,5%
350 - 500	47,6%	16,0%
> 500	44,3%	9,6%

TABLICA V. ANALIZA UDJELA IZMJERENIH BAZNIH POSTAJA I ZNAČAJNIH IZVORA DCS SUSTAVA U UKUPNOM BROJU IZVORA DCS SUSTAVA

Opseg izračene snage - ERP [W]	Udio izmjerenih baznih postaja sustava DCS u ukupnom broju DCS postaja	Udio značajnih izvora baznih postaja DCS sustava u ukupnom broju izmjerenih postaja
< 10	0%	0%
10 - 100	50,0%	0%
100 - 350	19,3%	18,2%
350 - 500	29,7%	0,1%
> 500	28,9%	11,9%

TABLICA VI. ANALIZA UDJELA IZMJERENIH BAZNIH POSTAJA I ZNAČAJNIH IZVORA UMTS SUSTAVA U UKUPNOM BROJU IZVORA UMTS SUSTAVA

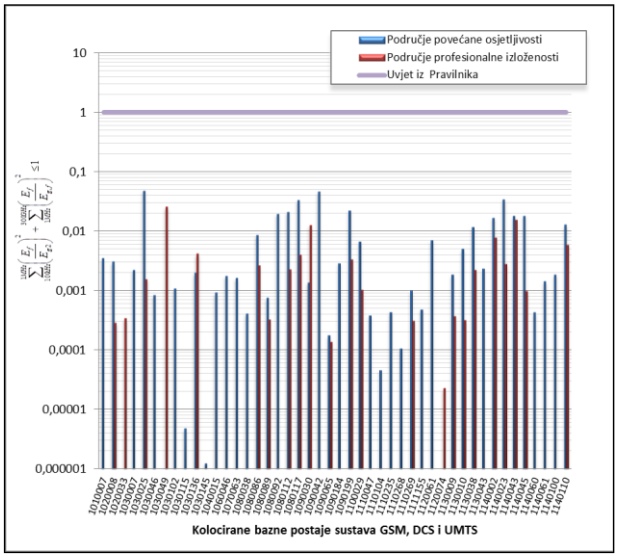
Opseg izračene snage - ERP [W]	Udio izmjerenih baznih postaja sustava UMTS u ukupnom broju UMTS postaja	Udio značajnih izvora baznih postaja UMTS sustava u ukupnom broju izmjerenih postaja
< 10	51,3%	0%
10 - 100	42,8%	0%
100 - 350	50,4%	1,5%
350 - 500	48,8%	2,8%
> 500	36,1%	0%

Statistički podaci dani u tablicama IV., V., VI. pokazuju da je analiziran relevantan broj mjerenja po svim sustavima i da se na temelju obavljene analize mogu dati zaključci primjenjivi na ukupan broj izvora u vlasništvu Hrvatskog Telekoma.

Analiza rezultata mjerenja pokazala je da izmjerene razine elektromagnetskog polja u svim mjernim točkama udovoljavaju graničnim razinama za područja povećane osjetljivosti i područja profesionalne izloženosti prema zahtjevima Pravilnika o zaštiti od elektromagnetskih polja.

S obzirom da su gotovo na svim lokacijama Hrvatskog Telekoma postavljeni kolocirani izvori, obavljena je i analiza zajedničkog učinka GSM, DCS i UMTS sustava na 92 lokacije, zajedničkog učinka GSM i UMTS sustava na 420 lokacija, te zajedničkog učinka GSM i DCS sustava na 93 lokacije. Svi rezultati pokazuju da je uvjet Pravilnika o zaštiti od elektromagnetskih polja vezan za zajednički učinak svih kombinacija tri sustava Hrvatskog telekoma ispunjen na svim izmjerenim lokacijama.

Na slici 5. dan je primjer izračuna za nekoliko kolociranih baznih postaja GSM, DCS i UMTS sustava. Ordinata pokazuje vrijednost izračuna uvjeta Pravilnika o zaštiti od elektromagnetskog polja, a na apscisi su dane identifikacijske oznake baznih postaja.



Slika 5. Grafički prikaz rezultata kolociranih sustava GSM, DCS i UMTS sustava na pojedinim baznim postajama

Iz slike 5. je vidljivo da su izračunate vrijednosti značajno manje od vrijednosti dozvoljene prema zahtjevima Pravilnika o zaštiti od elektromagnetskih polja.

Kolociranost s ostalim tehnologijama koje se nalaze u istom frekevencijskom spektru nije promatrana.

V. ZAKLJUČAK

Pravilnik o zaštiti od elektromagnetskih polja, koji se primjenjuje u Republici Hrvatskoj, postavlja granice referentnih veličina za frekvencijska područja u kojima emitiraju izvori elektromagnetskih polja u mobilnim komunikacijama (400-2000 MHz, 2-300 GHz). S obzirom na vrlo stroge granice, koje su određene visokim faktorom umanjenja u odnosu na preporuku Međunarodne komisije za zaštitu od neionizirajućeg zračenja, u radu se analiziraju rezultati mjerenja referentnih veličina u okolini izvora elektromagnetskog polja u mobilnim mrežama Hrvatskog Telekoma za GSM, DCS i UMTS tehnologiju. Faktor za koji su postrožene razine iznosi 2,5 za jakost električnog polja, jakost magnetskog polja i gustoću magnetskog toka, dok je za gustoću snage taj odnos čak 6,25.

Nakon analize 1069 izvora GSM sustava, 565 izvora UMTS sustava i 144 izvora DCS sustava, utvrđeno je da su svi izmjereni rezultati unutar granica propisanih Pravilnikom o zaštiti od elektromagnetskih polja.

Na lokacijama na kojima dolazi do istovremenog rada više izvora elektromagnetskih polja različitih frekvencija izvršena je analiza zadovoljenja uvjeta iz Pravilnika o zaštiti od elektromagnetskih polja (njihov zajednički učinak). Identificirane su 92 lokacije na kojima zajednički djeluju GSM, DCS i UMTS izvori, 93 lokacije na kojima zajednički djeluju GSM i DCS izvori te 420 lokacija na kojima zajednički djeluju GSM i UMTS izvori.

Analizom je utvrđeno da sve lokacije udovoljavaju zahtjevima članka 13. Pravilnika o zaštiti od elektromagnetskih polja koji se odnosi na zajednički učinak više izvora koji istodobno rade.

LITERATURA

[1] Measuring the Information Society, http://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-ICTOI-2012-SUM-PDF-E.pdf, ITU, 2012

[2] Pravilnik o zaštiti od elektromagnetskih polja, NN 98/11, 2011

[3] Guidelines for Limiting Exposure to Time-Varying Electric, Magnetic, and Electromagnetic Fields (up to 300 GHz). Health Physics 74 (4): 494-522; 1998

[4] Zakon o zaštiti od neionizirajućeg zračenja, NN 91/10, 2010

[5] http://www.who.int/docstore/peh-emf/EMFStandards/who-0102/Worldmap5.htm

[6] HRN EN 61156:2001, Hrvatski zavod za norme

Kolocirani izvori visokofrekvencijskih elektromagnetskih zračenja

B. Tanatarec*
* Centar za mjerenja u okolišu d.o.o., Zagreb, Hrvatska
btanatarec@cmo.hr

Sažetak - Izvori visokofrekvencijskih elektromagnetskih zračenja često puta mogu biti kolocirani. Kolocirani izvori uključuju dva ili više izvora visokofrekvencijskih elektromagnetskih polja unutar određenog frekvencijskog spektra. Kolocirani izvori visokofrekvencijskih elektromagnetskih polja iziskuju veće promišljanje o izloženosti ljudi utjecaju elektromagnetskim poljima u odnosu na one izvore koji nisu kolocirani. Pregledan je uzorak izmjerenih elektromagnetskih polja u blizini antena mobilne telefonije i ustanovljena je raspodjela tih izvora prema vrsti izvora, mikrolokaciji izvora i činjenici je li izvor kolociran s drugim izvorima ili nije. Iz uzorka su izdvojeni oni izvori koji prelaze granične vrijednosti na snazi u Republici Hrvatskoj.

I. UVOD

Elektromagnetski val se rasprostire u vremenskoj i prostornoj domeni, a sastoji se od električne i magnetske komponente. Čovjek je oduvijek izložen elektromagnetskim poljima prirodnih izvora - sunčevom zračenju, magnetskom polju Zemlje, te poljima nastalim uslijed atmosferskog pražnjenja. Uz navedene prirodne izvore zračenja u čovjekovom okruženju danas djeluju i umjetno stvorena polja koja mnogo većim intenzitetom i drugačijom spektralnom slikom bitno odudaraju od prirodnog elektromagnetskog okružja [1]. Ovisno o njegovoj frekvenciji, elektromagnetsko zračenje se dijeli na neionizirajuće i ionizirajuće. Obzirom da je u ovome radu prvenstveno riječ o visokofrekvencijskom elektromagnetskom zračenju, frekvencijskog područja od 100 kHz do 6 GHz, zapravo se radi o neionizirajućem zračenju. Navedeni dio spektra koristi se za tradicionalne komunikacijske sustave, zatim za primjenu u bežičnim, mobilnim, personalnim, satelitskim komunikacijama, zatim za razne industrijske, znanstveno-istraživačke i medicinske primjene, te za radar, mobilni radio i znanstvene primjene [2]. Zagađenje iz tih izvora sastavni je dio elektrosmoga, sveukupnog zagađenja nastalog od izvora elekromagnetskih zračenja u nekom području. Njihov utjecaj na zdravlje ljudi iskazuje se u nekoliko oblika. Kumulativan je u odnosu na izvore odnosno štetnost se zbraja iz svih izvora. Također, kumulativan je u odnosu na frekvencije odnosno štetnost se zbraja po svim frekvencijama. Kumulativan je u odnosu na period izloženosti elektromagnetskom zračenju [3].

II. ZAKONSKA REGULATIVA

U Republici Hrvatskoj na snazi je *Zakon o zaštiti od neionizirajućih zračenja* (Narodne novine 91/10) [4] na osnovu kojeg je donesen *Pravilnik o zaštiti od elektromagnetskog polja* (Narodne novine 98/11) [5] Ministarstva zdravlja Republike Hrvatske. Pravilnik [5] u smislu zaštite ljudskog bića od elektromagnetskog zračenja razlikuje područje povećane osjetljivosti (PoO) i profesionalne izloženosti (PrI). Područja povećane osjetljivosti su prema Pravilniku [5] područja stambenih zona u kojima se osobe mogu zadržavati i 24 sata dnevno, škole, ustanove predškolskog odgoja, rodilišta, bolnice, smještajni turistički objekti, dječja igrališta te površine neizgrađenih parcela namijenjene prema urbanističkom planu za jednu od prethodno navedenih namjena. Područja profesionalne izloženosti su Pravilniku [5] područja radnih mjesta koja nisu u području povećane osjetljivosti i na kojima se pojedinci mogu zadržavati do 8 sati dnevno, pri čemu je kontrolirana njihova izloženost elektromagnetskim poljima. Navedeni pravilnik u Prilogu 1 donosi temeljna ograničenja, a u Prilogu 2 granične razine referentnih veličina za oba slučaja dok se u Prilogu 3 nalaze uvjeti koji moraju biti zadovoljeni u slučaju elektromagnetskih polja više frekvencija za temeljne i referentne veličine. S danom 15. svibnja 2012. na snagu je stupio *Pravilnik o posebnim uvjetima postavljanja i uporabe radijskih postaja* (Narodne novine 45/12) [6] Hrvatske agencije za poštu i elektroničke komunikacije. Prema članku 6. navedenog Pravilnika izračun i mjerenje veličina elektromagnetskog polja obavlja se prema metodama i postupcima koje Agencija objavljuje na svojim internetskim stranicama (www.hakom.hr).

Na međunarodnoj razini dane su smjernice *Guidelines for Limiting Exposure to Time-Varying Electric, Magnetic and Electromagnetic Fields (up to 300 GHz)* [7] od strane Međunarodnog povjerenstva za zaštitu od neionizirajućeg zračenja (ICNIRP, *International Commission on Non-Ionizing Radiation Protection*) u kojima su definirane granične vrijednosti za iznose vremenski promjenjivih električnih i magnetskih polja u slobodnom prostoru, zasebno za opću i radnu populaciju odnosno za područja povećane osjetljivosti i profesionalne izloženosti. Te su smjernice donesene na temelju znanstvenih istraživanja i spoznaja o štetnom učinku elektromagnetskog zračenja na zdravstveno stanje ljudi. Također, Europska unija izdala je

direktive 2004/40/EC [8], 2008/46/EC [9] i preporuku 1999/519/EC [10] koje se odnose na minimum zahtjeva za zaštitu ljudi od rizika za zdravlje i njihovu sigurnost. U Republici Hrvatskoj na snazi su europske i međunarodne norme, propisane Pravilnikom [5], a koje su prihvaćene od Hrvatskog zavoda za norme (HZN). Dakle, mjerenja te proračun i procjena parametara visokofrekvencijskih elektromagnetskih polja obavljaju se u skladu s hrvatskim normama HRN EN 50383:2005, *Osnovna norma za proračun i mjerenje jakosti elektromagnetskih polja i gustoće apsorbirane snage koja se odnosi na izloženost ljudi radijskim baznim postajama i nepokretnim krajnjim postajama u bežičnim telekomunikacijskim sustavima (110 MHz - 40 GHz)* [11] te HRN EN 61566:2001, *Mjerenje izlaganja radiofrekvencijskim elektromagnetskim poljima – Jakost polja u frekvencijskome opsegu 100 kHz do 1 GHz[12]*.

III. RASPODJELA IZVORA PREMA VRSTI I MIKROLOKACIJI

Prema veličini područja koje izvori mobilne telefonije pokrivaju, razlikuju se nekoliko vrsta ćelija (eng. cell) – makroćelija, mikroćelija, pikoćelija i femtoćelija. Makroćeliju u mreži mobilne telefonije pokriva bazna postaja karakterističnog iznosa emitirane izlazne snage od nekoliko desetaka vata. Naziv bazna postaja odnosi se na odgovarajuću antensku strukturu i pripadnu elektroničku opremu. Najveća pokrivenost signalom u okviru makroćelija, od nekoliko stotina metara do nekoliko desetaka kilometara, postiže se postavljanjem antena na samostojeće stupove, tornjeve, silose, krovove kuća i drugih građevina. Mikroćelijskim baznim postajama nastoji se ostvariti pokrivenost signalom u gušće naseljenim urbanim područjima, a koja se nalaze u sjeni u odnosu na obližnju makroćeliju. Antene se postavljaju na vanjske zidove građevina te su znatno manje od onih u makroćelijama. Mikroćelijama se ostvaruje pokrivenost na udaljenostima od nekoliko stotina metara do jednog kilometra, a izlazna snaga im je uglavnom do nekoliko vata. Makroćelijska bazna postaja se često privremeno postavlja na određena mjesta za vrijeme velikih događanja i drugih prilika, gdje se predviđa potreba dodatnih kapaciteta mobilne telefonije. Pikoćelije pokrivaju bazne postaje postavljene unutar poslovnih zgrada, trgovačkih centara, industrijskih postrojenja, gdje je pokrivenost signalom loša. Antene su vrlo ograničenog opsega, pokrivaju nešto manje od nekoliko desetaka metara i uglavnom se ugrađuju pričvršćivanjem na unutarnje zidove te su vrlo niske izlazne snage. Trenutno najmanje područje pokrivanja može se ostvariti putem femtoćelijske bazne postaje i to do deset metara udaljenosti. Koriste se u domovima ili malim poslovnim prostorima gdje je signal slab ili ga uopće nema [13].

Mikrolokacija bazne postaje određena je načinom postavljanja antena, kako bi se ostvarilo željeno pokrivanje područja koje bazne postaje opslužuju. Da bi se omogućila usporedba različito postavljenih antenskih sustava, ćelije su podijeljene u tri kategorije, i to redom na samostojeći stup, postav na građevini te unutar građevine. Pod kategorijom „unutar građevine“ uključene su bazne postaje u tunelima, podzemnim garažama, unutar industrijskih postrojenja i poslovnih zgrada i dr.

Ovisno o odabranom senzoru i mjernom instrumentu, mjerenje razina elektromagnetskog polja može se provoditi u širem ili užem pojasu frekvencija. Pri širokopojasnom mjerenju mjeri se ukupna razina svih elektromagnetskih polja koja ulaze u izabrano područje frekvencija. Na taj se način mogu mjeriti kako vršne tako i efektivne vrijednosti polja i to bez naknadnih korekcija. Kod uskopojasnih se mjerenja mjeri iznos polja na svakoj pojedinačnoj frekvenciji, a rezultati se zatim zbrajaju na prikladan način. Zbog refleksija i drugih pojava pri rasprostiranju vala, u prostoru se pojavljuju komponente vektora električnog polja sa svim polarizacijama. Pri mjerenju potrebno je uzeti u obzir sve polarizacije da bi se mogla izmjeriti rezultantna najveća razina polja koja je potrebna za procjenu štetnosti po ljudsko zdravlje. Širokopojasna sonda u pravilu je izotropna i mjeri sve tri osi polarizacije odjednom. Električni dipol je pri mjerenju polariziran u tri ortogonalna smjera polarizacije elektromagnetskog vala (X,Y,Z). Time su obuhvaćene sve komponente vektora električnog polja (***E***) [2]. Rezultantna razina jakosti električnog polja dobiva se prema (1).

$$\boldsymbol{E} = \sqrt{\left(E_X^2 + E_Y^2 + E_Z^2\right)} \qquad (1)$$

IV. PREGLED KOLOCIRANIH IZVORA BAZNIH POSTAJA

Kako nema prave zaštite od elektromagnetskog zračenja, zakonodavac koji uređuje ovo zakonsko područje propisuje kontrolu razina elektromagnetskog polja koje ne smiju prelaziti određene vrijednosti. Prema Pravilniku [5], izvor koji premašuje 10% iznosa granične razine utvrđene za frekvenciju na kojoj izvor radi, smatra se značajnim izvorom elektromagnetskog polja. Značajni izvori su, dakle oni na koje se treba obratiti izvjesna pažnja i u čijoj blizini se vrlo oprezno mogu postavljati drugi izvori elektromagnetskog zračenja. Takvi značajni izvori se nalaze uglavnom u naseljenim mjestima ili na rubovima velikih gradova kako bi se jednim antenskim sustavom pokrilo što šire područje, a s time što veći broj korisnika.

Ponegdje se koristi više izvora elektromagnetskog zračenja za potrebe mobilne telefonije postavljenih na jednome mjestu tzv. kolokacije ili u neposrednoj blizini od recimo desetak metara koji daju kumulativno zračenje. Potrebno je prema Pravilniku [5] izvršiti provjeru zadovoljavaju li kolocirane bazne postaje uvjete za istovremeni rad više izvora elektromagnetskog polja na različitim frekvencijama prema (2).

$$\sum_{i=100\,\mathrm{kHz}}^{1\,\mathrm{MHz}} \left(\frac{E_\mathrm{f}}{E_\mathrm{g2}}\right)^2 + \sum_{1\,\mathrm{MHz}}^{300\,\mathrm{GHz}} \left(\frac{E_\mathrm{f}}{E_\mathrm{g,f}}\right)^2 \le 1 \qquad (2)$$

U jednadžbi (2) je E_f efektivna vrijednost jakosti električnog polja u V/m na frekvenciji *f*, $E_{g,f}$ efektivna vrijednost jakosti granične razine električnog polja u V/m na frekvenciji *f* (Pravilnik [5] - Prilog 2, Tablica 1 i 2) i E_{g2} vrijednost jakosti granične razine električnog polja, koja za uvjete iz Pravilnika [5] - Priloga 2, Tablice 1 iznosi 87/ $f^{1/2}$ V/m, a za uvjete iz Priloga 2, Tablice 2 iznosi 17,4/ $f^{1/2}$ V/m.

U posljednje vrijeme koristi se radi smanjenja troškova gradnje antenskih stupova već postojeći antenskih stupovi drugih mobilnih operatera ili drugih organizacija koje su izgradile svoje odašiljače. Najčešće se pored kolociranja s drugim izvorima mobilne telefonije kolocira i s radijskom ili s televizijskom mrežom. Ponekad se prilikom kontrolnih mjerenja cijelog spektra visokofrekvencijskog elektromagnetskog zračenja mogu pojaviti zračenja koja ne dolaze od izvora mobilne telefonije, ali koja sama za sebe ili kumulativno s izvorima mobilne telefonije predstavljaju određeni rizik za ljudsko zdravlje. Stoga se mora posebna pažnja obratiti mjerenjima i reguliranju snage pojedinih odašiljača unutar područja gdje se javljaju takve vrste zračenja iz različitih izvora.

Pregledano je ukupno 595 izvještaja mjerenja visokofrekvencijskih elektromagnetskih izvora (baznih postaja mobilne telefonije) u Republici Hrvatskoj s ukupno 3.570 mjernih točaka. Među izvorima uočeno je ukupno 134 kolociranih izvora. Između tih izvora izdvojeni su na temelju zakona i propisa izvori koji prelaze zakonske granice i predstavljaju potencijalnu opasnost po ljudsko zdravlje. Izmjerene su vrijednosti koje ne zadovoljavaju zakone i propise u blizini 5 postaja mobilne telefonije na ukupno 8 točaka, za različite vrste izloženosti visokofrekvencijskom elektromagnetskom zračenju.

Iz tablice I vidljivo je da na osam mjernih točaka (MT) u blizini pet kolociranih baznih postaja postoje prekoračenja zakonskih uvjeta propisanih Pravilnikom [5]. U blizini bazne postaje E (zadarsko područje) nalaze se čak tri mjerne točke koje prelaze vrijednost iz uvjeta i to dvije u području povećane osjetljivosti (PoO) i jedna u području profesionalne izloženosti (PrI). U blizini bazne postaje A (ličko područje) nalazi se mjerna točka s najvećim odstupanjem od vrijednosti iz propisanog uvjeta u području povećane osjetljivosti.

Daljnjom analizom rezultata mjerenja i detekcijom izvora elektromagnetskih polja u ukupnom mjerenom spektru frekvencija koji prelaze dopuštenu vrijednost iz uvjeta, pokazalo se da nedopuštene visoke vrijednosti dobivene mjerenjima na licu mjesta, a koje utječu na ukupan kumulativni rezultat cijelog spektra visokofrekvencijskog elektromagnetskog polja, uzročno ne potječu iz izvora u frekvencijskom području mobilne telefonije već su u području radiofrekvencijskog spektra za radio.

U tablici II prikazane su izmjerene vrijednosti jakosti električnog polja na pojedinim mjernim točkama u blizini kolociranih baznih postaja navedenih u tablici I. Navedene vrijednosti su veće od dopuštenih vrijednosti graničnih

TABLICA I. MJERNE TOČKE (MT) U BLIZINI KOLOCIRANIH BAZNIH POSTAJA KOJE PRELAZE VRIJEDNOST IZ UVJETA PROPISANOG PRAVILNIKOM [5]

MT	***Lokacija***	***Uvjet*** ≤ 1	***Područje izloženosti***
1	A	6,262367	PoO
2		1,192726	PrI
3	B	1,963900	PoO
4	C	1,309414	PrI
5	D	1,673034	PoO
6	E	2,250786	PoO
7		1,228398	PoO
8		2,034098	PrI

razina električnog polja definiranih u prilogu 2 Pravilnika [5] i to u frekvencijskom pojasu od 87,5 MHz do 108,0 MHz. Granična razina jakosti električnog polja u navedenom frekvencijskom pojasu, za najnepovoljniji slučaj ljudske izloženosti elektromagnetskom zračenju, u području povećane osjetljivosti (PoO) iznosi 11,2 V/m, a u području profesionalne izloženosti (PrI) iznosi 28 V/m. Također, u tablici II prikazan je odnos izmjerenih vrijednosti jakosti električnog polja i zakonski propisanog ograničenja kako za područje povećane osjetljivosti tako i za područje profesionalne izloženosti.

TABLICA II. DETEKCIJA FREKVENCIJSKOG POJASA UNUTAR KOJEG DOLAZI DO PREKORAČENJA VRIJEDNOSTI IZ UVJETA PROPISANE PRAVILNIKOM [5]

MT	***Frekvencijski pojas [MHz]***	***Ograničenje jakosti el. polja [V/m]***	***Izmjerena vrijednost jakosti el. polja [V/m]***	***Odnos [%]***
1	87,5 - 108,0	11,2	27,92	249,27
2		28,0	30,52	109,00
3		11,2	15,59	139,19
4		28,0	31,86	113,80
5		11,2	14,36	128,25
6		11,2	16,77	149,76
7		11,2	12,40	110,71
8		28,0	39,62	141,49

V. ZAKLJUČAK

Uslijed nedopušteno visokih kumulativnih vrijednosti jakosti električnog polja, koje su izmjerene u okolini pet baznih postaja, narušen je spektar visokofrekvencijskog elektromagnetskog polja u području radiofrekvencijskog spektra za radio u frekvencijskom pojasu od 87,5 do 108,0 MHz, a izvan frekvencijskog pojasa mobilne telefonije.

U ovakvom slučaju kolociranih izvora, koji u cjelokupnom mjerenom spektru prelaze granične vrijednosti, potrebno je smanjiti snagu izvora nakon što se utvrde izvori koji sudjeluju u tom prekoračenju. Ponekad je potrebno i izmještavanje tih izvora s istog stupa ili drugačiji teritorijalni raspored svih izvora u okolini. Kako izvori mogu biti odašiljači različitih organizacija, nakon detaljne analize svih doprinosa tom prekoračenju, regulator (u Republici Hrvatskoj je to HAKOM – Hrvatska agencija za poštu i elektroničke komunikacije) treba poduzeti pravedne i efikasne mjere u odnosu na sve sudionike i njihove doprinose u tom prekoračenju.

Iako sve izmjerene kolocirane bazne postaje sustava mobilne telefonije (GSM 900, UMTS 900, GSM 1800 i UMTS 2100 i/ili LTE) zadovoljavaju uvjet za istovremeni rad više izvora visokofrekvencijskih elektromagnetskih polja na različitim frekvencijama prema uvjetu iz Pravilnika [5], za bazne postaje kod kojih postoji kolokacija različitih izvora visokofrekvencijskog zračenja pokazala se potreba za mjerenjem šireg spektra frekvencija visokofrekvencijskog elektromagnetskog polja u svrhu zaštite od kumulativnog djelovanja visokofrekvencijskog elektromagnetskog polja na zdravlje ljudi.

LITERATURA

[1] Poljak D., *Izloženost ljudi neionizacijskom zračenju,* Kigen, Zagreb 2006.

[2] Tanatarec, B.: *Poboljšanje kvalitete mjerenja u svrhu evaluacije izloženosti ljudskog tijela visokofrekvencijskim elektromagnetskim poljima,* doktorski kvalifikacijski rad, Fakultet elektrotehnike i računarstva – Sveučilište u Zagrebu svibanj 2010.

[3] Tanatarec, B.; Nikolić, N.: *Norme u području visokofrekvencijskog elektromagnetskog zračenja u svrhu zaštite od štetnog djelovanja*, Hrvatsko društvo za zaštitu od zračenja, Krk, travanj 2011.

[4] *Zakon o zaštiti od neionizirajućih zračenja*, Narodne novine 91/10

[5] *Pravilnik o zaštiti od elektromagnetskog polja,* Narodne novine 98/11

[6] *Pravilnik o posebnim uvjetima postavljanja i uporabe radijskih postaja*, Narodne novine 45/12

[7] International Commision on Non-Ionising Radiation Protection (ICNIRP) Guidelines: *Guidelines for Limiting Exposure to Time-Varying Electric, Magnetic and Electromagnetic Fields (up to 300 GHz)*, Health Physics, Vol. 74,4 (1998) 494-522

[8] *Directive 2004/40/EC of the European Parliament and of the Council of 29 April 2004 on the minimum health and safety requirements regarding the exposure of workers to the risks arising from physical agents (electromagnetic fields),* Official Journal of the European Union L 159 of 30 April 2004

[9] *Directive 2008/46/EC of the European Parliament and of the Council of 23 April 2008 amending Directive 2004/40/EC on the minimum health and safety requirements regarding the exposure of workers to the risks arising from physical agents (electromagnetic fields),* Official Journal of the European Union L 114/88 of 16 April 2008.

[10] *European Council Recommendation of 12 July 1999 on the limitation of exposure of the general public to electromagnetic fields* (0 Hz to 300 GHz) (1999/519/EC), Official Journal of the European Communities, L 197 of 30 July 1999, pp 59-70

[11] Hrvatska norma HRN EN 50383:2005, *Osnovna norma za proračun i mjerenje jakosti elektromagnetskih polja i gustoće apsorbirane snage koja se odnosi na izloženost ljudi radijskim baznim postajama i nepokretnim krajnjim postajama u bežičnim telekomunikacijskim sustavima (110 MHz - 40 GHz)* (EN 50383:2002), Hrvatski zavod za norme, Zagreb, 2005.

[12] Hrvatska norma HRN EN 61566:2001, *Mjerenje izlaganja radiofrekvencijskim elektromagnetskim poljima – Jakost polja u frekvencijskome opsegu 100 kHz do 1 GHz*, Državni zavod za normizaciju i mjeriteljstvo, Zagreb, 2001.

[13] Tanatarec, B.: *Elektromagnetsko onečišćenje okoliša,* Udruga hrvatskih laboratorija – CROLAB, Šibenik, listopad 2012.

Analysis methodologies of assessment of impact teleworking on the decrease carbon emissions of physical mobility

Mirza Berković dipl.inž., Mr.sc. Amel Kosovac dipl.inž., Almedin Džebo dipl.inž.

Faculty of Transport and Communications, Sarajevo, Bosnia and Herzegovina
Communications Regulatory Agency, Sarajevo, Bosnia and Herzegovina
mirza.berkovic@gmail.com
dzebo.almedin@gmail.com
amelkosovac@gmail.com

Abstract -Work has been done on the subject of "Analysis methodologies of assessment of impact teleworking on the decrease carbon emissions of physical mobility." Work represents professional paper in which is independently, using several methods with general approach to the problem, was analysed of methodology for assessing of the impacts information and communication technologies on decrease carbon emissions. Today's development of standard assessment procedures takes place in two directions for products and services ICT and for organizations that use these technologies. Our analysis will be executed on the example service teleworking, like the representative ICT solution with potential significant possibility reductions of carbon emissions in the part of transportation activities. In the analysis part attentions will be aimed at possibilities definings of indicators evaluation of influence teleworking on the decrease of carbon emissions, and given recommendations about the data acquisition.

I INTRODUCTION

Information and communication tehnology can have a significant impact on creating a sustainable future through the creation of new opportunities for economic growth and welfare, reduce operating costs and minimize negative effects on the environment. When we talk about the effects on the environment, according to rough estimates, the further development of information and communication technologies can create opportunities for the substitution of physical travel through electronic communication, what can reduce for 15% in carbon emissions until 2020. year. This is known as a "business as usual - BAU" scenario [1]. Further support of this process also provide various institution that have joined efforts in finding adequate, but primarily standardized procedures for assessment (quantification, monitoring and reporting) the emission of pollutants for which is responsible sector of information and communication technologies. Secondly, it is known that ICTs enable start of the bits (information) rather than the physical launch of the atom (matter), which is the essence of the process of dematerialization. Telecommunications system, as a system for the distribution of these bits, offers an alternative for today's crucial problems of translation points of physical distance. Telecommutation, as a service, and also the subject of our analysis, is a typical example of a networked solution that we defined as the practice of performing regular work from home or from a location close to home instead of travel to work. In the literature for this kind of "telework" is used terms teleworking and e-work. Teleworking in the function of reducing the number of trips as the idea emerged of 70-years, but the growth of this type of service since then have been progressed relatively slowly. Although it is estimated that despite the rapid development of information and communication technologies, wider implementation of telecommutation will go further slowly. Reducing the number of trips to work is the fact which will bring social benefits, as a way of solving the problems of congestion and reducing carbon emissions. Transport activity as a subject of substitution with advantages that brings to society is causing huge social and environmental costs, such as accidents, health risks, pollution and emissions of greenhouse gases. Strategies which lead to less intensive transport requires consideration of potential carbon reduction refers to various measures, including a measure teleworking.

II METHODOLOGY ASSESSMENT OF IMPACT ICT ON DECREASE CARBON EMISSIONS

The methodology for assessing the impact of ICTs on the reduction of carbon emissions involves two aspects: first is related to the assessment of emissions technology which is caused directly (in the processes of production of components and systems functioning), and the other on the quantification and evaluation of the effects of emission reduction due to the application of this solution. Both effects must be taken into account to estimate the overall utility of the application. Process of the methodological assessment in both of these domains is based on the use of life cycle approaches [1]. Existing standard of the life cycle belong to family ISO standards 14040 series, which is the basis for developing specific methodologies intended for ICT.

Methodology called ICT Enablement Methodology (ICT-EM), which is briefly analyzed below, was developed under the roof of an international strategic partnership of information and communication companies and associations-GeSI [1]. It is actually the specific instructions on how to identify and evaluate the effects of application of information and communications technologies. It was created as an expression of the need to standardize procedures for quantifying the effects of

applying information and communications technologies. With a focus on simplifying assessment, through the applicable general approach, different actors of this industry, the economy, and policy makers get a practical guide to assessment procedures. ICT is based on approach a lifecycle and estimate the changes that occur in the system BAU, and which as result of the adoption of various ICT solutions. It consists of three main steps: (1) defining the objectives and scope of the research, (2) limiting / reduction evaluation (3) assessment and interpretation of the overall effects. These steps are shown in Figure 1. and briefly explained.

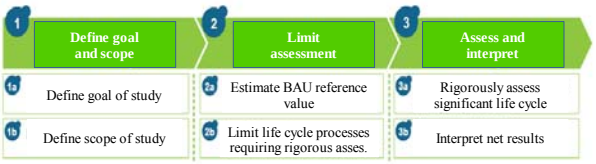


Figure 1. Steps of ICT Enablement Methodology

The first step (defining the objectives and scope of the research) requires consideration of the entire potential of the application of information and communication systems, including the following negative effects. At the same time should be taken into account all the individual components of the information and communication systems, which are necessary for its functioning.

In the first phase (1a) is defined purpose of the research, which requires the documentation of the purpose of the research (for example, whether it relates to the assessment of the environmental impacts of a product or requires to foresee potential of ICT solutions to reduce carbon emissions across the country), and users who are intended for solutions. Defining the types of users is important for assessing the wider use of a particular solution and therefore measurable effects assessment.

In the second phase (1b) is determined the scope of the research, which involves the identification of information and communication systems, BAU system and all potential positive and negative effects associated. These effects include: direct ICT emissions, primary and secondary positive effects, and primary and secondary accompanying adverse effects. Primary positive effects is occur as a direct result of the introduction of ICT solutions refer to the reduction of energy consumption and to reduce the number of trips. In contrast to them, secondary positive effects can occur after a long period of time and usually they are the result of extension of the user. Primary precede negative effects are related to increase in emissions which is usually the result of additional activities / behavior changes, which reduces the positive effects. The primary negative following effects can occur immediately as a result of implementation of ICT solutions. Secondary effects is occur later and they are often the result of cumulative impacts.

The goal of the second step (reduction estimates) is to limit the assessment of relevant effects for the components of the BAU and ICT systems allocate only for those processes within the life cycle of the product will significantly affect carbon emissions. Basically, the purpose of this step is to rule out processes for which it is estimated and it will not significantly modify the final grade.

In the first stage (2a) is determined by the so-called. BAU reference value that should be the basis for comparison of the individual contributions to the process of BAU system. This value represents change in the amount of emissions in the life cycle of that BAU process, which is assumed to be the main driver of emissions reductions. BAU reference value can be in the form of numerical values, numerical scale or order of magnitude.

In the second phase (2b) for each LC process (Lyfe cycle) is determined the expected change in emissions. As for determining the baseline, quantification of the impact is a complex problem that can be solved in different ways- using existing LCA estimates for the components of the information and communication systems solutions and BAU, or by collecting data and using the results of small pilot study. At the end of this stage, we can distinguish processes which do not have a impact, where the choice of limit values made decision maker (for example, the decision maker may determine to exclude all LC processes whose influence is less than 1% from baseline).

The last step (evaluation and interpretation of the effects) ICT-EM consists also of two phases. First we performed more accurate assessment of relevant effects and processes identified in steps 1 and 2, and then their aggregation (3a). In contrast to the previous step, the evaluation of emissions for LC processes that have been selected as relevant at this stage requires reliable data. For example, if estimate the volume level of implementation, can not be determined reliably enough, it is necessary to consider different scenarios. In the final stage we have to interpret the results in the context of the assumptions in terms of potential uncertainty (3b). Figure 2 shows the concept of the overall impact of information and communication systems on the emission which represents the cumulative effect on reduction of emission in the BAU systems and increase the information and communication system.

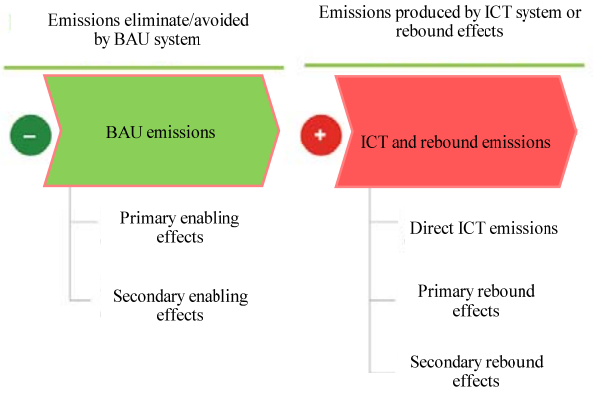


Figure 2. Calculation of net enabling effect

III ANALYSIS OF ASSESSMENT THE IMPACT TELEWORKING ON REDUCE CARBON EMISSIONS

The described methodology is a good basis for a better and more comprehensive understanding of the positive and negative effects teleworking and real possibility of its wider use.

In the first step of ICT-EM is required to define the goal as the basis for defining indicators, and in the case of teleworking it could be defined as an exploration of the possibilities in information and communication sector to

reduce emissions in the transportation sector, bearing in mind the increased energy consumption in the household. Analysis can be performed at the level of companies and employees. The scope of the research involves defining the components of the system. This system contains a PCs, printer and the necessary infrastructure such as servers and networks. BAU system components are vehicles, houses and buildings (Table I).

TABLE I. ICT AND BAU SYSTEM, AND THEIR COMPONENTS

System	Description	Components of systems	
ICT	Teleworking system	1. PCs 2. Printers 3. Data centers	4. Servers 5. Networks connections
BAU	Home to office commute	1. Private vehicles 2. Public vehicles	3. Home 4. Office building

In the next step of ICT-EM,we should consider all positive and negative effects and isolate only those who are estimated to have a significant impact. In the case of switching, that can be potential effects of teleworking and indicators in the analysis given in Table II.

TABLE II. TELEWORKING EFFECTS

Categories	Identified effects
Direct ICT emissions	Emissions from ICT equipment required for telecommuting
Primary enabling	Reduced private vehicle use
Secondary enabling	Reduced public vehicle use
	Reduced building use
	Reduced building construction
	Reduced road congestion and constructio
Primary rebound	Increased home energy use
Secondary rebound	Urban sprawl

If you choose effects that are expressed in a relatively short period of time and at a relatively small scale of implementation (at company level), the final list of processes that require evaluation are decreases. In this case value of BAU reference is determined based on the potential maximum effect that can be achieved [2]. This means that you need to calculate the level of emissions which is broadcast using passenger cars owned by employees, public transport per employee, etc., between place of housing and workplace/factory (Figure 3).

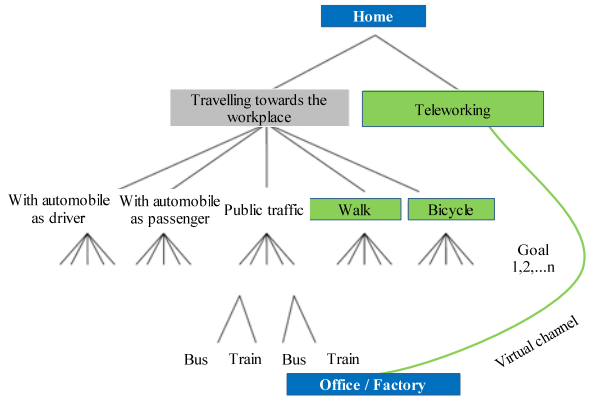


Figure 3. The structure of regional model of physical travelling and virtual channel on the relation of houses - work place

At the same they can use existing assessments and make extrapolation of data for the number of employees in the company. Changes in emissions for ICT components are considered insignificant, and the work of some components of the system like a server and computer centers does not have also influence because they work regardless of where they are employed. It is also assumed that employees already possess the necessary equipment (computers and printers) at home, and the purchase of new equipment due teleworking does not influence. In this case, as significant effects we can identified the following: reducing the number of trips, an increase in household energy consumption and reduce equipment and energy use in commercial buildings. All of this processes need to be mathematically quantified only in the phase of use (LC process is not complete) [3].

For further calculation, it is necessary to determine mean values of distance to work, then determine the percentage of employees who use passenger cars and those who use public transportation or use existing emission factors for the final showdown. Realized effect (A) is expressed in quantity (kt) of CO_2 emissions. The same units are expressed in the emission of CO_2 which is reduced due to energy savings in commercial buildings / offices (B), as well as emissions resulting from the additional energy consumption in households (C). Total effect (A + B - C) gives a quantitative value and the answer to the question of the potential usefulness teleworking on reducing carbon emissions.

Thus, for quantify the total effects of introducing teleworking in business systems is necessary, based on the life cycle, execute evaluate environmental impacts, how its constituent parts of the system and service teleworking, so and reference business system.

For the analysis effects of a potential reduction of environmental impact by introducing service teleworking, ITU recommendations include two cases of comparative analysis and to [4]:

First case: comparison of the reference production system and service teleworking. In this case, there are assessment of environmental impacts or BAU reference ("business as usual") production systems and systems teleworking who replaces him.

Second case: comparison between two services or products. The goal is to identify the differences between the network, services and products teleworking, usually between old and new technologies.

IV CHARACTERISTICS OF CARBON INFLUENCE OF TRANSPORTATION ACTIVITY THE JOURNEY TO WORK

The effects of transport activity can be largely classified into to direct effects resulting from the current transport activity, with the reason and effect of relationship of these effects are usually clear and well understood. Then indirect impacts (secondary and tertiary effects of transport activities), which have more serious consequences than the direct effect, whereby the their connections and relationships with the environment often incomprehensible and difficult to ascertain. The third type of influence transport activity are cumulative impacts that have additional, multiplier or synergistic effects. Under these we

understand the different direct and indirect effects on the ecosystem, and they are largely unpredictable.

Because of its importance, four variables, A (total transport activity with the vehicle in km), S (participation of transport type in the way), I (energy intensity) and F (content of CO_2 in fuels), they are called indicators of transport activities on the travel to job, while the variable which are consists of these indicators are called factors.

General set of indicators for the carbon impact of transport activities of travel to work can be a tool that will contribute to the identification of potential applications for carbon reduction activities. Carbon indicators can be used in the ex-ante evaluation, for example, to plan future actions, including research and development. After the introduction of transport measures, policies or programs, indicators can be used for their evaluation. Indicators for the carbon impact can be used for monitoring / control targets for the reduction of carbon, to construct forecasting models or field comparisons for development of low-carbon transport.

For determine the savings in CO_2 emissions transport activities by using service teleworking not only significant estimated CO_2 emissions after the measurement, but also the "business as usual" (BAU) or baseline projections on projected emissions without measuring. For the budget savings in the program is adequately to use the ASIF treatment, as this method provides a "quick and easy" way of quantifying carbon emissions [5]. It is a bottom-up procedure for the calculation of CO_2 emissions in any form of transportation (private car, public transport). CO_2 emissions are displayed as a product of four factors: activity (A), division (S), intensity (I) and fuel (F), which are essential for evaluating the transport sector and the possible impact. Formula for CO_2 transport system is [5]:

$$CO_2 = \sum_{i=1}^{n} G_i = \sum_{i=1}^{n} A_i S_i I_i \sum_{j=1}^{k_j} \lambda_j F_{ij} \qquad (1)$$

where are:

CO_2 - total emissions of all types of transport used by the user on the way from home to work;

G_i - the CO_2 emissions of the type of transport used by the beneficiary;

n - the number of transportation types;

k_j - the number of types of fuels in the type of transportation;

A_i - is the total transport activity in vehicle-km by the type of transport;

λ_j - is the participation of k_i fuel in the type of transportation (%)

j - participation of k_i fuel in the type of transmission (%);

S_i - participation of the types of transport and;

I_i - the energy intensity of the transport type and (MJ/km)

F_{ij} - is CO_2 emissions from fuels j her type of transport and (g / MJ).

According ASIF model, emission savings of transport activity travel to work m is calculated as [5]:

$$\Delta CO_2 = A_{BAU} * S_{BAU} * I_{BAU} * F_{BAU} - A_m * S_m * I_m * F_m \quad (2)$$

A_{BAU} represent transport activity under the BAU scenario and A_m is transport activity if is measures undertaken (by analogy with the other variable). Savings of CO_2 from the impact on the transport activity after t years is calculated as the difference between the annual basic values of CO_2 emissions and annual emissions (project value) of CO_2 emissions in a case which it was undertaken as realization of teleworking in a particular area. Alternatively, the value of emission savings for some kind of transport *i* can be determined on the basis of the formula [5]:

$$\Delta C0_2 = t * (CO_{2\,old} - CO_{2\,new}) - D = \sum_{i=1}^{n} \Delta G_i - D \quad (3)$$

For each of the components A (activity), S (separation of type), I (energy intensity) and F (fuel), a single set of factors is needed to determine the impact of the carbon transport measures. Additionally, D as a value of constant emissions from transport measures we are also taken into account. When combining various ASIF measure, emission savings can not be easily determined due to a possible increase and adverse effects of any measure which it has on other measures. This is because of the interaction of elasticity of the measures types with all other measures what should be taken into account with the inclusion of effects, such as unwanted and induced transport effects.

The transport sector is a very complex system and small changes in one area can have a big effect on the whole system. In this way, individual transport measures can be evaluated with the exception of this interaction. If we take the measure to reduce CO_2 emissions, there are always side effects that can affect the result of this measurement. This can be done in the same procedure as the original effect and on that way increases the impact (they are called multiplied effects) or doing the reverse procedure diluted the original effect (they are called unwanted effects).

An example of unwanted effects is "induced traffic" that can be viewed with infrastructure measures to increase road capacity and to reduce congestion - because of better road conditions, the people next to the option teleworking, want to drive more, taking the car increases, the traffic problems is induced and congestion is not resolved. If several measures are undertaken at the same time, estimates about the overall impact on emissions become very difficult or even impossible. Interaction between the different measures are very general and sign of these interactions can be positive or negative.

Example for multiplication effect is the interaction between the measure of public transport infrastructure and price measures for individual, motorized transport, which can lead to multiplication of these effects opposite to collecting. Another example can be the impact of measures to increase urban densities, which will reduce transport demand.

V ANALYSIS OF INDICATORS EVALUATION OF CONTRIBUTIONS TELEWORKING

Indicators for evaluation of contributions teleworking can be developed at the international, national or regional, and local level. Monitoring of indicators assessment is the most appropriate at the local level (or project level). Monitoring at this level can bring real benefits by facilitating decision making based on real data. Since one of the objectives of the methodology is based on indicators of comparability of indicators at different levels, the indicators should be available to all interested parties. Indicators, which are derived from the data, are the basic tool for the analysis of changes and trends [6]. They are a better analytical tool than the data for several reasons. First, they can serve as the basis for evaluation by providing information on the status and trends of development solutions. Second, as the basis of these assessments, the indicators provide input for policy formulation teleworking. Third, presenting more data with one number which is easier to interpret than complex statistics, give the possibility to facilitate communication between the different groups.

A. Goals of services

The goals of teleworking include the application description of what should be achieved in order to attain the desired look of the new state programs and goals as a function of the emission reductions in the transport sector (reducing the number of trips, reduction kilometers per vehicle, reducing transport emissions, etc.). Objectives should be defined for each individual problem, which is marked as important in the projection of future conditions. For defined goals of teleworking, we determine the possible indicators to measure performance for sustainability solutions.

B. Developing and defining a set of indicators of the impact to service reduction of carbon emissions of transport operations

In accordance with established goals to reduce emissions by using solutions teleworking, we can develop and define the indicators of performance solutions. These indicators should be related to performance that indicate the degree of compliance with the established emission reduction targets. Development and definition of the indicators should be made through the following steps [6]:

- defining criteria for the selection of evaluation indicators,
- identification of potential indicators,
- evaluation of potential indicators,
- final selection of indicators.

1. Defining criteria for the selection of evaluation indicators

Criteria for the selection of evaluation indicators are defined in accordance with the circumstances and the priorities identified in the strategies to reduce carbon emissions. Criteria for the selection of indicators that have been proposed are:

- scientific / theoretical acceptance indicator assessment,
- representativeness of indicators for specific problem teleworking,
- the ability of indicators to explain the connection between cause and effect,
- the ability to change the indicator to show the value of performance,
- the importance of user-friendliness and indicators for the service user,
- comparability of indicators with objectives, standards and tolerance limits,
- Compatibility with indicators of other solutions,
- the adequacy of the indicators in relation to the size of the area of application solutions teleworking,
- the adequacy of the indicators in relation to the accuracy, type, volume and time data availability,
- cost eligibility indicators.

2. Identifying potential indicators of assessment

In accordance with one or more of the adopted criteria we shoud identify all set of potential indicators of impact assessment teleworking on reducing emissions. We can come to the set of potential indicators with complete research literature available that can be applied in the field of application teleworking on reducing carbon emissions of transport operations. Examples of possible indicators teleworking are shown in Table III.

TABLICA III. INDICATORS OF TELEWORKING

	Indicators of travel reduction
Decrease of travel	kilometres per teleworking day
	kilometres per teleworking week
	numbers of trip by individual
	vehicle-kilometres
Type of vehicle	kilometres per week for car users
	kilometres per week for rail users
Distance (km)	short, medium i long distance
	Indicators of assessment
The classification of time	full-time teleworks
	full and office out teleworks time
	part-time teleworks
	"ad hoc" teleworks

3. Evaluation of potential indicators

For each of the identified potential indicators, it is necessary to perform the evaluation in relation to the criteria adopted for the selection of indicators. Evaluation should also be carried out in accordance with the relevance of performance to which the indicator is related to the defined goals.

4. Final selection of indicators

On the basis of the type of evaluation of potential indicators we made the final selection of indicators. Final selection of indicators leads to a set indicators whose values will be monitored and measured in the context of reducing emissions in a particular area under consideration.

C. Data acquisition and processing

At this stage it is necessary to collect, process and interpret the data so that it allows the identification of effects and spotting trends of application teleworking Data collection can be done in two ways:

a) passive collection - collection of existing data from other database sources

b) active collection - collection of new data, which is related to the generation of data necessary for determining the value of the adopted indicators, which can not be obtained from the databases of other sources.

Collection and processing of data usually includes the following stages:

1. accurate identification of specific needs for data on which it is based reporting,
2. identifying reliable sources from which it is possible to get the data passively,
3. determining the data that is necessary to provide active collection,
4. establishing a process of passive and active data collection
5. establishing a procedure for the treatment of collected data.

It is recommended, that at this stage are documented procedures of data collection and processing, with defined responsibilities, scope, frequency and other relevant factors relating to the collection and processing of data (Table IV) [7]. Rating of carbon efficiency measures of teleworking and estimates critically carbon footprint depend of the collected data. Obtaining data is particularly difficult for the transport sector, as it was composed of many small and mobile units. If the development of indicators for sustainability solutions is used for comparisons across countries or communities, or in the period of the development of a spatial unit, must be provided comparable data. Harmonization of statistical standards is extremely important to assess the state of the system, especially in terms of sustainability and carbon emissions. The final version of the report shows a tabular view of data quality for the relevant indicators. Data quality is significantly changed by the indicators because of the importance and the fact that they are not required by all countries simultaneously.

TABLE IV. DATA - MOST WHO TELEWORKING DO SO DAILY

	Total	Germany	Italy	France	U.K.	Spain	U.S.
Sample size	1021	205	158	120	164	173	201
Every day	31%	30%	25%	39%	34%	27%	31%
Two to four days per week	24%	28%	25%	18%	23%	25%	24%
One day per week	18%	14%	21%	18%	18%	22%	18%
One day every two weeks	6%	7%	7%	9%	5%	4%	6%
One day per month	5%	5%	7%	3%	5%	6%	5%
Less than one day per month	15%	15%	15%	14%	15%	15%	14%

D. Analysis of research results

After processing the "raw" data, we should analyzing them in terms of value and trend of observed indicators. The results of the analysis must be sufficiently representative to allow discussion and comparability with previous periods. In order to representativeness of data should take into account the following principles:

- always start from the same definitions, ie., units of measure,
- use the same method in determining the value of indicators,
- monitor the performance of the indicators in the same periods of time,
- use the same procedure for the budget, as the site, and control methods.

Particular attention should be paid to possible trends of the observed indicators in the case of the existence of data over a longer period of past time. That will allow you to understand what is happening and what can be expected in any of the coming periods. You might discover connection between the first glance unrelated factor group.

VI CONCLUSION

The results of analysis and research of parameters teleworking which define combination of the characteristics of this system sufficient to determine positive impact of teleworking on reducing carbon emissions of physical commuting, it can be concluded that the current assessment methodology as a set of parameters or values derived on the basis of the positive effects of teleworking describes partially condition, and does not include the effects for which are wholly beyond the parameters of the carbon impact to transport operations. The analysis showed that in addition to the absence of a set of relevant indicators in the case teleworking that take into account such as the analysis of national mobility, the additional problem of a lack of models is created by collecting data on the effects teleworking. The indicators proposed in the literature can only be an initial impact assessment indicators teleworking of the reduce emissions of these activities. For a more comprehensive and detailed approach to this issue is necessary to have wider access to research and data from terrain. Also, the paper presents two approaches that represent domains with a good starting point, but whose application is still far from routine. Complexity assessment procedures as well as the required level of information bases make the commitment of a regular assessment of environmental impacts teleworking still only part of the individual projects. In order to maintain overall positive environmental impact teleworking on reducing carbon emissions physical mobility travel to work it is essential that there is an integrated and standardized procedure for its evaluation. As in any scientific work, given the framework of time and space, and here opened up a number of additional issues that need to be further processed, such as the evaluation and ranking of effects, use of appropriate methods of assessing the effects of the model and the definition of collecting data on the effects teleworking.

LITERATURA

[1] GeSI, Evaluating the carbon-reducing impacts of ICT – An assessment methodology, Global e-Sustainability Initiative Report, February 2011, Available: http://www.gesi.org/ReportsPublications/AssessmentMethodology.aspx

[2] P. Nelson, Safirova, and M. Wells (2007). "Telecommuting and environmental policy: Lessons from the ecommute program. Transportation Research D 12, pp. 195-207.

[3] N. Bojkovic i M. Petrovic, „Mogućnosti evaluacije doprinosa Informaciono – komunikacionih tehnologija smanjenju karbonskih emisija," Zbornik radova „Postel 2011", Beograd, pp. 78-86, 2011.

[4] Recomendation ITU-T L.1410, Methodology for the assessment of the environmental impact of information and communication technology goods, networks and services; ITU, March, 2012

[5] REACT: Izveštaj o razvoju opšteg skupa indikatora za karbonski uticaj, Završni izvod iz Izveštaja o razvoju opšteg skupa indikatora za karbonski uticaj, D.3.3 iz EU FP7 REACT projekta Od Dr. Susanne Böhler-Baedeker (project coordination), Thorsten Koska, Robert Gruber, Carolin Schäfer-Sparenberg, Frederic Rudolph, Ulrich Jansen (Wuppertal Institute for Climate, Environment and Energy) i Lorraine Whitmarsh, Dimitrios Xenias (Schoof of Psychology, Cardiff University), U.K.

[6] Guidelines for the development of sustainability indicators, Michael Ditor, avgust 2001., http://www.ec.gc.ca/soerree/English/Scip/guidelines.cfm,

[7] Global e-Sustainability Initiative – GeSI uz podršku kompanija British Telecom, Deutsche Telekom, Ericsson i Verizon, 2010

C E

International Conference on

COMPUTERS IN EDUCATION

Steering Committee

Chairs: Marina Čičin-Šain, University of Rijeka, Croatia
Jadranka Sunde, Defence Science and Technology Organisation, Salisbury, Australia
Ivan Uroda, University of Rijeka, Croatia
Ivanka Sluganović, I. Tehnička škola Tesla, Zagreb, Croatia

Members: Michael E. Auer, Carinthia Tech Institute Villach, Austria
Ivan Gerlič, University of Ljubljana, Slovenia
Marija Marinović, University of Rijeka, Croatia
Slavomir Vukmirović, University of Rijeka, Croatia
Željka Požgaj, University of Zagreb, Croatia

INVITED PAPER

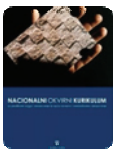
Tablet learning environment in iSchool Vežica

Violeta Nikolić*, Loredana Zima Krnelić*, Davor Kolarić* i Mihael Nikolić **
* Primary school Vežica, Rijeka, Croatia
**Primary school for classical ballet and contemporary dance, Rijeka, Croatia
os-vezica@ri.t-com.hr

Summary - In this paper the authors present the results of implementing tablets (iPads) in the learning process. The tablet (iPad) is transforming teaching and learning at all levels and completely reorganizes the educational environment. It provides interactive learning supported by the Internet and other multimedia, thousands of free educational applications which makes learning easier and more fun, real time (online) communication between teachers, pupils and parents sharing digital teaching materials as well as support of iBooks, interactive digital books that teachers can very easy create for their own subject (theme). By now we have already created five iBooks for geography. The most important is that in the tablet learning environment pupils become active learners, by practicing team work, solving problems, exploring and learning how to use knowledge instead of reproducing one, also learning necessary digital skills. This challenge based learning increases pupils' engagement, their grades and prepare them for the modern digital age and lifelong learning. The tablet learning environment motivates teachers to implement ICT in all subjects, requires constant education of teachers using new technology and leads to a new modern curriculum we all aim to.

I. INTRODUCTION

Educating young pupils is one of the most beautiful jobs in the world but it has been becoming very hard for the last few years or, should I say, decades. As primary school teachers we have been noticing that pupils are mostly not motivated enough, they are more passive listeners then active learners and they are very often not interested in knowledge acquisition at all. We cannot blame them because time has changed and the way we produce, access, visualize and share information and knowledge is very different from what it was decades ago. The amount and type of information produced, shared and dealt with every day is growing exponentially and many of the skills required in today's working environment are not taught in formal school systems. In order to form a multidisciplinary approach to teaching and learning we have to provide a learning environment based on problem solving activities, challenge and discovery.



That is what our National Curriculum, the fundamental document for writing any school curriculum, points out as well.
The National Curriculum is an official document established by the Ministry of Education, Science and Sport which determines the educational policy for public schools so, it represents a fundamental document when writing the school curriculum.

II. PROJECT ISCHOOL

This school year, we have started a new project named iSchool. The idea came from our headmistress Violeta Nikolić who gathered a project team. The team members are our teachers Loredana Zima Krnelić, Davor Kolarić and Mihael Nikolić.

We are the first primary state school in Croatia which implemented iPads in the learning process. By using the iPad technology, we design a whole new educational environment for our pupils and teachers. Our main goal is to completely reorganize the educational environment and create a new modern curriculum.

At first we started with 32 iPads and one class. At the beginning of the school year we chose the 8th form which had had the lowest school success in order to see whether their motivation and achievements would improve with the introduction of iPads. That was, in fact, what we expected to happen and that is what has been happening since then.

Two months later we bought other 25 iPads, so now every teacher in our school has got an iPad and our plan for the next school year is to introduce iPads in the three classes of fifth graders (50 pupils).
The project goals are:

- hightened school success, increasing pupils engagement, pupils as active learners prepared for lifelong learning
- improved competency for living in the modern digital era and the development of social awareness and responsibility towards new technologies
- equalizing our pupils' abilities with pupils from developed countries
- lighter school bags
- new learning process
- digital learning materials
- a new modern curriculum

The last but not the least important project objective is becoming an Apple Lighthouse School and being able to share our knowledge and experience with other schools in Croatia.
To ensure that we achieve our goals, our project is evaluated by the Department of Pedagogy and Psychology at the Faculty of Philosophy in Rijeka.

III. TABLET LEARNING ENVIRONMENT

The tablet (iPad) is transforming teaching and learning at all levels.

In what way?

First of all, it provides interactive learning with lots of multimedia and Internet access enabling pupils to became active learners. In the classroom they practice team work, solving problems, exploring and learning how to use knowledge instead of reproduce one. They also acquire necessary digital skills.

For example, pupils learn where they can find good educational web sources for every subject. They learn how to make a good research on the Internet for different topics as well as they make critical reviews of different web pages which is very important because they not only use the Internet for learning but also they learn how to distinguish the quality of web sources.

In this environment teachers also spend more time online than they did before, mostly searching digital teaching materials to share with their pupils. So the implementation of ICT in every subject is inevitable. As one of the project goals is to teach our pupils how to be responsible users of the Internet, this is for sure, one of the ways to learn that.

The App Store offers thousands of free educational applications for every subject to make learning easier and funnier for pupils. Interactivity and multimedia in applications make content alive and fun so pupils can learn through play. We assume this will make learning easier for everyone. The proof is the fact that our pupils started to heighten their school success in a very short period.

Even though the applications available in App store are in English, pupils understand most of the terminology and, if they don't, they learn new English words. One of our teachers said that it can't hurt if pupils learn English units of measurement such as feet, mile, pound and so on. This is a multidisciplinary approach of teaching in which pupils learn applicable knowledge by linking scholarly content to everyday life.

Some schools do not have a chemistry laboratory and all the necessary chemical substances with the consequence that pupils can not see what they look like or how they react. Tablet applications make it possible to do virtual experiments with all chemical substances. Pupils can do them on their own having the opportunity to experiment and repeat it as many times they want. This positively affects pupils' involvement and motivation for learning.

The implementation of tablets (iPads) is a so called 1:1 program because it personalizes learning by giving the opportunity to pupils to use tablets (iPads) as a tool to aid learning according to their knowledge and skills. The result is that they are more active in the classroom and finally they become partners with their teachers in the learning process. Some of the pupils literally woke up and thrilled the teachers.

A 1:1 program allows teachers to easily customize assignments to each student individually, especially those for children with special needs.

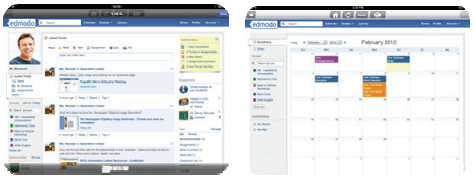
The most important advantage of this collaborative environment is that our pupils are online with their teachers and parents and can get help with their assignments any time of the day. In our school pupils who were particularly shy during English classes started sending emails to their teacher while some of them finished their homework at recess time. This surprised our English teacher who also commented that thanks to the use of tablets (iPads) she got to know her pupils much better and that helps her be more objective when evaluating them.

Every subject is specific, for example in English class pupils who are shy often tend not to communicate. Now our teacher shows them applications for exercising pronunciation at home. In that way their English as well as their achievements improved.

In this new environment except from their teachers pupils learn a lot from each other and in some cases teachers learn from their pupils. This collaboration allows teachers to change their role in the classroom and become mentors to their pupils.

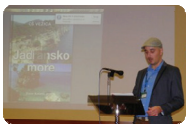
Pupils and teachers communicate by email and through Edmodo, an application very similar to Facebook. It allows pupils to blog, post comments, upload assignments, email and share links and videos.

Teachers can very easily create their own online tests and send assignments. Parents can also communicate with teachers and pupils through Edmodo and are for the first time directly involved in the learning environment.



In order to have 21st century materials /learning environment our teachers started with digitalization: In addition to common digital materials our teachers use in the classroom such as Power Point Presentations, some of our teachers create iBooks – interactive books which are much more engaging for pupils and closer to the way they want to read and learn – having audio, video and interactivity that makes content alive, not forgetting the fact that information in these books is always up to date which is important in our information society.

Below you can see what our first digital interactive books look like. Our teacher of geography, Davor Kolarić, made those iBooks so pupils don't have to buy paper books for geography.



He said: „Geography, with the use of iPad, is much more dynamic and interesting to students. In the past few months, I made a few "digital books" that are completely different from those in paper form. The advantage of these "books" is that they are interactive and provide the student a lot more information and a variety of multimedia materials in one place. No need to look for more multimedia content on CDs and DVDs, because they are available and integrated into the content. Students with a single click can visualize content on a short film, for example description of the size of the waves. Motivation of students is at a high level as well as their activity. Students explore more and are active participants in the educational process, rather than passive observers. Communication between student and teacher is much simpler and faster, and takes place at all times, not just during they stay in school. A simpler way of checking the knowledge through interactive online tests, and also make it easier to the teacher correcting, grading and making of the test statistics. Student's bags are not heavy anymore because they do not have to carry a book, a workbook and a notebook to school."

Finally pupils have started to use technology for learning and not only for playing games. They learning how to maximize the use of technology without being used by it. They become aware of what they upload, download, of the way they communicate online (net etiquette), of how to communicate on mailing lists and forums as well as they learn what cyberbulling and stolen identity is or the truth about Facebook. In other words what the traps of the virtual world they live in are. (they are part of).

Our pupils can take tablet (iPad) at home so they can choose how, when and where they learn. Using the Calendar and the Remainder they can easily organize their school duties and exercise organizational skills.

The tablet (iPad) enables teachers to be more mobile in the classroom. He doesn't have to stand in front of the class but he can sit with pupils and they can work together. This creates teacher-pupilds connections of a different level. By sharing new learning experiences pupils and teachers became more attached to each other.

IV. CONCLUSION

It is true that technology can make a profound difference in the way teachers teach and kids learn. We have succeeded to increase pupils engagement and their grades. We have succeeded to motivate teachers in implementing ICT in every subject very successful. Some of our teachers strated to create their own interactive books (iBooks). Next school year we will continue with our project in three classes of 5th graders.

At the end we can conclude that implementation of iPads in the learning process has positive effects on pupils and teachers, enables discovery, problem and challenge based learning environment which leads to a reform of the formal school system in order to make the school curriculum alive.

PAPERS

Are We Trendy?

Hannu Jaakkola*, Bostjan Brumen **, Jaak Henno *** and Jukka Mäkelä****
* Tampere University of Technology/Pori Department, Pori, Finland
** University of Maribor/FERI, Maribor, Slovenia
*** Tallinn University of Technology/Information Technology, Tallinn, Estonia
**** University of Lapland/Method Sciences, Rovaniemi, Finland
hannu.jaakkola@tut.fi; bostjan.brumen@uni-mb.si; jaak@cc.ttu.ee; jukka.makela@ulapland.fi

Abstract ï A trend is the general course or prevailing tendency, a move in a certain direction over time, the general direction ï it is an expected continuum of the current state. A trend is based on history and provides a view of the future. Understanding trends helps us to be prepared for the future ï to be proactive instead of being reactive. So ï life should be easy: just recognize the trend and be prepared for the future. There is, unfortunately, a minor problem. Sometimes the trends do not coincide with the future reality. The reasons for this are manifold: we have followed the wrong trend, the conditions behind the trend have changed, ñtrend makersò have changed the trend, etc. Regardless of uncertainty, there is not much really new that is not already recognizable today if we are talking of the near future, 5-10 years from now. We seek the answer to the question in the title by analyzing the messages of reports by the most commonly known trend analysts. The focus of our paper is on the phenomena related to the information society. The main trend enabler is technology, especially information technology. The paper discusses trends as a means for future analysis first on the level of theory and then it provides a synthesis of the some recognized trends affecting our future.

I. INTRODUCTION

A trend is defined as "*the general course or prevailing tendency*", "*a move in a certain direction over time*", "*the general direction*". It is an expected continuum of the current state - by knowing the past we are able to predict the future. *Trend based prediction* expects that everything with an effect on the (near) future is already available and that environment changes are manageable or stable; we just have to recognize what is important in the mass of information chunks surrounding us and combine / interpret these in a beneficial way.

Trend analysis is applicable in several contexts and especially used in connection with analyzing the adoption speed of certain products or changes in technology. This topic was handled extensively by the first author of this paper in his PhD Thesis "An Analysis of Diffusion of Information Technology in Finnish Industry" [1] and several conference / journal papers published thereafter (see e.g. [2; 3]). The thesis covers the basis of trend analysis from a wide-ranging perspective. It first focuses on *mathematical diffusion models* and finally introduces the concept of the "*Heuristic Diffusion Model - HDM*." The aim of a mathematical diffusion model is to find the best mathematical formula (distribution) that fits the real trend of the past. The HDM is a collection of principles aiming to increase understanding of the trend as a phenomenon. The term "*diffusion*" describes the phenomenon in the background of the trend and is applied in two contexts: (1) to describe how widely the potential user population has adopted a certain product and (2) how a new innovation is applied in products / society / industry.

Since early 1991, when the thesis was published, diffusion-related research has been quite active, on one hand in the field of marketing research and on the other hand in technology studies, and it is the latter that is the focus of the authors of this paper. In spite of the activities of this researcher community it is interesting to notice that the basis of trend analysis has remained quite stable. To pour oil on the fire, the same basic ideas that were introduced by Gabriel Tarde in his book "Les Lois de l'Imitation" [4] in 1890 are still built into the methods used by researchers. Tarde modeled the diffusion (spread) of political ideologies among the population and used a mathematical formula to explain the changes and to predict future changes. Since then, large numbers of new models - variations of the original idea - have been published and applied in new contexts, and the work is still continuing. In spite of the fact that the first author of this paper has not been active in this field of research in more than a decade, understanding the basics of the diffusion phenomenon has proved its usefulness in analyzing and understanding the changes in the information society, which has also been one of the topics of earlier joint publications by the authors of this paper (see e.g. [5; 6; 7]). The authors come from different cultures and from different scientific traditions, which provides an excellent opportunity to combine a variety of viewpoints in information society analysis.

The purpose of this paper is first to give a short introduction (theoretical view) to trend-based analysis in the context of products and technology; however, the main focus is on the latter. The aim is not to go into detail on the modeling itself, but to give a basis for understanding trend-based behavior, including the opportunities and threats it presents. This framework will be applied in introducing a synthesis of the future trends most to be expected which will have an effect in the near future. The synthesis is based on publicly available data from selected market research companies that publish regular technology trend-related reports.

The rest of the paper is organized in the following way. Section II is focused on the basics of trend analysis. Section III introduces the findings of the selected technology trend analysis and opens a discussion on them. Section IV concludes the paper.

II. UNDERSTANDING TRENDS

A. Terminology

The *diffusion process* is a phenomenon in which an innovation spreads in a social system (target population) over time using certain channels [8, p. 5]. *The diffusion model* describes the spread of an innovation (which may be a product, a service, or an ideology) in the potential population according to the selected dimension (which in most cases is time). As an example, we can analyze the spread of the use of tablets among all mobile terminal users. As a result, we get the *penetration rate*, e.g. 57 % of all mobile terminal users use a tablet. The term *potential population* may be generalized and replaced by the term "*capacity or capability*" or just "*potential*". If the diffusion process is discussed in the context of *technology innovations*, the potential is seen as the *innovation (renewal) capacity* available when this technology is used; this capacity is "consumed" in time until the innovation potential of this certain technology is finished – nothing new can be expected anymore. An example: In the context of mobile terminals we can analyze the spread of 3G technology among its potential user population and get a curve describing the diffusion of this technology; as a result we see that e.g. 3G technology has an 88% penetration level in mobile terminals used. As mentioned above, the difference between the *diffusion of a product* and *diffusion of technology* is only in the viewpoint – we are looking at the same phenomenon in both cases. We can also see the diffusion process as a *learning effect* and call the adoption curve a *learning curve* – "potential users have learned to use the product" and "producers have learned to use the potential of a technology in a product."

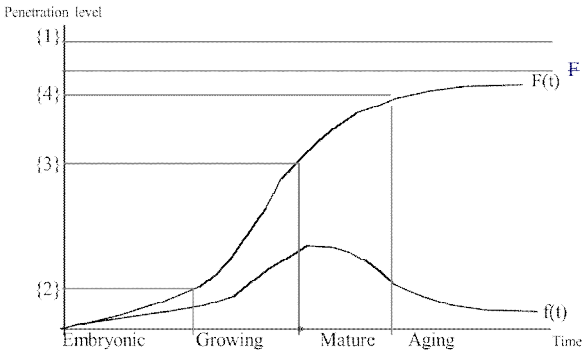
The diffusion process seems to follow – in the product and technology context – a regular curve shape. *Mathematical diffusion analysis* applies the pre-selected mathematical distribution in the data available from the past on the best-fit basis. The most commonly used mathematical distributions are Logistic distribution (also called Pearl Curve), Normal distribution, Weibull distribution, Gompertz distribution, Exponential distribution, and modifications / combinations of these. The data available from the real phenomenon (past time) is fitted in the mathematical formula of the model to evaluate its parameters. The formula is then used to extrapolate future behavior. Sometimes the past values can be replaced by the values of another phenomenon that is known and expected to show similarity in diffusion. In this paper we will not go into details of mathematical diffusion models.

The *heuristic diffusion model - HDM* [1] applies the basics of mathematical models. However, it avoids producing exact predictions; instead it focuses on *understanding the arguments* affecting the diffusion process. In spite of its regularity, the real process includes a lot of *irregularity*, which leads to incorrect long-term predictions. Instead of numbers and percentages, the HDM increases understanding and creates knowledge of the changes that happened (why the changes have been as they are) and provides the means to find arguments that will affect the coming changes.

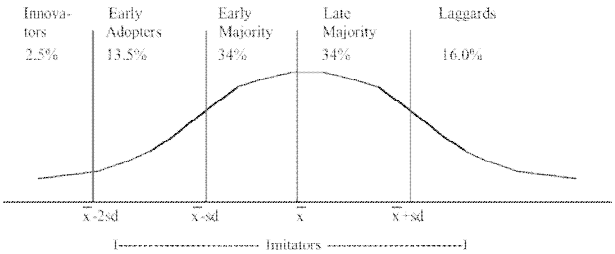
The (mathematical) regularity of the diffusion process is discussed in sub-section B and the principle of HDM in sub-section C. Sub-section D introduces an additional "theoretical" approach, called the Hype Slope or Hype Curve.

B. The S-Shape

As discussed above, the diffusion process is modeled by selected mathematical distributions. Figure 1 represents a generalized model known as the *S-Shape*, because the cumulative (penetration) function has an S-shape (with different variations). Respectively, the non-cumulative (frequency) function has a "bell-shape." Two viewpoints – diffusion and the adoption approach are included.



a. Diffusion



b. Adoption

Figure 1. Diffusion and Adoption

The upper part of Figure 1 (a) uses the following notations:

- f(t) non-cumulative adoption function: amount of new user / usage in time;
- F(t) cumulative adoption function (diffusion): sum of all users / usage in time;
- ~~F~~ potential adopter population: maximum amount of users / capacity available;
- {1} the whole population, in which the potential adopter population is a subset;

- {2} the *lower threshold level* of penetration: diffusion before this point is not regular;
- {3} on the *inflection point* t=t*, f(t) has its maximum value: the growth of new users per time starts to decrease;
- {4} the *upper threshold* level, maturation level: diffusion after this point is no longer regular.

The lower and upper threshold points and the inflection point divide the *life cycle* into four parts. In the *embryonic* phase, the diffusion process is not regular and adoption / usage decisions are not based on rational (modellable) decision making. In the *growing* phase, the number of new adopters grows, and after the inflection point, in the *mature* phase, it decreases. In the *aging* phase, the full penetration potential is in practice reached and the process is no longer regular. In fact, the process would also continue so that the penetration level starts to decrease, because the technology / product analyzed will be replaced by a newer one (*decline* phase).

Without going into the details, the potential sources of analysis errors occur at three points:

- the real data before the lower threshold point {2} (or after the upper threshold) is used in the analysis (this topic is discussed in connection with the Hype Slope later on in this paper);
- the trend includes irregularity that cannot be explained and the real data is fitted by force to the selected mathematical distribution – the reasons are discussed in connection with HDM;
- the potential ~~F~~ is estimated wrongly – this is also discussed in sub-section C.

The best fit in mathematical analysis is usually calculated using the least square (or related) methods.

The lower part of Figure 1 (b) approaches the diffusion process from the *adopter's point* of view; it also explains the diffusion phenomenon. Rogers [8, p. 247] divides adoption (and adopters) into two main clusters: *innovators*, who are the users that eagerly test everything new and want to have experiences of new products, and *imitators*, who are users that base their adoption decision on the experiences of others (word-of-mouth effect). According to Rogers, the innovators are the first users of an innovation. Majahan et al. in [9] point out that both innovators and imitators exist along the adoption process and the only difference is the mechanism behind the adoption decision. The imitators (in both approaches) are further divided into four sub-categories according to the adoption speed. From the mathematical modeling point of view, the frequency function is expected to follow the *normal distribution*, which leads to the regular S-shape cumulative function. This model gives heuristic analysis a tool for understanding the differences in decision making. It is important to recognize the role of "lead users" and "trend-makers" from the population in speeding up the adoption process.

What more needs to be said about mathematical diffusion modeling? In this context we would like to refer to the text of a Finnish author, Veikko Huovinen. In his book "Havukka-ahon ajattelija" [10] (in English "Backwoods Philosopher") the main character, Konsta Pylkkänen, says:

"Far-seeing wisdom [Kaukoviisaus] – it is that everything is considered in advance and all situations are visualized in our mind so that whenever it happens, everything is clear. This kind of wisdom is given only to few of us. The one who has it, would be happy for it. But, there are two problems, even with this kind of wisdom: the expected situation does not happen at all or it happens in a different way. To those who are wise enough to take also these unexpected situations into account, life will be easy to live."

The book has not been translated into English; as a result, the translation was made by the authors and may be slightly inaccurate (the original text is available in the endnote[i] at the end of the paper). It crystallizes the main problem of prediction: in spite of careful modeling and prediction, something unexpected may still happen. If we had more understanding of the phenomenon itself, life would be easier. Because of that we need *heuristics* and *understanding* of the phenomenon. The main problem of trend-based analysis is that the models usually expect the environment of the phenomenon to be stable, which is quite often not true in real situations; changes in the environment affect the parameters of the mathematical formula used in trend analysis and lead to the wrong predictions. Consequently, in some / most cases instead of mathematical exactness process understanding would be a more valuable alternative. Even in this case, an understanding of the principles of mathematical trend modeling is needed.

C. The Heuristic Model

In the *Heuristic Diffusion Model* (HDM), the features of the process are disassembled from empirical data rather than having the goal of mathematical accuracy. This gives useful information, which is hidden inside the mathematical formula. The goal is to *increase understanding* of the diffusion process and to *provide knowledge* for further analysis, even in the case of short time series and short product lifecycles. HDM can also be used for benchmarking purposes – data encountered from one usage context is applied in the situation of another. One example: if we know the adoption trend of 3G mobile phones, we can use this trend to predict the adoption trend of 4G mobile phones; if we know the adoption trend of a certain product in the USA (leader), we can use this trend to build a prediction for Germany (follower). The result is a *diffusion process description* based on features and information important for the analyst. The disassembling principles are discussed in brief below (derived from [2] and [3]; illustrations are given in Fig. 2).

Understanding the life cycle model principles (A): The lifecycle model of an innovation (sub-section B, Fig. 1) establishes a basis for analysis. Every phase has its typical features. Life cycle models can be used even in finding an analogy between the process already known and the process being analyzed.

Categorization of the user population (B): The adoption intensity of a user population is not homogenous.

The basic model classifies users in five adopter classes (innovators, early adopters, early majority, late majority, and laggards; Fig. 1). In addition to the basic classification, the adoption process can be analyzed based on other factors: the purchasing power of the users (heavy, moderate or light buyers), the users' role (home, work; see [2]), the role of multi-adaption, level of uncertainty in the adoption process, negative diffusion (decline) as a part of the life cycle.

Recognition of meta, product, and process level of innovations (C): There is an interrelation and a time lag between changes in the meta level (availability of *elementary technology*), diffusion of *products* (the normal diffusion process), and changes in production (related) technology. These follow each other in the order: meta -> product -> process; in analysis we may find indications from meta level changes to product level, and from product level to process level. This can also be seen when analyzing the competitive factors, which are first based on advanced technology, after that on product properties, and finally on effective production and logistics.

User potential is flexible (D): Mathematical diffusion models expect the number of potential users to be fixed (or known). However, there are a huge number of factors affecting supply and demand that have either a direct or indirect effect on adoption by changing the structure and the extent of the user potential. Instead of being stable, the number of potential users follows some (modellable) curve.

Understanding the role of technology generations (E): In the current world it is very typical that existing technology is replaced by newer generation solutions providing the same (or extended) functionality. An easy example comes again from the mobile telephone world: 1G analog, 2G digital, 3G UMTS, 4G wide band data have followed each other. The adoption decision of the potential adopter is controlled by the expected benefit of the updated technology. Better technical properties, lower price, user experiences that are available etc. lead to the situation in which the duration from breakthrough point to maturation phase is shorter in a new generation and the volume of the potential is more extensive.

The diffusion process is modular and constructed of key technology diffusions (F): The (macro level) diffusion process consists quite often of a sequence of *micro* level diffusions (representing the key technology of a certain period). These follow each other in time. Every micro level process can be analyzed as an individual diffusion process, but at the same time it forms a part of the macro level process.

The process may have discontinuity points (G): The macro level diffusion may have an irregularity that is caused by trend discontinuity - in practice, it is a transfer point from one trend to another. Two different discontinuity classes affect the analysis: (1) *phase discontinuity*; and (2) *environmental discontinuity*. The former occurs on the macro level, when the key technology is changing, and the analyst must notice that the micro level trend has changed. The latter occurs when some radical changes happen in the environment of the technology (e.g. changes in the national economy).

The importance of the depth (quality, significance of use) dimension of adoption (H): The usual approach to diffusion analysis is bi-fold. The *quantitative* growth tends to decrease in time (according to the S-Shape). Simultaneously, the *significance* of the use (*infusion* – importance of the innovation) grows and its explanatory power in the total process (diffusion + infusion – three-dimensional diagrams) becomes dominant. Qualitative growth is a sum of several factors, e.g. better performance of tools and equipment, a better ability to benefit from existing facilities (after a learning period), etc.

Benchmarking of the leader's experience to predict the follower's behavior (I): In diffusion we can recognize two roles: the leader – as an early adopter – and the follower – as a later adopter. This can be seen in the basic principles of the diffusion process (Fig. 1), which is explained by imitation behavior. The same phenomenon is also included in technology generation analysis (E above). As a general phenomenon we can use leader-follower analysis in the situation where one target group (leader) has progressed in adoption at a high penetration level, as another target group (follower) is just starting it. In this case, the adoption process of the leader may be used as an indication of the adoption process of the follower.

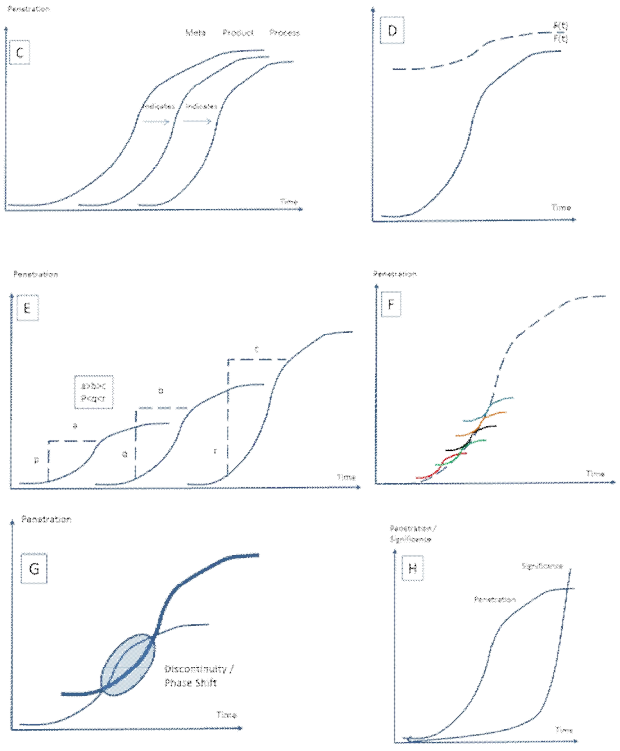


Figure 2. Some principles of HDM illustrated

HDM provides elements for extracting knowledge from the phenomenon analyzed. The principles (phenomena A-I) listed above are examples of situations that are worth recognizing in trend analysis. The list is not intended to be complete, but it includes typical situations. One of the terms used often in the context of future prediction is "*weak signal*". Heuristic analysis of trend extracts these hidden signals from the phenomenon analyzed to provide a solution for the second part of Konsta Pylkkänen's definition of wisdom (see above).

D. *The Hype Slope*

A typical adoption curve follows the S-shape (or modifications of it) and the prediction of future behavior is based on the data available from the early life cycle phases. As discussed in sub-section B, uncertainty is very high in the embryonic phase (as well as in the aging phase). One difficulty that has not yet been discussed relates to recognizing when the inflection point is reached and the trend becomes stable and analyzable. As early as the embryonic phase, the analyst may notice a promising first part of an S-curve (far left of Fig. 3) and start to use it for predicting the future trend. In reality, the phenomenon analyzed is still in the embryonic phase providing no evidence for the future.

The Gartner Group approached the solving of this problem in the form of the Hype Slope [11]. The Hype Slope (Fig. 3) models the adoption process of the Embryonic Phase of the diffusion curve (in Fig. 1), which is an inference-sensitive part of the S-shape.

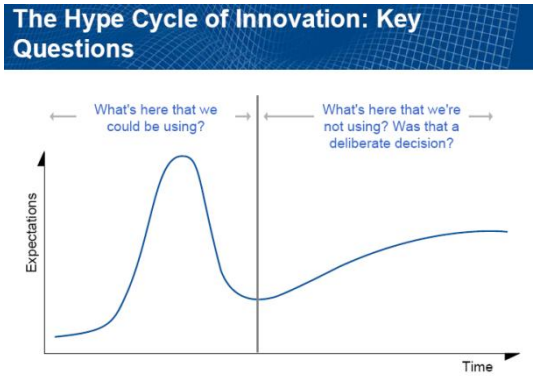


Figure 3. The Hype slope [11]

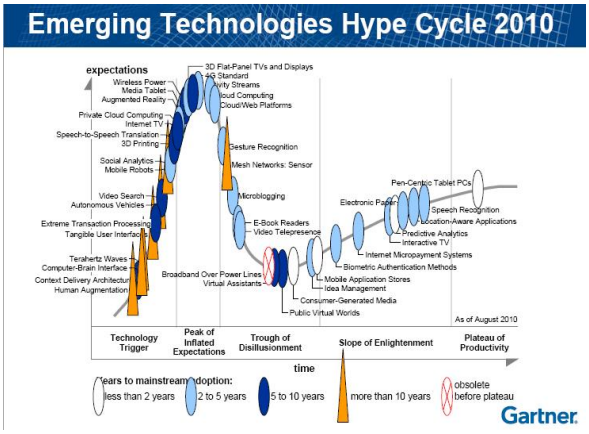
The Hype Slope is divided into two parts: the left-hand side "*hype*" and right-hand side "*non-hype*." Gartner uses this tool for categorizing the importance of expectations and the time span to the mainstream adoption of selected technologies. As a tool it is defined [13]: "*When new technologies make bold promises, how do you discern the hype from what's commercially viable? And when will such claims pay off, if at all? Gartner Hype Cycles provide a graphic representation of the maturity and adoption of technologies and applications, and how they are potentially relevant to solving real business problems and exploiting new opportunities. Gartner Hype Cycle methodology gives you a view of how a technology or application will evolve over time, providing a sound source of insight to manage its deployment within the context of your specific business goals.*" These are discussed in Section III.

III. THE TRENDS OF TODAY

A. *General trends of Emerging Technologies*

The future can be seen as a continuum of the past – in most cases. As a result, several organizations are monitoring the changes in technology to understand coming changes (threats / opportunities). One of these monitoring companies is Gartner Inc., which publishes (since 1995) an annual analysis "*Hype slope of Emerging Technologies*" both on a general level [11] and also in a business sector dedicated form (not available to the authors, commercially available products). The general trends are illustrated in Fig. 3.

Fig. 4 applies the Hype Slope method in the context of Emerging Technologies. The slope (a) illustrates the position of selected technologies in the phases of the curve: *Technology triggering, Peak of expectations, Trough of disillusionment, Slope of enlightenment* (first step on the real diffusion curve included in the embryonic phase) and *Plateau of productivity*. The slope also visualizes the years before mainstream adoption (figure shape / color), which is easier to see in the lower (b) part of the figure; the table format adds the expected value of the innovation as a new dimension to the figure. A short explanation of selected (most interesting) technologies is given below (based on [12]).



a. Hype slope

Years to ms adoption -> / Benefit	<2 years	2-5 years	5-10 years	>10 years
Transformational		• **CloudComputing** • **Cloud Web Platforms** • Media Tablet	• 3D Printing • Context Delivery Architecture • Extreme Transaction Processing	• Autonomous vechiles • **Human augmentation** • Mobile robots • Terahertz waves
High	• Mobile application stores • **Predictive Analytics**	• **Activity streams** • E-Book readerrs • Electronic paper • Interactive TV • Internet micropayment systems • **Location aware applications** • Private cloud computing • **Social analytics**	• **Augmented reality** • Internet TV • **Virtual assistants** • Wireless power	• Mesh networks: sensor
Moderate	• **Consumer generated media** • Pen-centric tablet PCs	• 3D flat TV and displays • Biometric authentication method • Gesture recognition • Idea management • **Microblogging** • Speech recognition • **Video telepresence**	• 4G standard • **Public virtual worlds** • Speech-to-speech translation	• Computer Brain interface
Low				• Tangible UI

b. Priority Matrix (modified)

Figure 4. Emerging Technologies 2010 [11]

Predictive analytics encompasses a variety of techniques from statistics, data mining, and game theory that analyze current and historical facts to make predictions about future events.

Consumer-generated media indicates the opportunity for producing content in social networks to fill the individual needs of consumers.

Cloud computing indicates the trend to move applications from individual computers to the Internet – providing easy availability and scalability.

Activity stream builds a life stream for individual persons by aggregating their social activities all in one place.

Microblogging is a broadcast medium in the form of blogging. Microblogs allow users to exchange small elements of content such as short sentences, individual images, or video links.

Augmented reality (AR) is a term for a live direct or an indirect view of a physical, real-world environment whose elements are augmented by computer-generated sensory input, such as sound or graphics. It is related to a more general concept called *mediated reality*, in which a view of reality is modified (possibly even diminished rather than augmented) by a computer. As a result, the technology functions by enhancing one's current perception of reality.

Virtual assistant is a (human) entrepreneur who provides professional administrative, technical, or creative (social) assistance to clients from a home office on a demand basis.

Human augmentation is used to improve the capability of a human being by the assistance of technology (e.g. wearable robots).

Telepresence refers to a set of technologies which allow a person to feel as if they were present, to give the appearance that they were present, or to have an effect, via telerobotics, at a place other than their true location

The technologies discussed above emphasize the growing role of virtual reality, the growing importance of networking in different levels, the integration of human being and technology, the need for intelligent filtering, and the ability to combine existing data and knowledge. The next sub-sections widen the analysis of Gartner by synthesizing it with "future analysis" from other sources.

B. *Mobile dominance is changing the ICT ecosystem*

Mobile technology has grown in importance fast and its progress seems to be accelerating. The activities used to perform in a networked environment are being transferred to an increasing extent to the wireless environment.

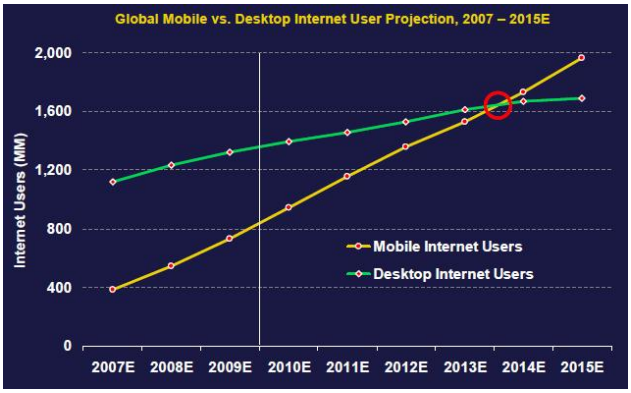


Figure 5. Changes in Internet User Base [14]

The trend is confirmed by Fig. 5 (Morgan Stanley, [14]). Email usage has been one of the main network applications. The trend seems to be towards dominance of mobile users with a crossing point in 2014. This trend is accelerated especially by the fast growth of usage in less developed economies having a poor infrastructure of traditional computer networks. One indication of the fast growing importance of mobile usage is email, which is expected to grow from 354M in 2009 to 713M in 2014 – i.e. more than double.

In principle, almost everything that is available for users in traditional networks is also available for mobile terminals in ubiquitous use. A remarkable observation is that up until now – because of restricted transmission speed and processing capacity – mobile terminals have been in the role of follower compared to networked terminals (leaders). Mobile application users have suffered from restricted functionality and poor user interfaces. At the moment, thanks to the fast adoption of advanced mobile terminals, the direction has changed: mobile users' needs are primarily taken into account in application development. User interfaces of networked and mobile terminals are simultaneously being unified. This can be seen e.g. in the case of Windows 8: first the renewed user interface was available in mobile terminals and after that in desktop and laptop computers. Additional innovation related to this progress is Cloud Technology: instead of local memory, applications are supported in increasing numbers by file space provided by cloud services (Skydrive, Dropbox, GoogleDrive, ICloud). Even this progress has been strongly supported by the strategic decisions related to mobile terminal platforms. Piece by piece, the symmetry between traditional and mobile solutions is becoming reality.

Hardware progress seems to follow the same trend: the primary focus at the moment is on the design of fast multi-core processors with low power consumption – the primary target is the mobile terminal. The diversity of mobile devices is growing: light portable PCs, smartphones and tablets provide alternatives for different use situations and needs. It seems to be time also for mobile information systems, based on renewed application architecture and implementation principles. As a consequence, the concept of software is changing.

What is happening then on the software market? Instead of software ownership, the fast growing availability of SaaS (Software-as-a-Service) based products and processing services seems to be becoming dominant – first in routine applications, but finally also in more strategic ones. Instead of high acquisition prices and licenses, users pay for volume – time, transactions, storing capacity etc. – and get scalable access to enormous resources. Another growing business, coming from the mobile world, is the existence of application stores. Apple Store, Android market (Play), Nokia Store, etc. provide thousands of beneficial applications for mobile users (smartphones and tablets) at reasonably cheap prices (earning money from software sales is based on multiple low-cost purchases rather than a few expensive licenses). The numeric data gives evidence of this: the estimated number of downloads from App Stores in 2014 is 70 billion (Gartner, [24]). The software market is changing

from a centralized to brokering market model. This trend also changes application architectures dramatically: instead of monolithic large information systems the future is one of simple solutions used as components to assemble more complex ones. The time of downgraded traditional software for less powerful terminals is over.

Related to the topic of this Mipro / CE conference, an interesting mobile application area is "Mobile School" [25]. This Stanford University project involves the assessment, design, development, implementation, and evaluation of mobile learning technology to provide underserved and marginalized children around the world with equitable access to basic education and literacy exposure in health and environmental safety. The primary framework for the design and implementation addresses situation specificity, cultural sensitivity, practical usability, theoretical applicability, economic scalability, and viable sustainability. Creative content and mobile applications are being created through the Stanford Mobile Empowerment Developers Network.

As a short summary of the discussion above, Table 1 includes some typical software-related trends (changes) and their environment in a comparison between traditional and mobile platforms.

TABLE 1. APPLICATION CONVERGENCE AND ENRICHMENT

Traditional	Mobile & Ubiquitous
High speed networks	Slower speed networks
High processing capacity	Lower processing capacity - Proxy processing needed
Diversified visual capacitty	Limited visual capacity
Closed application base	Open application base - Open application market
From legacy to cloud	Cloud built-in to the concept
Stable context	Dynamic context -context sensitivity
…	…

One additional trend related to mobility relates to network topology. In this new era of Information Systems, network topology, which traditionally has been based on Client-Server (C/S) architecture is changing to Server-Server (S/S) architecture. The role of an individual node may vary from client to server in different situations. The nodes (terminals) are context sensitive and adapt in the use situation. Context includes dimensions, like computing, user, physical environment, time. Chen and Kotz [15] define context as "*the set of environmental states and settings that either determines an application's behavior or in which an application event occurs and is interesting to the user*". Because of the higher degree of freedom, the mobile context is also richer than the context of the traditional network environment. It is easy to conclude that this is also the time of context sensitive information systems.

Related to mobile application users, Forrester uses the term "dumb smartphone users" [16]: "*Thanks to handset subsidies, smartphones will be available to the masses. Expect new smartphone users to be less engaged and active than the first cohorts of Android and iPhone early adopters. The good news is that thanks to customer education and the convenience that these devices offer, even "dumb" smartphone users will consume more mobile media than ever before and will have incremental usage of mobile data*!" This indicates the fast growing usage of mobile services, which requires wide geographical network coverage and high technical quality access to the services for acceptable prices. Forrester also points out that the term "mobile" will mean a lot more than mobile phones. Consumers are connecting more and more (different kind of) devices wirelessly to the Internet.

A technology that is both emergent but also suitable to discuss briefly in this sub-section is the *Internet of Things* (IoT). The term itself has several definitions; a short and simple one is given by SAP in [17]: *"A world where physical objects are seamlessly integrated into the information network, and where physical objects can become active participants in business processes. Services are available to interact with these 'smart objects' over the Internet, query and change their state and any information associated with them, taking into account security and privacy issues*." Active components of IoT are identification technology (e.g. RFID), device to device, and device to service messaging. This technology is highly analogous with the role of mobile devices in a ubiquitous world – devices are recognized by services, which react to the demand for service without human control. Even mobile devices provide increasing opportunities for interaction with the physical environment by using QR codes, NFC, augmented reality services etc. Ubiquity is also providing new kinds of opportunities for the media – "TV everywhere" is a phenomenon that had great importance in distributing videos from the demonstrations during the "Arab Spring." Growing network capacity provides means for telepresence applications (virtual meetings).

C. *More about trendy things*

The sub-sections above focused on the two most important sources of trends – emerging technologies and mobility. In writing this paper it was clear from the beginning that one conference paper does not provide the opportunity for wide coverage of all relevant topics – not even for all the most important ones. Therefore we have to be satisfied with switching the presentation style to listwise mode rather than an explanatory one.

People and organizations are the key elements of society. Along with the changes of generations, the capabilities, attitudes, and skills of people are changing. In public discussion the generations are used to classify by the letters X (born in 1960s and 1970s), Y (born in 1980s and 1990s), and Z (born in 2000-); for details see e.g. [18]. A big change will happen in society when the current X (and older Baby Boom) generation is replaced by the younger ones. *Generation Z* could be characterized by the following behavioral patterns: Used to being active members of social communities; Used to using the internet as a tool of knowledge mining; Used to collecting and synthesizing knowledge from small chunks of information; Used to handling data in multiprocessing mode; Solve problems in a bottom-up manner; Are well educated; Are international and global – used to working and communicating in a global context, used to traveling and meeting people from foreign cultures, but also local –

giving value in their own culture; Used to using foreign languages; Have adopted green values. Just a short conclusion in relation to this: they seem to be a lot different from our generation, i.e. those of the generation leading the society and organizations today. A management revolution is to be expected and as a consequence the transfer to a new enterprise culture is also to be expected – this is discussed in the next paragraph.

Andrew McAfee [19] introduced the term Enterprise 2.0 as "*the use of emergent social software platforms within companies, or between companies and their partners or customers*." His book [20] provides the means for managers to develop their organizations to adopt the principles of the concept, which include: flat organization structure (instead of hierarchical), interaction sensitivity (instead of automated processes), the use of emergent (instead of dictated) information systems, agility (instead of bureaucracy), dynamics and adaptivity (instead of static and rigid), the use of user-driven technology (instead of IT-driven), short time-to-market cycles (instead of long), bottom-up work culture (instead of top-down), distributed organization structure (instead of centralized), self-organizing teams (instead of management-driven), open borders (instead of boundaries and silos), transparency (instead of controlled communication), folksonomies (instead of taxonomies), simplicity (instead of complexity) and open (instead of closed) standards. The ideology of Enterprise 2.0 fits well into the value and skill profile of the coming employee generation, which also indicates a kind of organizational revolution.

Other trends collected by the authors are listed – without detailed comment – below. The sources are not listed for two reasons: some of the findings are not from original sources, and there are overlapping ones. The sources reviewed cover web material from Forrester, Infoworld, ITBusinessEdge, The Wall Street Journal, Stanford University Reports, Techarta, CS Monitor, BBC, CNN, Computerworld, Fjord, (references listed in Appendix 1). This was collected by a search in Google and Google Scholar using the key words "mobile technology trends" (349 million hits) and "mobile information systems" (3 million hits). The following trends / technologies (among others) are mentioned in the material: the growth of mobile marketing in many forms and supported by the recognition of the user's context, the growing role of social networking in the distribution of on-time data, services supported by interaction with the physical environment, Web 2.0 adoption in business (Enterprise 2.0), mobile payment, mobile search connected to actions (immediate reaction of the search findings), mobile payment, improved user experience based on context awareness (knowing his/her interests, intentions, priorities, etc,.), object recognition – sensors and processing capacity for recognizing the surroundings, next generation analytics (in-line embedded, predictive, collaborative), AI and robots (e.g. Siri), user interface revolution (touch, gesture, search, voice, video – adopted from mobile terminals), social gestures (all you use, read, listen, watch, etc. is automatically posted to the profile, which is used by services as your context).

IV. CONCLUSION

When the authors started to write this paper, the goal was to give a wide report of trends affecting our lives in the near future. The analysis was started by executing Google searches to find suitable material. The extent of findings was chaotic on the one hand – because of the amount of interesting material - but on the other hand comforting, because the similarity of the findings soon became clear.

Our paper consisted of two parts: the aim of the first part (Section II) was to increase general level understanding of the trends and trend analysis, and the second part (Section III) opened up a discussion on the important trends that can be recognized today. We hope that the first part helped the reader to interpret the progress of the phenomena discussed in the second part.

Our paper listed several innovations that are either already here or imminent. The main driver and enabler of the changes seems to be *technology* itself. The key elements behind the trends discussed in this paper are the growth of processing and transmission capacity. As a consequence, networking (on physical and logical level), large data repositories (Big Data), innovative user interfaces, built-in intelligence, embedded sensors etc. provide new types of innovations first on a technology level and then inbuilt in daily life and business processes. The progress seems to be accelerating: according to Moore's law [21; 22] the computing hardware, i.e. the number of transistors on integrated circuits is doubling approximately every two years. The period is also often quoted as "18 months," which refers to the period for a doubling in chip performance (being a combination of the effect of more transistors and their increased speed). This trend – or "law" – might be one of the most often cited trends. It has been true since the 1960s and the trend of the innovation power of transistor and VLSI technology looks likely to continue in the future as well. The Wikipedia material on Moore's law [22] includes a long list of references that are interesting for further studies. The book of Endres and Rombach [23] provides interesting reading on other ICT-related laws.

To conclude this paper it is worth going back to the question of the title: "Are we trendy?" Yes, we are. We cannot avoid the fact that changes follow some regularity, which makes it easier for us to be prepared for adoption and to embrace changes, when the time is right. The right moment, however, is not the same for all – some of us are leaders, some are followers, some are early adopters and some laggards. Understanding trends also makes it easier for us to find the right moment.

REFERENCES

[1] Jaakkola Hannu, An Analysis of Diffusion of Information Technology in Finnish Industry. PhD Thesis. Tampere: Tampere University of Technology Publications 70, 1970.

[2] Jaakkola Hannu, Kajanto Markus, Neuvo Yrjö, Analysis of Mobile Telephone Diffusion. In Kocaoglu D., Anderson, T. (eds.), Proceedings Vol 1: PICMET '99 Summaries. Piscataway, NJ: IEEE, 1999.

[3] Jaakkola H., Gabbouj M., Neuvo Y.(1998), "Fundamentals of Technology Diffusion and Mobile Phone Case Study". Circuits Systems, Signal Processing, Vol. 17, 3. pp. 421-446.

[4] Tarde, Gabriel, Les Lois de l'Imitation. Original edition in 1890. Second Edition in 1895. Republished, based on the second edition. Paris: Kimé, 1993. Available in http://classiques.uqac.ca/classiques/tarde_gabriel/lois_imitation/tarde_lois_imitation_1.pdf. Retrieved January 8th, 2013.

[5] Hashimoto, T., Henno, J., Jaakkola, H., Sasa, A. and Thalheim, B. (2012). "Infrastructures for Knowledge Systems Environments". In Information Modelling and Knowledge Bases XXIII. Henno, J., Kiyoki, Y., Tokuda, T., Jaakkola, H. and Yoshida, N., Eds. Amsterdam:, IOS Press, pp. 369-398.

[6] Jaakkola, H., Linna, P., Henno, J., Mäkelä, J. and Welzer-Druzovec, T. (2012). "A Path Towards Networked Organizations - in the Push of Digital Natives or in Pull of the Needs? ." International Journal of Knowledge Engineering and Soft Data Paradigms (IJKESDP) 3(3/4), pp. 240-260.

[7] Jaakkola, H., Linna, P. and Henno, J. "(Social) networking is coming - are we ready?" In the Proceedings of MIPRO 2011, Cicin-Sain, M., Uroda, I., Prstacic, I. T. and Sluganovic, I., (Eds.). Opatija, Mipro and IEEE, 2011, pp. 170-176.

[8] Rogers E., Diffusion of Innovations. New York: The Free Press, 1983.

[9] Majahan Vijan, Muller Eita., Bass Frank M., "Dynamics of Innovation Diffusion: New Product Growth Models in Marketing". Journal of Marketing 54,1 (Jan 1990), pp. 1-26.

[10] Huovinen Veikko, Havukka-ahon ajattelija (Backwoods Philosopher). Helsinki: WSOY, 1952.

[11] Gartner Group, Emerging Technologies Hype Cycle 2010. http://www.gartner.com/it/content/1395600/1395613/august_4_whats_hot_hype_2010_jfenn.pdf. Retrieved Aug 8th, 2011.

[12] Sasaki, S., Vojtas, P., Jannaschk, K., Thalheim, B., Jaakkola, H. and Kiyoki, Y., "Multimedia Information Systems for Social, Cross-Cultural and Environmental Computing". In Information Modelling and Knowledge Bases XXIV. Vojtas, P., Kiyoki, Y., Tokuda, T., Jaakkola, H. and Yoshida, N., Eds. Amsterdam: IOS Press, 2013.

[13] Gartner Group, Research Methodologies. http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp. Retrieved January 9th, 2013.

[14] Morgan Stanley, Internet Trends. http://www.morganstanley.com/institutional/techresearch/pdfs/Internet_Trends_041210.pdf. Retrieved August 3rd, 2012.

[15] Chen Guanling , Kotz David, A Survey of Context-Aware Mobile Computing Research. https://mmlab.snu.ac.kr/courses/2005_advanced_internet/handout/ppt/36%20-%20context_aware.pdf. Retrieved January 3rd, 2013.

[16] Mobile Marketing Watch, Mobile Technology Trends For 2011, According To Forrester Research. http://www.mobilemarketingwatch.com/mobile-technology-trends-for-2011-according-to-forrester-research-12801/. Retrieved January 9th, 2013. Original material available in http://www.forrester.com/.

[17] Wikipedia, Internet of Things. http://en.wikipedia.org/wiki/Internet_of_Things. Retrieved January 9th, 2013.

[18] Wikipedia 2011, List of Generations. http://en.wikipedia.org/wiki/List_of_generations#List_of_generations. Retrieved Aug 8th, 2011

[19] McAfee, Andrew, Enterprise 2.0. Enterprise 2.0, Version 2.0. http://andrewmcafee.org/2006/05/enterprise_20_version_20/. Retrieved January 8th, 2013.

[20] McAfee, Andrew, Enterprise 2.0: New Collaborative Tools for Your Organization's Toughest Challenges. Harvard Business Press, 2009.

[21] Moore, Gordon E., "Cramming more components onto integrated circuits". Electronics 38, 8 (April 1965), 4 pages. Available in http://download.intel.com/museum/Moores_Law/Articles-Press_Releases/Gordon_Moore_1965_Article.pdf

[22] Wikipedia, Moore's Law. http://en.wikipedia.org/wiki/Moore's_law. Retrieved January 17th, 2013.

[23] Endres, A. and Rombach, D., A Handbook of Software and Systems Engineering - Empirical Observations, laws and Theories. Pearson and Addison Wesley, 2003.

[24] Gartner Group, Top 10 Strategic Technologies for 2012. http://www.gartner.com/it/page.jsp?id=1826214. Retrieved May 12th, 2012.

[25] Stanford University, Mobile School. http://gse-it.stanford.edu/research/project/pocketschool. Retrieved January 18th, 2013.

Appendix 1. List of www-material sources

Fjord:

http://www.slideshare.net/fjordnet/fjord-digital-trends2012final

Forrester:

http://www.mobilemarketingwatch.com/mobile-technology-trends-for-2011-according-to-forrester-research-12801/

http://www.forrester.com/2011+Mobile+Trends/fulltext/-/E-RES57183?objectid=RES57183

Computerworld

http://www.computerworld.com/s/article/352068/5_Trends_to_Watch_in_2011

Infoworld

http://www.infoworld.com/d/mobile-technology/5-smartphone-trends-mobile-world-congress-187497?page=0,1

IT Business Edge

http://www.itbusinessedge.com/slideshows/show.aspx?c=87261

Wall Street Journal

http://online.wsj.com/article/SB10001424052748704774604576035611315663944.html

Techarta

http://techarta.com/top-trends-in-mobile-technology/

CS Monitor

http://www.csmonitor.com/Business/2012/0101/Technology-2012-Four-tech-trends-to-watch/Mobile-technologies-in-2012-will-free-us-and-save-lives

CNN

http://edition.cnn.com/2011/12/19/tech/innovation/top-tech-trends-2012/index.html

---

[i] Original text: "Se on sitä, että asiat harkitaan etukäteen ja kuvitellaan tapaus sikseenkin elävästi, että kun se kerran tapahtuu, on reitit selvät. Tätä lajia on harvalle suotu, jolla sitä on, niin pitäköön hyvänään! Mutta tässä lajissa on kaksi vikaa: asia jää huvikseen tapahtumatta tai se tapahtuu eri tavalla. Joka arvaa ottaa nämäkin huomioon, sille on maailmanranta kevyt kiertää."

Concept maps in computer-assisted knowledge assessment

Juraj Petrović, Branko Jeren and Predrag Pale
University of Zagreb, Faculty of Electrical Engineering and Computing
Department of Electronic Systems and Information Processing, Zagreb, Croatia
juraj.petrovic@fer.hr

Abstract – Concept maps are a learning tool that has been successfully applied in the educational process for more than three decades now. Their original aim was to serve as a learning aid, offering a more efficient overview of the learning content and more successful integration of new knowledge to the learner's schemata. Later, thanks to the popularization and increase in availability and applications of computers in learning and assessment they emerged into an efficient knowledge assessment tool as well. Concept maps are therefore today widely perceived and used not just as an e-learning but also e-assessment tool enabling assessment of knowledge as they explicitly unveil domain concepts and their relationships. It is the aim of this paper to provide an overview of most important aspects, practices and achievements in using of concept maps in computer-assisted knowledge assessment.

I. DEFINING A CONCEPT MAP

The invention of concept maps is usually attributed to John D. Novak, who in 1979, together with his students Stewart, Van Kirk and Rowell, invented them to help children learn biology by representing knowledge in form of hierarchical, node-link diagrams where nodes represented domain concepts and links implying relationships between those concepts [1], later marked by a label describing the relationship [2]. The basic unit of knowledge or meaning in a concept map is called a *proposition*, and refers to a pair of concepts and a label describing their relationship [3]. A concept map is therefore a "*visual representation of the relationships between concepts held by an individual"* [4]. Figure 1 is an example of some of the first Novak's concept maps.

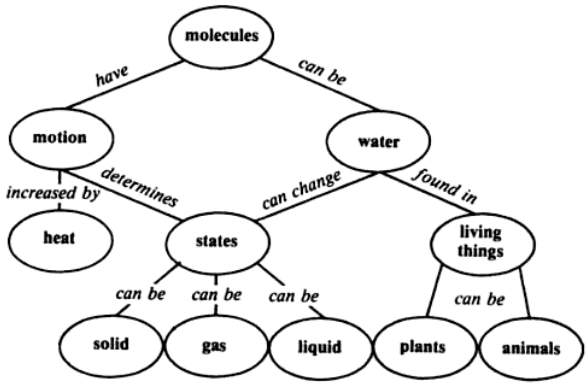


Figure 1. One of Novak's first concept maps [2]

This simple, yet efficient idea of representing knowledge in a hierarchy of concepts was inspired by schemata theory [5], [6] and works of David Ausubel [7], [8], who suggested internal representation of knowledge is "*hierarchically organized in terms of highly inclusive concepts under which are subsumed less inclusive subconcepts and informational data*" [9]. Consequently, the most important aspect of acquiring new knowledge is the existing knowledge (or prior knowledge) structure since it provides anchoring sites for the new knowledge [10]. And that is what concept maps describe – the existing or to-be-acquired knowledge so a learner can recognize how it relates to what he already knows. The idea that knowledge is internally represented as an organized network of interrelated concepts, ideas and procedures and that one's understanding of a domain is directly related to the number of connections relating them has remained supported by cognitive psychologists [11]–[13].

Concept map improvements and refinements have been introduced in the research following their invention. As different propositions can be of different importance for the whole map, an idea of *weighted concept maps* was introduced somewhat later [14]. The authors suggested proposition weights could range from 0 to 1 and thereby define proposition importance. Also, although concept maps were originally intended to be hierarchical [15], different kinds of non-hierarchical maps were proposed.

- *Cyclic concept maps*, allowing loops formed through relationships among concepts have been introduced in the literature mostly to describe cyclic relationships among concepts typical in domains like cybernetics, system modeling and social systems [16].
- *Net-like* or *relational concept maps* were suggested to represent domains where there is no key concept and the important knowledge is represented in concept relationships which can be defined between two concepts in one, two or none directions ([17] cited by [18]).
- Finally, *chain-like concept maps* are suggested to represent series of linear cause-effect concepts ([19] cited by [3]).

Additional improvements include technical concept map aspects such as arrowhead pointers in concept links to suggest direction of the link, avoiding hierarchical

structure if not necessary, not using any concept more than once, and suggesting a direction for reading the map (for example top-bottom or bottom-top) [20]. An example of a cyclic concept map is presented in the Figure 2.

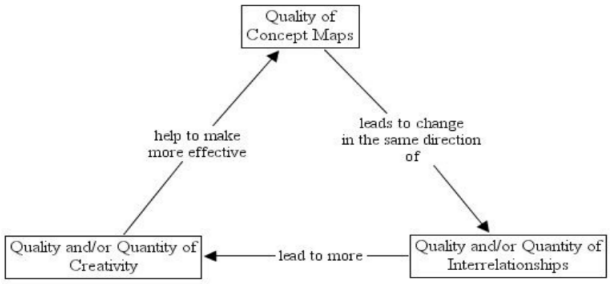


Figure 2. A simple cyclic concept map proposed in [16]

II. FORMAL DEFINITION

A concept map can also be defined formally as a triplet $CM=\{C,R,T\}$ [21], where C is a set of concepts, R is a set of relationships among concepts defined as (c_i,c_j,r_k), meaning that concepts c_i and c_j are in a relationship described by label r_k. Finally, each element of T is a list of concepts from C that share the same level of generalization in the map. As concept maps do not necessarily need to have hierarchical structure, they can also be defined without the set T simply as $CM=\{C,R\}$ [22].

III. CONCEPT MAP CONSTRUCTION

There are two main approaches proceeding to a concept map construction: a normative and a descriptive one [23]. The descriptive approach suggests there might be more than one concept map correctly describing a domain of interest [24]. Still, the normative approach starting from the simplified assumption that there is only one concept map describing the desired domain from the desired perspective or in regard to a given aspect is more common, especially in context of concept map based computer-assisted assessment. In both cases, at least one correct concept map must be provided to enable comparison and grading of other maps in the assessment process. The referent map can be constructed either manually or automatically.

In case of a manual concept map construction there is no single correct construction procedure, since domains' inner structure and aspects that need to be stressed vary. A most commonly cited approach is explained in [8] where Novak suggests the process of manual construction of a concept map as iterative. One should first determine the domain and context of the concept map, identify and list domain concepts (recommended 15-25) starting from more general to more specific, organize them graphically and finally define their relationships. The map should then be refined until it is providing information about all relevant concepts. Although identifying and relating relevant concepts in a domain of interest can be a demanding process even for domain experts [25], it is still the most commonly used method for construction of a concept map. There is also a wide variety of free online and desktop applications supporting this task including *IHMC Cmap Tools*, *yEd*, *VUE*, as well as applications that are originally intended for ontology design like *Protégé* or *KAON* that can be helpful. Most of these tools store a concept map in an XML file format.

Since manual construction of concept maps can be a demanding process, a number of approaches to automatize this procedure have been proposed. Automatic construction of concept maps or *concept map mining* (CMM) refers to such procedures for extracting concepts and concept maps without or with limited need for user's intervention, usually from textual documents [21]. Extraction methods are based on dictionaries and linguistic tools [26], or statistical text properties and data mining [27], and as their result usually include concepts and their connections without labels [28]–[30], concepts and labels describing their relationships but no hierarchy [31], or both concepts and their labeled relationships [32]–[34]. Concept maps resulting from such procedures are often used just to obtain a starting map which is later refined either by an expert or another automatized construction procedure [25]. Unfortunately, there are currently no implementations of automatic concept map construction algorithms available to be tried online or locally.

IV. ASSESSMENT TASKS

Once the referent concept map describing the knowledge one would like to assess is defined, ways of checking for this knowledge or assessment tasks need to be defined. Three main aspects which define a concept map assessment task are *task demand*, *task constraints* and *task content structure* [3].

- Task demand refers to the process or assignment the student needs to fulfill in order to complete the task. Some commonly used task demands include constructing the whole concept map, filling in the missing concepts in the map, defining the connections between concepts, rating their importance, or correcting a proposed concept map.
- Task constraints present limitations while accomplishing the task demand and usually refer to the presence or absence of a predefined list of concepts or relationships that have to be used in the task resulting in closed-ended and open-ended concept maps respectively [35]. There can be only one general concept list or several lists defining concepts that can be used for different parts of the map. Additional misleading or incorrect concepts can also be included in the list.
- Finally, task content structure describes the structure of the concept map domain, usually referring to weather it is a hierarchical, cyclic, net-like or chain-like domain [36].

Varying these three aspects of assessment tasks, numerous different student task examples have been proposed in the literature. It has been suggested that 739 different tasks can be developed just by varying given

map structure (complete, partial or none) and given sets of concepts and links (complete, partial, none, misleading), which can be given in a list or in the map [36]. However, a significantly smaller number of tasks have been commonly used in the literature, mostly based on filling in the missing concepts or relationships provided on a list into an incomplete map [37]–[39], constructing a concept map using a list of concepts or relationships [40], [41] or constructing a concept map from the beginning with no predefined elements [38], [40], [42], [43].

All of the concept mapping applications identified in the previous section can be used to support these assessment tasks, since they can be used to create and provide students with any combination of concept map structure and its predefined elements or their value constraints. Their XML output can simply be parsed and processed with an implementation of desired scoring scheme.

V. SCORING ELEMENTS AND SCHEMES

Concept map scoring scheme is defined as an accurate and consistent procedure for a concept map's evaluation [44]. However, the problem of defining a satisfying scoring scheme is, as identified earlier, the possibility for differences in students' concept maps that are not necessarily incorrect. This problem was noted by Novak already in the early days of concept mapping [2]. Since the scoring method can also depend on teacher's preferences, domain aspects, and concept map based assessment tasks, there is no single widely available and recommended scoring recommendation [45]. Scoring of the map can be performed by having an expert evaluate the map, automatically by comparing the map with the expert's map, or by combining those two approaches. The assessed concept map attributes and scoring method are dependent on the assessment task, yet are based on some qualitative and quantitative criteria [36], [46].

- Quantitative criteria include: number of concepts, links, propositions, branchings, hierarchy levels, concepts per hierarchy level, cross-links, graph diameter, the degree of the maximum-degree node in the map, and examples.
- Qualitative criteria include: semantic value, correctness, completeness, or importance of used concepts' labels, links' labels, and propositions.

Identified assessment criteria can be combined to score the map based on their importance, position or count in the map.

Several types of more advanced quantitative measures have been proposed in the literature, valuing not only the count of given elements in the map, but also their structure in the map. *Closeness index* [47] is one of the first measures proposed and utilized to assess similarity of two concept maps in terms of using same concepts and their connectedness. To calculate closeness index *CI* between a student's concept map $CM^S=\{C^S,R^S\}$ and an expert's map $CM^E=\{C^E,R^E\}$, one should first calculate the closeness index CI_m for every of M unique concepts appearing in any of the two maps ($M=|C^E \cup C^S|$). To calculate the closeness index CI_m for a given concept K_m, element of $C^E \cup C^S$, one should define N^s, the set of concepts related to K_m in the student's map, and N^e, the set of concepts related to K_m in the expert's map. Closeness index CI_m for m-th concept is than equal to $CI_m=|N^s \cap N^e| \div |N^s \cup N^e|$. The overall closeness index *CI* is calculated using the formula (1) and is a numerical value between 0 and 1. The higher is the value of *CI*, the greater is the closeness of two maps.

$$CI = \frac{1}{M}\sum_{i=1}^{M} CI_i \,. \tag{1}$$

Fuzzy closeness index [48] and more recently proposed s*imilarity index* [49] are two measures introduced to account for the weight of propositions which was originally ignored. To calculate the similarity index *S*, each of L proposition $p_{l=}(c_i, c_j, r_k)$ in the referent expert's map, needs to be assigned a value $w(p_l)$ ranging from 0 to 1 describing its importance. Each of P propositions $p_{p=}(c'_i, c'_j, r'_k)$ in a student's map is then evaluated by searching the expert's map for the same proposition. If the corresponding proposition exists in the expert's map, it is assigned its full value $\text{score}(p_p)=w(\text{p}_\text{l})$ and is considered to be learned. If the corresponding proposition exists in the expert's map, yet it is described with a different label, it is considered to be partly learned and is assigned half of the defined proposition score $\text{score}(p_p)=0.5 \cdot w(\text{p}_\text{l})$. Otherwise, $\text{score}(p_p)=0$ is assigned to the target proposition. When all the propositions in the student's map have been assigned a score, the similarity index can be calculated using formula (2). The higher is the value of *S*, the greater is the similarity of two maps.

$$S = \sum_{i=1}^{P} \text{score}(p_i) \div \sum_{j=1}^{L} w(p_j) \,. \tag{2}$$

Some other proposed algorithms assess not only related concepts or propositions in a concept map, but chains of propositions [22], score particular patterns specific for the concept map domain [50] or even automatically score open-ended concept maps [35].

Automatic evaluation or comparison of a concept map with a referent one presents the main advantage of conducting computer-aided concept map assessments, since automatic evaluation can not only save time but also enable instant feedback. Those two parameters are related since manual correction of concept maps requires a greater amount of time and thereby causes delayed feedback. Feedback is especially important in case of formative assessments. The positive effects of providing learners with an appropriate feedback immediately following the assessment process for learning and motivation has been emphasized many times [51]–[53].

VI. ASSESSMENT OBJECTIVES

The last but definitely not least important aspect of concept maps in knowledge assessment is what kind of knowledge can be assessed using concept mapping tasks. Literature about concept map based assessment includes their application in assessment of understanding [54], [55], conceptual knowledge [56], [57], propositional knowledge [58], declarative knowledge [38], and structure of abstract knowledge [59]. The large variety of loosely defined knowledge attributes and qualities has been noted already earlier in the literature about learning and instruction [60].

Although different concept map assessment tasks could be used to assess different types of knowledge, authors of this paper suggest using the terminology and knowledge types defined in Krathwohl's revision of Bloom's taxonomy [61], suggesting there are four types of knowledge: *factual* and *conceptual* (sometimes also referred to as *declarative* knowledge), *procedural* and *meta-knowledge*. Concept maps visualize and assess knowledge of concepts and their relationships, which can commonly be described as factual knowledge or specific facts (propositions), but they also assess knowledge of connections of facts to other facts at a higher level of abstractness (conceptual knowledge). It should also be possible to represent procedural knowledge or knowledge about procedures required to achieve a goal using cyclic or chain-like concept maps.

Understanding, which is both in the original [62] and revised Bloom's taxonomy of learning outcomes defined using a set of common actions enabling its assessment, could on the other hand be defined as a state defined by the meta-knowledge about concepts enabling their generalization and expansion to include other concepts.

The suggested knowledge framework should therefore be able to provide terminology for concept-map based assessment or learning outcomes. Even if not, there is a need for consistent terminology considering types and attributes of knowledge assessed using concept maps.

VII. CONCLUSION

To conclude, main aspects of concept maps' application in assessment are their construction procedures, which can at least partly be automatized but still for the most part remain a time demanding task for the domain expert, assessment task characteristics, which define how is the assessment task structures, and finally quantitative and qualitative scoring elements and algorithms, which provide support for map assessment. Practical examples of those properties are discussed in the paper.

Although many successful computer-based solutions to concept-map generation and assessments have been proposed, one still open question that hasn't been answered in the literature so far is the common terminology referring to types of knowledge being assessed. It is hoped that the simple suggestion to apply the revised Bloom's taxonomy proposed in this paper would help to resolve this issue.

ACKNOWLEDGMENT

The work described in this paper was conducted under the research project "Complex System Modeling" (036-0362214-1987), supported by the Ministry of Science, Education and Sports of the Republic of Croatia.

REFERENCES

[1] J. Stewart, J. van Kirk, and R. Rowell, "Concept Maps: A Tool for Use in Biology Teaching," *American Biology Teacher*, vol. 41, no. 3, pp. 171–75, Mar. 1979.

[2] J. D. Novak and D. B. Gowin, *Learning how to learn*. Cambridge University Press, 1984.

[3] M. A. Ruiz-Primo and R. J. Shavelson, "Problems and issues in the use of concept maps in science assessment," *Journal of Research in Science Teaching*, vol. 33, no. 6, pp. 569–600, 1996.

[4] J. D. Novak, "Introduction to concept mapping." Cornell University.

[5] R. C. Anderson, R. J. Spiro, and M. C. Anderson, "Schemata as Scaffolding for the Representation of Information in Connected Discourse," *American Educational Research Journal*, vol. 15, no. 3, pp. 433 –440, Jun. 1978.

[6] D. E. Rumelhart, "Schemata: the building blocks of ognition," in *Comprehension and teaching: Research reviews*, In J. Guthrie., International Reading Association, 1982, pp. 3–26.

[7] D. P. Ausubel, J. D. Novak, and H. Hanesian, *Educational psychology: a cognitive view*. Holt, Rinehart and Winston, 1978.

[8] J. D. Novak and A. J. Canas, "The theory underlying concept maps and how to construct and use them," *Technical Report IHMC CmapTools*, 2008.

[9] D. P. Ausubel, L. C. Robbins, and E. Blake Jr., "Retroactive inhibition and facilitation in the learning of school materials," *Journal of Educational Psychology*, vol. 48, no. 6, pp. 334–343, 1957.

[10] V. F. Rossner, "A description and critique of David P. Ausubel's model of learning," Thesis (M.A. (Ed.)), Simon Fraser University, 1982.

[11] J. Hiebert and T. P. Carpenter, "Learning and teaching with understanding," in *Learning and teaching with understanding*, New York: Macmillan., 1992, pp. 65–97.

[12] J. Hiebert and P. Lefevre, "Conceptual and procedural knowledge in mathematics: An introductory analysis," in *Conceptual and procedural knowledge: The case of mathematics*, Hillsdale, NJ: Erlbaum, 1986, pp. 1–28.

[13] J. M. Royer, C. A. Cisero, and M. S. Carlo, "Techniques and Procedures for Assessing Cognitive Skills," *Review of Educational Research*, vol. 63, no. 2, p. 201, 1993.

[14] R. J. Mislevy and D. H. Gitomer, "The role of probability-based inference in an intelligent tutoring system," *User Modeling and User-Adapted Interaction*, vol. 5, no. 3–4, pp. 253–282, 1996.

[15] Novak, "Learning, creating, and using knowledge: Concept Maps as facilitative tools in schools and corporations." Mahwah, NJ: Erlbaum, 1998.

[16] F. Safayeni, N. Derbentseva, and A. J. Cañas, "A theoretical note on concepts and the need for Cyclic Concept Maps," *Journal of Research in Science Teaching*, vol. 42, no. 7, pp. 741–766, 2005.

[17] R. Shavelson, Lang, and Lewin, "Concept maps as potential alternate assessments in Science: CRESST Technical report." .

[18] S. Puntambekar, A. Stylianou, and R. Hübscher, "Improving navigation and learning in hypertext environments with navigable concept maps," *Hum.-Comput. Interact.*, vol. 18, no. 4, pp. 395–428, Dec. 2003.

[19] D. L. Harnisch, T. Sato, P. Zheng, S. Yamaji, and M. Conell, "Concept mapping approach and its implications in instruction and assessmen," *Computers in the Schools*, vol. 5, no. 3, pp. 132–168, 1996.

[20] M. Åhlberg, "Varieties of Concept Mapping," in *Poster presented at The First International Conference on Concept Mapping, Public University of Navarra, Spain*, 2004.

[21] J. J. Villalon and R. A. Calvo, "Concept Map Mining: A Definition and a Framework for Its Evaluation," in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*, Washington, DC, USA, 2008, pp. 357–360.

[22] Q. Yao, K. Yang, G. Zhao, and R. Huang, "A Concept Map Scoring Algorithm Based On Proposition Chains For Concept Mapping," in *Concept Maps: Theory, Methodology, Technology. Proc. of the Second Int. Conference on Concept Mapping*, Universidad de Costa Rica, San José, Costa Rica, 2006.

[23] D. Albert and C. Steiner, "Representing Domain Knowledge by Concept Maps: How to Validate Them?," in *T. Okamoto, D. Albert, T. Honda & F. W. Hesse (Eds.), The 2nd Joint Workshop of Cognition and Learning Through Media-Communication for Advanced e-Learning (JWCL), September 28-30, 2005, Tokyo, Japan.*, 2005, pp. 169–174.

[24] H. . Herl, H. . O'Neil Jr, G. K. W. . Chung, and J. Schacter, "Reliability and validity of a computer-based knowledge mapping system to measure content understanding," *Computers in Human Behavior*, vol. 15, no. 3–4, pp. 315–333, May 1999.

[25] K. Zubrinic, D. Kalpic, and M. Milicevic, "The automatic creation of concept maps from documents written using morphologically rich languages," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12709–12718, Nov. 2012.

[26] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[27] M. S. Binwahlan, N. Salim, and L. Suanmali, "Fuzzy Swarm Based Text Summarization," *Journal of Computer Science*, vol. 5, no. 5, pp. 338–346, May 2009.

[28] R. Clariana and R. Koul, "A computer-based approach for translating text into concept map-like representations," presented at the Proceedings of the First International Conference on Concept Mapping, 2004.

[29] R. Richardson and E. A. Fox, "Using concept maps in digital libraries as a cross-language resource discovery tool," in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, 2005. JCDL '05*, 2005, pp. 256 –257.

[30] S.-M. Chen and S.-M. Bai, "Using data mining techniques to automatically construct concept maps for adaptive learning systems," *Expert Syst. Appl.*, vol. 37, no. 6, pp. 4496–4503, Jun. 2010.

[31] A. Zouaq and R. Nkambou, "Building Domain Ontologies from Text for Educational Purposes," *IEEE Transactions on Learning Technologies*, vol. 1, no. 1, pp. 49 –62, Mar. 2008.

[32] A. Zouaq, D. Gasevic, and M. Hatala, "Ontologizing concept maps using graph theory," in *Proceedings of the 2011 ACM Symposium on Applied Computing*, New York, NY, USA, 2011, pp. 1687–1692.

[33] J. Villalon and R. A. Calvo, "Concept Extraction from Student Essays, Towards Concept Map Mining," in *Ninth IEEE International Conference on Advanced Learning Technologies, 2009. ICALT 2009*, 2009, pp. 221 –225.

[34] A. M. Olney, W. L. Cade, and C. Williams, "Generating concept map exercises from textbooks," in *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, Stroudsburg, PA, USA, 2011, pp. 111–119.

[35] E. M. Taricani and R. B. Clariana, "A Technique for Automatically Scoring Open-Ended Concept Maps," *Educational Technology Research and Development*, vol. 54, pp. 65–82, Feb. 2006.

[36] M. Strautmane, "Concept Map-Based Knowledge Assessment Tasks and Their Scoring Criteria: an Overview," in *publication.editionName*, 2012, vol. 1, pp. 80–89.

[37] C. Schau, N. Mattern, M. Zeilik, K. W. Teague, and R. J. Weber, "Select-and-Fill-in Concept Map Scores as a Measure of Students' Connected Understanding of Science," *Educational and Psychological Measurement*, vol. 61, no. 1, pp. 136–158, 2001.

[38] R. J. Shavelson, M. A. Ruiz-Primo, and E. W. Wiley, "Windows into the Mind," *Higher Education: The International Journal of Higher Education and Educational Planning*, vol. 49, no. 4, pp. 413–430, Jun. 2005.

[39] I. C. Sas, "The Multiple-choice concept map (MCCM): An interactive computer-based assessment method," University of Nevada, Las Vegas, 2010.

[40] K. Plummer, *Concept-Map Assessments: The Reliability and Validity Of Classroom Accessible Concept-Map Assessments*. VDM Verlag Dr. Muller Aktiengesellschaft & Co. KG, 2008.

[41] P.-H. Wu, G.-J. Hwang, M. Milrad, H.-R. Ke, and Y.-M. Huang, "An innovative concept map approach for improving students' learning performance with an instant feedback mechanism," *Br. J. Educ. Technol.*, vol. 43, no. 2, pp. 217–232, Mar. 2012.

[42] E. Gouli, A. Gogoulou, K. Papanikolaou, and K. Grigoriadou, "Compass: An Adaptive Web-Based Concept Map Assessment Tool," in *Proc. of the First Int. Conference on Concept Mapping*, Pamplona, Spain, 2004.

[43] Ş. K. İngeç, "Analysing Concept Maps as an Assessment Tool in Teaching Physics and Comparison with the Achievement Tests," *International Journal of Science Education*, vol. 31, no. 14, pp. 1897–1915, 2009.

[44] M. A. Ruiz-Primo, S. E. Schultz, and R. J. Shavelson, "Concept Map-Based Assessment in Science: Two Exploratory Studies." CSE Technical Report 436, CRESST/Stanford University.

[45] U. Park and R. A. Calvo, "Automatic Concept Map Scoring Framework Using the Semantic Web Technologies," in *Eighth IEEE International Conference on Advanced Learning Technologies, 2008. ICALT '08*, 2008, pp. 238 –240.

[46] B. E. Cline, C. C. Brewster, and R. D. Fell, "A rule-based system for automatically evaluating student concept maps," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2282–2291, Mar. 2010.

[47] T. E. Goldsmith, P. J. Johnson, and W. H. Acton, "Assessing structural knowledge," *Journal of Educational Psychology*, vol. 83, no. 1, pp. 88–96, 1991.

[48] S.-W. Chen, S. C. Lin, and K. E. Chang, "Attributed concept maps: fuzzy integration and fuzzy matching," *Trans. Sys. Man Cyber. Part B*, vol. 31, no. 5, pp. 842–852, Oct. 2001.

[49] K. E. Chang, Y. T. Sung, R. B. Chang, and S. C. Lin, "A new assessment for computer-based concept mapping," *Educ. Technol. Soc.*, vol. 8, no. 3, pp. 138–148, 2005.

[50] J. E. Sims-Knight, R. L. Upchurch, N. Pendergrass, T. Meressi, P. Fortier, P. Tchimev, R. VonderHeide, and M. Page, "Using concept maps to assess design process knowledge," in *Frontiers in Education, 2004. FIE 2004. 34th Annual*, 2004, pp. F1G – 6–10 Vol. 2.

[51] S. Narciss and K. Huth, "Fostering achievement and motivation with bug-related tutoring feedback in a computer-based training for written subtraction," *Learning and Instruction*, vol. 16, no. 4, pp. 310–322, Aug. 2006.

[52] P. Denton, J. Madden, M. Roberts, and P. Rowe, "Students' response to traditional and computer-assisted formative feedback: A comparative case study," *British Journal of Educational Technology*, vol. 39, no. 3, pp. 486–500, 2008.

[53] S. Jordan and T. Mitchell, "e-Assessment for learning? The potential of short-answer free-text questions with tailored feedback," *British Journal of Educational Technology*, vol. 40, no. 2, pp. 371–385, 2009.

[54] K. Chang, Y. Sung, and I. Chen, "The Effect of Concept Mapping to Enhance Text Comprehension and Summarization," *The Journal of Experimental Education*, vol. 71, no. 1, pp. 5–23, 2002.

[55] K. Sanders, J. Boustedt, A. Eckerdal, R. McCartney, J. E. Moström, L. Thomas, and C. Zander, "Student understanding of object-oriented programming as expressed in concept maps," in *Proceedings of the 39th SIGCSE technical symposium on Computer science education*, New York, NY, USA, 2008, pp. 332–336.

[56] C. G. Williams, "Using Concept Maps to Assess Conceptual Knowledge of Function," *Journal for Research in Mathematics Education*, vol. 29, no. 4, p. 414, Jul. 1998.

[57] W. A. Ku, "Using concept maps to explore the conceptual knowledge of technology students: an exploratory study," thesis, Ohio State University, 2007.

[58] E. Hegarty-Hazel and M. Prosser, "Relationship between students' conceptual knowledge and study strategies-part 1: student learning in physics," *International Journal of Science Education*, vol. 13, no. 3, pp. 303–312, 1991.

[59] J. M. Laffey and J. Singer, "Using mapping for cognitive assessment in project-based science," *J. Interact. Learn. Res.*, vol. 8, no. 3–4, pp. 363–387, Dec. 1997.

[60] T. de Jong and M. G. M. Ferguson-Hessler, "Types and qualities of knowledge," *Educational Psychologist*, vol. 31, no. 2, pp. 105–113, 1996.

[61] D. R. Krathwohl, "A Revision of Bloom's Taxonomy: An Overview," *Theory into practice*, vol. 41, no. 4, Autumn 2002.

[62] B. S. Bloom, *Taxonomy of educational objectives: the classification of educational goals*. Longmans, Green, 1956.

Tracking the Roots of Plagiarism
Results of the initial International Survey

Duško Lukač, Dr.
Rheinische Fachhochschule Köln gGmbH-University of Applied Sciences
Vogelsanger Strasse 295, 50825 Köln, Germany
Phone: +49-221-5468759 Fax: +49-221-5468755 E-mail: lukac@rfh-koeln.de

Marina Čicin-Šain, Ph.D., Prof.
Faculty of Economics, University of Rijeka
Ivana Filipovića 4, 51000 Rijeka
Phone: +385 51 355 111 Fax: +385 51 212 268 E-mail: marinacs@efri.hr

Abstract - **In already published joint papers written by the Rheinischen Fachhochschule Köln - University of to Applied Sciences and the University of Rijeka the reasons for the intellectual fraud were analyzed. This study about the intellectual fraud becomes extended by inclusion of other international partners which are contributing to the study by including the own national and organizational view to the subject. In this paper new results, which give an overview to the national legislation regarding the intellectual fraud policy in Austria, Serbia, Croatia and Germany is presented, as well as to the institutional (university) position to the subject of the intellectual fraud is presented. Moreover, the institutions of the project partners with regard to the university's plagiarism policy and honor code regulations as well as the national law regulations are explored.**

I. INTRODUCTION

The first step in this research is to get institutional and national insights into the subject from the individual perspective of the participants. In the later step of the project more comprehensive details from the participants are needed in order to take into consideration also individual aspect of the roots for academic more dishonesty. In previous publication we took a deeper insight into the topic of the intellectual fraud and the definitions of the academic dishonesty as for instance "Plagiarism (non academically acknowledged presentation of another person's work as the own one, collusion (collaborative plagiarism), falsification (presentation of the untrue data, Cheating (not allowed obtaining or giving the support and assistance or information), Deceit (resubmitting of the own previously assessed work), Personation (supposition of the characteristics of another person with intent to mislead or gain inequitable benefit)" [1, p. 1156]. Thereby we mentioned different examples of academic dishonesty by getting the overview of intellectual fraud in Croatia and Germany in the past few years and also we analyzed both cultures by using of methods of cross-cultural theories. From the German side, the trend of the discovering of the academic "wrongdoer" unfortunately keeps to be sustainable also in 2013. Actual topic in the German media is the case of the academic dishonesty among the public people, this time nobody else but the German Minister of Education Annette Schavan. The University of Düsseldorf has her doctoral degree she got 20 years ago denied and annulated. As stated by Schmoll [2, p.1], the faculty council of the university of Dusseldorf spoke about the „intentional deception by plagiarism" with her thesis. After consultation 12 members voted to for the denial and two against it, one of 15 eligible to vote was abstained. Bases for the consultations were the preliminary report of the promotion committee as well as the statement Mrs. Schavans with two enclosed education-scientific statements. It is to be stuck that the reference standards at the faculty of education science, where she wrote the thesis, were the identical ones like in the remaining faculty of philosophy to the time of origin point of the work. The frequency and construction of literal takeovers, also the non-mention of literature titles in footnotes or even in the literature list would have proved after conviction of the faculty council the general view, „that the doctoral candidate at that time, distributed gave mental performances systematically and intentionally about the complete thesis which she had not produced in reality." After University of Düsseldorf publicized the results, she will complain using the all juridical means and will keep her position as Minister. But 3 days after, she announced that she will leave the position of the minister [3]. It is a 5th case of plagiarism among the German politicians in the last 3 years. It is a similar development like in other European countries where many cases of plagiarism occurred. It supports the necessity for tracking the roots for academic dishonesty we try to analyze in this project.

II. METHODOLOGY OF THE RESEARCH

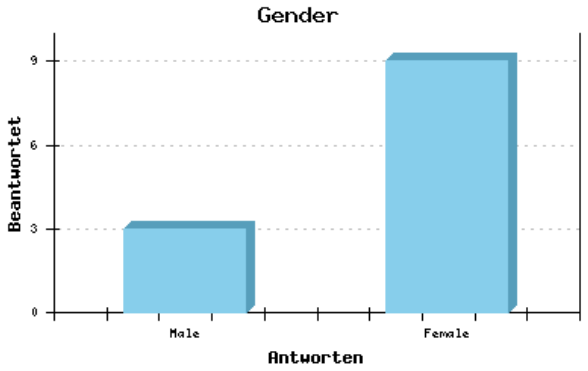
The project is currently supported by persons working at institutions in those countries: Croatia, Germany, Austria, Serbia, Macedonia, Greece, USA and India. The institutions currently involved in the project are University in Rijeka (Croatia), University of Delhi (India), University of Applied Sciences Upper Austria (Austria), Flinders University in Adelaide, South Australia (Australia), Grand Valley State University (USA), Rheinische Fachhochule Köln - University of Applied Sciences (Germany), Deutsches Jugendinstitut e.V. (Germany), University of Novi Sad (Serbia), University Ss. Cyril and Methodius in Skopje (FYR Macedonia), University of Xanthi (Greece), Georgia Southern University (USA), University of Zagreb (Croatia). The project should be carried out in different phases including the national, organizational and individual analysis in order to get deeply insight into the reasons for different attitudes regarding academic dishonesty. In initial step of the project, we carried out the survey consisting out of 12 questions, which are

giving the answers about the gender, nation, status (lecturer, student), using of the anti-plagiarism software, obligation to check the thesis or student papers concerning the plagiarism, national law/penalty regulations regarding academic dishonesty, existence of honor code at the universities as well as the accessibility of the information necessary to the lecturers or students in order to be appropriately informed about the topic of the academic dishonesty.

III. ANALYSIS OF THE INITIAL RESULTS

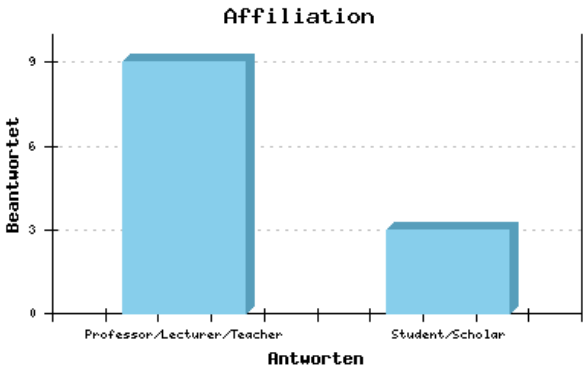
In this part the results of the analysis will be presented and discussed. Here only the overall data as a absolutely value will be presented. The correlation analysis and deeper possible relationships between the variables will be done in further step of research. This especially because for the reason that for reliable correlation analysis the further collection of the data is necessary. The survey can be accesses under [4]. It is anonymous survey which requires so called individual Survey Accessed Code and it is carried out online.

1. Question - What is your gender?



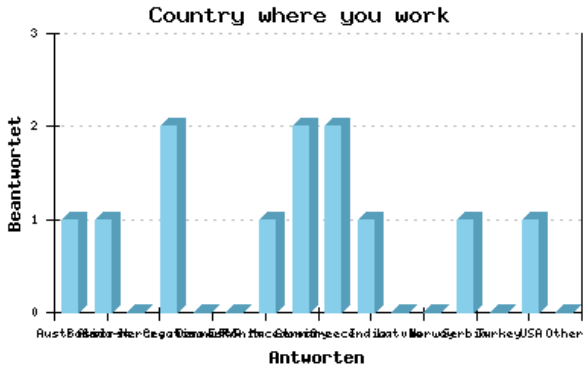
Most of the survey participants are male. This information is initially only for statistical purposes. Further inclusion of the gender impact into analysis, as for instance impact of feminine characteristics as e.g. caring for others and keeping the sustainable quality of life or masculine characteristic as tendency to competition, achievement and success will be carried out within the scope of cross-cultural and individual analysis.

2. Question - What is your professional status?



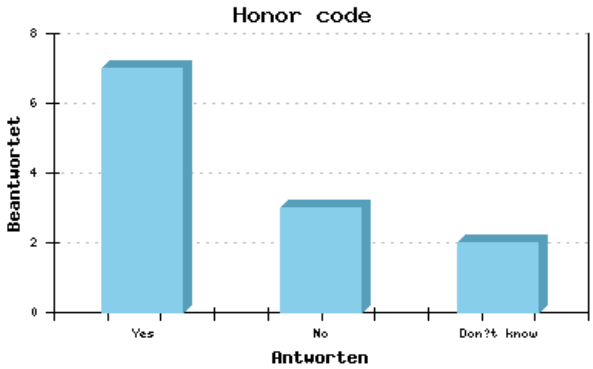
The most of the respondents are belonging to the group of the professors/lectures/teachers. The correlation e.g. between the affiliation and being fully and understandably clarified about the academic dishonesty policy at the own institutions, or other possible relationships will be analyzed.

3. Question - In which one of the following countries are you working (or studying) at the moment?



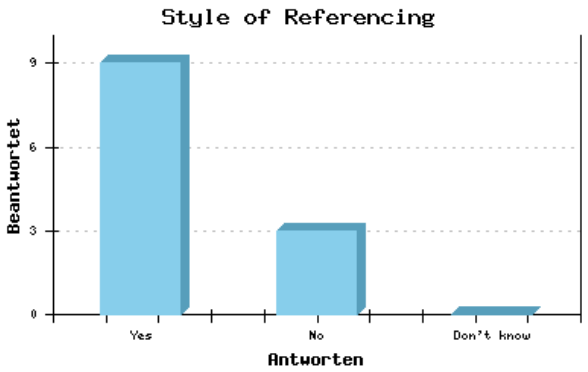
According the survey respondents from following have been participating (number of the participants is shown in the brackets): Australia (1), Austria (1), Croatia (2), FYR Macedonia (1), Germany (2), Greece (2), India (1), Serbia (1) and USA (1). By using of means of the cross-cultural analysis related to the national culture, further correlations will be investigated.

4. Question - Does at your educational institution/university so called "Honor code" exist? Honor code is most commonly used to deter (prevent) academic dishonesty.



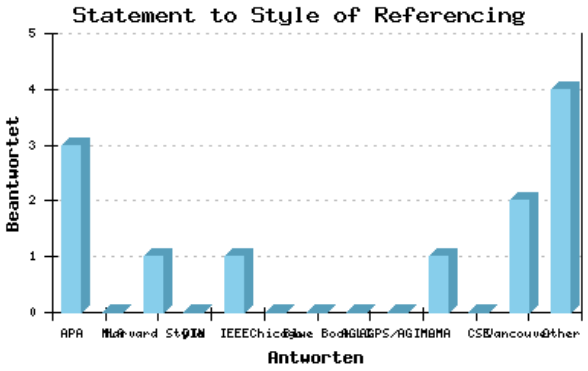
At the 7 educational institutions so called Honor code exist. Even at 3 institutions the Honor code does not exist, and 2 respondents are not informed about, meaning that possible insufficient application of the eventually existing honor code can be assumed. An honor code is a code of behavior or principled standards leading a group of people or institution based on ideals that characterize what comprises admirable manners within that group of people or institution. Such Honor codes are since 1779 typically used in the USA and most commonly used in the Anglo-Saxon countries to prevent academic dishonesty. In Europe it is still not widespread.

5. Question - Does at your educational institution/university is defined the Style of Referencing (e.g. Harvard, APA, IEEE, DIN etc.) the students/scholars have to use when writing the academic papers, thesis or seminar papers?



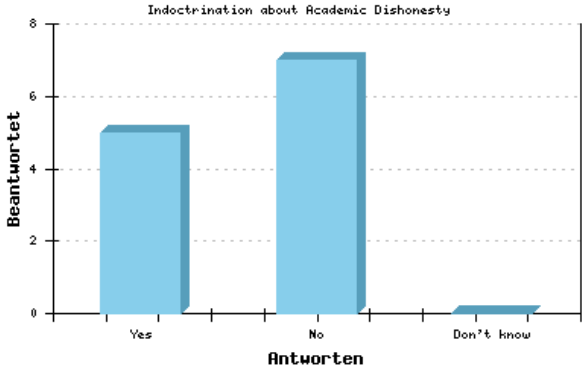
At most institutions already defined Style of Referencing exists. At three universities it doesn't exist at all. It can be assumed at the moment, that at those institutions clear rules to deter academic dishonesty are not defined and therefore also not applied.

6. Question - If yes, which Style of Referencing is used at your educational institution/university?



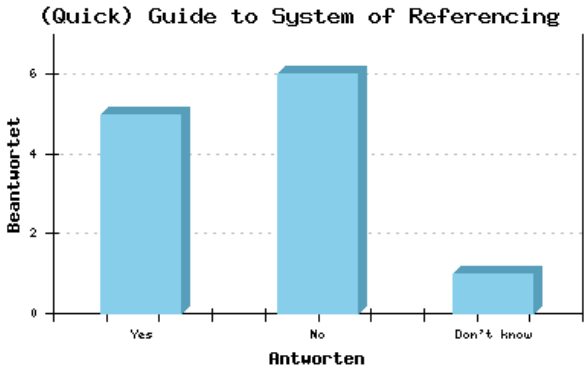
The most style of referencing used is APA, followed by Vancouver style. Here is to be mentioned that the answer "Other" was used for those which are not using any style of referencing at their institutions because it doesn't exist or the name of the style of referencing is not listed in the list of the possible answers.

7. Question - Are the students/scholars at your educational institution/university, well advised or accurately informed about the subject of academic dishonesty and about the consequences in the case of academic fraud?



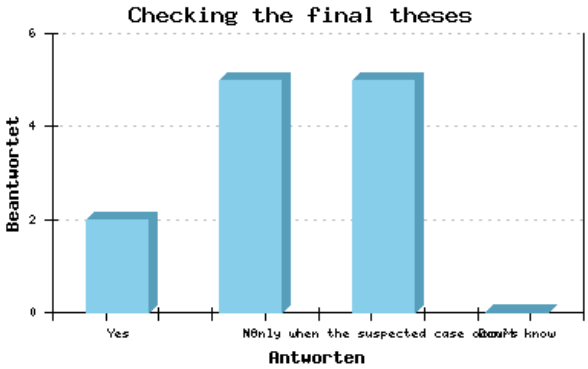
At more as a half institutions the students are not well and accurately advised or informed about the subject of academic dishonesty. It is a disappointing fact showing the discrepancy between the rules which are may be existent at the institutions but are not applied. If 7 institutions of all responds are have honor code it means that at least at 2 this is just formal existence which is not functional in real.

8. Question - Does at the educational institution where you work some kind of (Quick) Guide to System of Referencing is defined and does everybody can access/download this Guide?



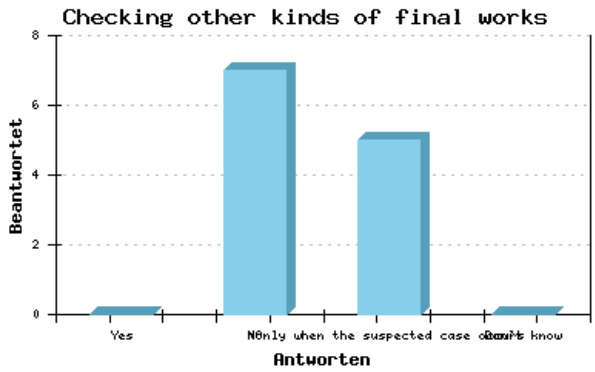
The answers correspond to the answers to the Question Nr. 7 about how the students are advised and informed about the matter of academic dishonesty. It seems to be that at some institutions honor code exists but there is no practical description, available for all, of it.

9. Question - Does at your educational institution/university, final thesis (Bachelor, Diploma, Master, PhD) is obligatory checked on plagiarism by using the anti-plagiarism software?



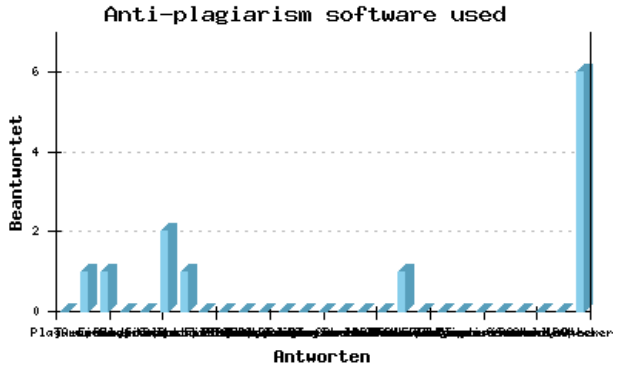
Very interesting answers are given to the Question Nr.9. Only at 2 educational institutions final theses becomes checked. This means that most of the institutions are relaying on the existing honor code and good will of the students to apply it, or most of the institutions may be advice the students about the problem of academic dishonesty but do not show the means how they avoid it in practical case. In such cases it can be implicated that students will try to use such unstrategic gap in negative way, no matter what for reasons are behind of such movements.

10. Question - Does at the educational institution at you work, other kinds of final works (final assignment, seminar paper, student research paper) are obligatory checked on plagiarism by using the anti-plagiarism software?



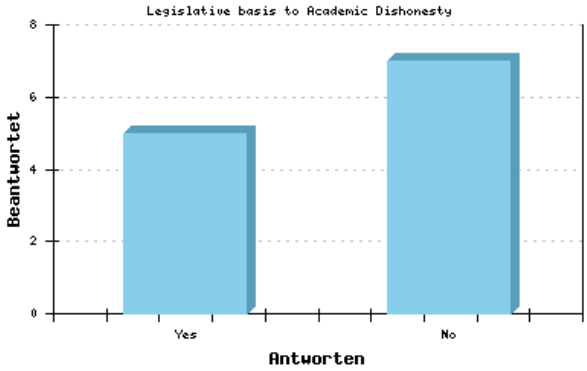
There is no obligatory check of the final student works at any of participating institutions. It is partly surprising result, but also not supporting fact for avoiding of academic dishonesty. At least 7 institutions are checking the final works for the suspected case and 5 institutions are letting the writing of the final works without any check. The question one can ask is how can really be expected from the students, to apply the rules for academic honesty with writing the final theses, if they haven't been proved by doing that, while writing different final works during the study?

11. Question - If at the educational institution at you work anti-plagiarism software is used, please give the exact name of anti-plagiarism software used? The name of anti-plagiarism software is...



Here is to be mentioned that the answer "Other" is used for the case that no anti-plagiarism software is used or the software used is not listed. According to the results the most of the institutions (6) are not using the anti-plagiarism software. It means that at least 1 institution checks the final thesis without the software. On the other side, is has to be mentioned according to [1] that event the best rated software can discover about 60% of the cases of academic dishonesty. The most used software among the respondents is Plagiarism Finder produced by one German company, followed by Turnitin, Ephorus, Docoloc and GenuineText . This software is according to the test made by University for Techniques and Economy in Berlin [5, p.1] rated as "almost insufficient" for use in the education. The software as PlagAware, Turnitin, Ephorus, PlagScan and Urkund are best rated but only with the mark "partly usable systems" [5, p.1].

12. Question - Is in your country academic dishonesty according to national law declared as severe offense (criminal offence)? (Please check the legislative basis (law background) to this question carefully and in the case of doubt please ask somebody else involved in the topic to support you in order to get correctly answer)



According to the answers of the participants in 7 of 12 participating countries there is no legislative basis which declares academic dishonesty as severe offense (criminal offence). It means that there is no a sufficient legal base, or even legal interest to treat academic dishonesty as serve offence. In further study the correlation between the countries with no sufficient legal base and the high rate of academic dishonesty will be investigated.

IV. CONCLUSION

In this study we carried out initial research on the topic of academic dishonesty. We got overall data from different countries getting answers to the basic institutional questions about the practical protection against academic dishonesty. In further step, more respondents each nation has to be involved and further questioner need to be developed. Correlation analysis based on existing answers needs to be carried out. This will be prepared in the next step of joint research.

V. REFERENCES

[1] Lukac, D and Cicin-Sain, M. (2012) "Cultural influence on the intellectual honesty and fairness of the German and Croatian students", Page(s): 1156 - 1160, 2012 Proceedings of the 35th International Convention MIPRO, IEEE Xplore,

[2] Schmoll, H (2013) "Schavans Doktortitel aberkannt" [Accessed: 05/02/13] http://www.faz.net/aktuell/politik/inland/plagiatsverfahren-schavans-doktortitel-aberkannt-12052730.html, 17:22h

[3] Zeitonline (2013) „Bildungsministerin Schavan tritt zurück„ [Accessed: 09/02/13] http://www.zeit.de/politik/deutschland/2013-02/ruecktritt-schavan-plagiat, 16:01h

[4] International survey supporting the investigation of the reasons for the intellectual fraud (2013) "Tracking the Roots of Plagiarism" [Accessed: 13/02/13] http://ilias.rz.rfh-koeln.de/ilias.php?baseClass=ilObjSurveyGUI&ref_id=213&cmd=infoScreen&lang=en, 15:53h

[5] HTW Berlin (2013) „Softwaretest 2010", [Accessed: 13/02/13] http://plagiat.htw-berlin.de/software/2010-2/, 15:42h

Considerations on the evolution of attention paid to education in Romania compared to EU, in the 1998-2010 period

M. Zaharia, D. Enachescu
"Petrol-Gaze" University/Modeling, Economic Analysis and Statistics Department, Ploiesti, Romania
marianzaharia53@gmail.com
denachescu22@yahoo.com

Abstract - The paper analyzes the evolution of attention paid to education in Romania, in the period after 2000. It shows the evolution of indicators such as expenditure on education as % of GDP, the index of total public expenditure on education, net investments in education, the percentage of computer use at place of education by all individuals, by individuals, 16 to 24 years old and by unemployed also.

To highlight the attention paid to education and the evolution of computer use in education in Romania, in the analyzed period, within the paper are presented comparisons with the analyzed indicators evolution in EU and European countries. The main conclusion is that, in the analyzed period, Romania has made significant efforts to bring education to European standards, about 70% Percentage growth of computer use at place of education by Individuals, 16 to 24 years old, from this period being determined by the increase in education expenditure

I. INTRODUCTION

Economic development and social progress in the European Union, and not only, depend on the training and education of all citizens. In this direction, attention to education at all levels has an essential role.

Before 1989, Romania was given special attention to education both technical equipment and professional quality of the teachers. This is evidenced by the large number of specialists and the quality of research in that time

Unfortunately, in 1990-2000 the governments concern for education was very low, which resulted in the backwardness of this very important activity.

After 2000, Romania started, the first few years shy, to narrow the gap by upgrading and computerization of the educational process.

During the recent years, governments and educational institutions are increasingly concerned with ensuring the conditions for widespread use of IT and C, endowment of education with the most modern techniques, as data mining, as a way to improve the quality of education[1].

The paper focuses specifically on the recovering process of the Romanian education at the EU level, a process that is ongoing.

For performing studies and the development work were used available Eurostat data series [2] and those published by Romanian National Institute of Statistics [3].

II. THE DINAMIC AND THE STRUCTURE EVOLUTION OF EXPENDITURE ON EDUCATION IN 2001-2009 PERIOD

A. The evolution of expenditure on education as percentage in GDP

The expenditure on education as% in GDP is an important indicator that shows the attitude of the State towards education and the place it occupies in comparison with other states. For these reasons, a first comparative analysis focuses on the development expenditure on education as% of Romania's GDP compared to other countries.

To have the image of Romania's position in all European countries in this regard, between European countries, was chosen Sweden, as representative of countries issuing the highest percentage of GDP on education, Bulgaria and Hungary, Romania neighboring countries and Croatia (figure 1).

EU average expenditure on education as a percentage of GDP has fluctuated between a minimum of 4.95% in 2007 to a peak of 5.41 in 2009. Sweden was above average, with levels ranging from 6.61% in 2007 and 7.36% in 2002, and Bulgaria and Croatia have evolved slightly upward on a trend approaching the EU average. In Croatia, the expenditure on education as a percentage of GDP, increased almost linearly from 3.71% in 2002 to 4.33% in 2009. In Bulgaria, however, the evolution of this indicator was more oscillating.

Regarding Romania, during 2001-2009, except 2007, it was at the lowest level of the percentage of GDP allocated to education. However, it is worth noting that if in 2001 it was 3.25% of GDP (by 1.74 percentage points below the EU average) in 2009 it reached 4.24% of GDP (by 1.17 percentage points below the EU average).

As results from Figure 1, particular attention was given to education by the governments of Romania in the

period 2004-2007. During this period the percentage of GDP allocated to education increased from 3.28% to 4.25%. Overall, except for the years 2003 and 2004, in Romania there is concern for the approximation of the average percentage of GDP allocated to education in the EU.

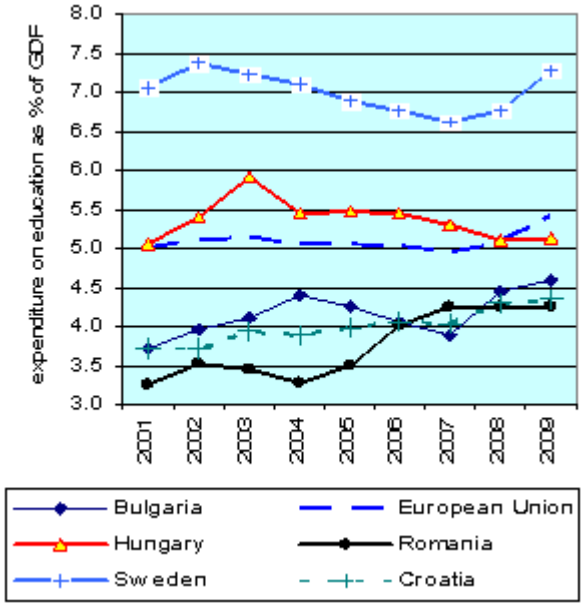


Figure 1 The evolution of expenditure on education as % of GDP in Romania and other fore European countries

However, the percentage of GDP allocated to education is an indicator that provides only an overview regarding the attention of state given to this activity. Moreover, due to significant differences between levels developed, on the one hand and the size of GDP on the other hand, comparisons between countries gives us only information on government policies in relation to education, and only to a limited extent of actual expenditure for education.

B. The evolution of the index of total public expenditure on education

For a clearer image on volume of expenditure on education carried in Romania in Figure 2 shows the evolution of total public expenditure index on education in constant prices (2000 = 100) compared with Bulgaria, Sweden and the EU average.

The situation shown in Figure 2 differs significantly from that described above (Figure 1). This highlights the concern to bring the Romanian education to European standards. Although Romania is the last place on the percentage of expenditure on education in GDP, the effort made is significant.

Thus, the annual rate of total public expenditure on education at EU level has evolved from a maximum of 4.54% in 2001 and a minimum of 1.08% in 2004, the average for the entire period being 2.63%. In Sweden the average rate of total public expenditure on education was 1.85% (by 0.78 percentage points below the EU average). In Bulgaria annual rate of total public expenditure on education has evolved from -0.62% in 2001 to 14.78% in 2004 to 1.31% in 2006 to 21.3% in 2008 and falling below zero (-2.41%) in 2009. Throughout the period, the average rate of total public expenditure on education in Bulgaria is of 6.69%.

In Romania the average rate of total public expenditure on education, in analyzed period, was 9.77%. In 2001-2008 the annual growth rates of total public expenditure on education was strictly positive above average in 2001, 2002, 2006 and 2007 (the maximum being 19.94% in 2001) and with values below average in 2004, 2005 and 2008 (minimum value being 3.20% in 2005).

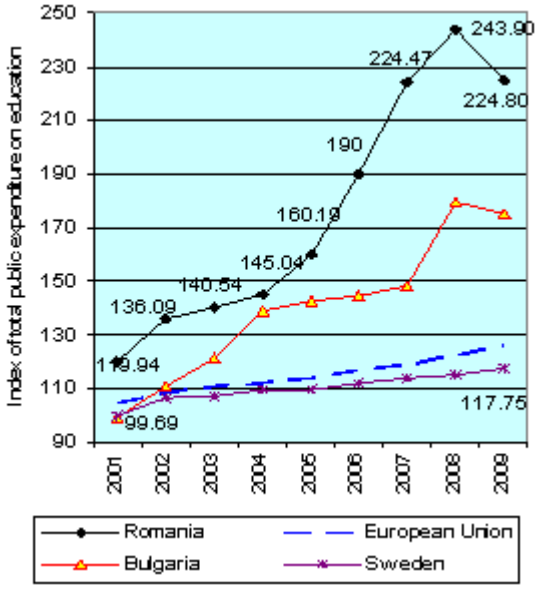


Figure 2 The evolution of the index of total public expenditure on education in constant prices (2000 = 100), for all levels of education combined

The economic crisis started in 2008 had implications also for expenditure on education, the growth registering the only negative value of the analyzed period (-7.83%). It should be noted, however, that the volume of expenditure on education has not fallen below its level in 2007, being the second volume of expenditure throughout the period.

C. Education and the net investment structure evolution by activity of national economy

In the period 2000-2009, as shown in the Figure 2, the total public expenditure on education (expressed in 2000 prices) increased by approximately 2.5 times. These expenses targeted both investment as well as staff salaries in this sector, with a series of other current expenditure.

Given, however, that we want to highlight the impact of policies implemented by governments that ruled Romania in 2001-2010 on the extent of the use of computers in educational processes in the following we will briefly review the evolution and place of education in total net investment allocated to principal activities of the national economy of Romania.

As can be seen from Table I, along the first decade of the 21st century, governments' attitude towards education has changed significantly. Thus, in 2001-2005 the net

investment in education accounted for less than 1% in total. In 2001, education was the second smallest (last place is occupied by health), from 2002 to 2005 it will occupy the last place with allocations between 0.35% and 0.85% of total net investments in Romania.

TABLE I. NET INVESTNENT , BY ACTIVITY OF NATIONAL ECONOMY (%)

Activity of national economy	Year		
	2001	2005	2008
Agriculture, hunting and forestry	6.36	3.14	3.36
Industry	40.03	36.99	31.78
Construction	6.46	9.54	10.91
Trade	10.73	15.33	14.51
Hotels and restaurants	1.34	1.96	2.00
Transport, storage and communications	24.16	10.88	12.75
Financial intermediation	2.72	3.51	1.39
Public administration and defense	3.04	3.17	6.21
Education	0.35	0.36	2.10
Health and social assistance	0.15	2.36	1.35
Other activities	4.67	12.75	13.64

Data source: http://www.insse.ro/cms/rw/pages/index.en.do.

After the 2004 general elections, the governments attitude towards education records a change and since 2006 the percentage of net investments exceeds 1% (1.40% in 2006) amounting in 2008 to 2.01% in total net investment. From this point of view, for education, 2008 was the best year, the share of net investment in the sector surpassing the percentages allocated to 'Health and social assistance "(1.35%)," Financial intermediation "(1.39%) and" Hotels and restaurants "(2.0%).

Finally, if we analyze the evolution of net investments allocated to education compared to the evolution of total net investment activities in the national economy, expressed in constant prices (2000 = 100) we note that between them are significant differences, net investment in education recording a dynamic almost sinuous (figure 3).

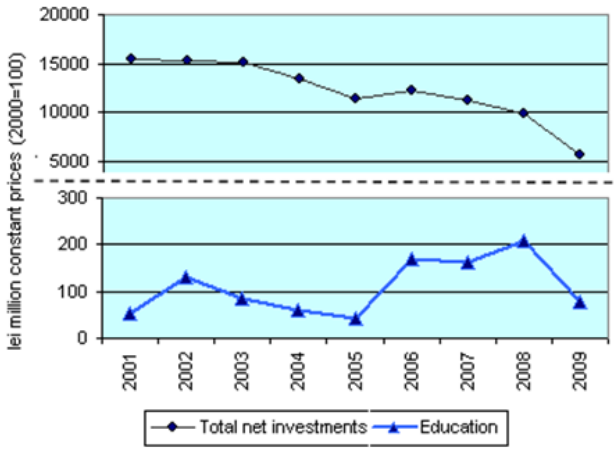


Figure 3 The evolution of total net and education investments

The attention to modernize Romanian education in 2006-2008 is evident. Thus, although the total net investment in the national economy (expressed in 2000 prices) declined from 12.222 billion lei in 2006 to 11.104 billion lei in 2007 and, respectively, 9.865 billion lei in 2008, the net investments in education increased from 41 million in 2005 to 171 million in 2006, respectively, to 207 million in 2008.

Economic crisis unleashed, in late 2008, had an impact on both the net investment volume and the percentage allocated to education in total net investment in the national economy. In 2009, total net investment in the national economy (expressed in 2000 prices) decreased by 43.44% compared to the previous year. The impact on education is higher, net investment in this sector decreased by 61.72%. These had the directly effect of reducing the percentage of net investments in education to 1.42%. As an indirect effect, it was the change of the place of education in relation to other activities of national economy. Education reaches the penultimate place being surpassed of "Health and social assistance" (1.70%), and "Hotels and restaurants" (1.81%).

As is also known, in an economic recession, the education is one of the most affected Romania's activity, and not only.

III. CONSIDERATIONS OF THE COMPUTER USE AT PLACE OF EDUCATION IN ROMANIA IN 2004-2011 PERIOD

A. Aspects of computer use in education by all individuals

The first aspect on which we stop in this chapter is the evolution of Romania's place in relation to other European countries and the EU average regarding to computer use at place of education by all individuals.

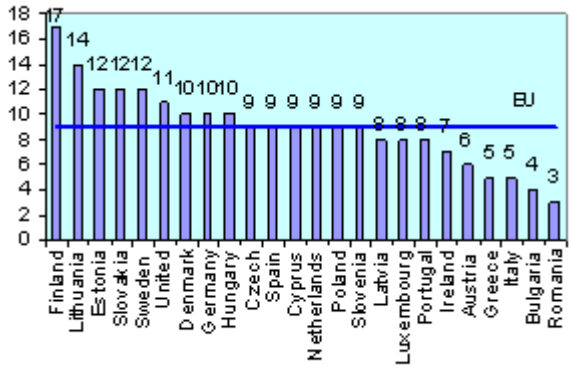


Figure 4 The place of Romania and the percentage of computer use at place of education by all individuals in 2004

As can be seen (Figure 4), of the countries considered, Romania was in 2004, at this indicator also, still on the last place. Below the EU average (9%) and far enough away from it, was still Bulgaria (4.0%), Italy (5.0%) and Grcia (5.0%). Among the countries included in this analysis, in 2004, well above the European average was situated Finland (17.0%) and Lithuania (14.0%).

If benchmarking the situations in Figures 1 and 4, leads to the conclusion that a significant percentage of GDP allocation for education, especially for investments in this sector of national economy lead to the increasing of the opportunities to learn and skills to use the computer, in these activities. Thus, Sweden, as shown in Figure 1, which allocated around 7% of GDP on education, register a percentage of computer use at place of education by all individuals in 2004 by 12%. Also, Hungary in which the percentage of GDP allocated to education, in the analyzed period is greater than or very close to the EU average, recorded in 2004 a percentage of computer use at place of education by all Individuals of 10%.

On the other hand, the efforts made by Romania and Bulgaria to bring education to EU standards outlined in the previous chapter (Figure 2 and Figure 3 for Romania) resulted in a significant increase of percentage of computer use at place of education by all individuals . The situation, for the year 2008, is shown in Figure 5.

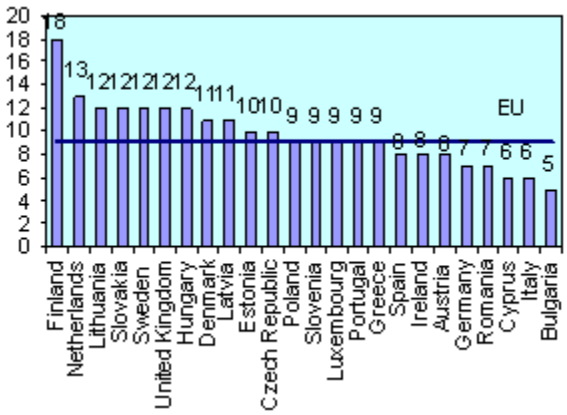


Figure 5 The place of Romania and the percentage of computer use at place of education by all individuals in 2008

The fact that in 2001-2008 the index of total public expenditure on education (Figure 2) in Romania exceeded the values from Bulgaria, has determined that, in 2008, Romania to reach the percentage of computer use at place of education by all individuals, recorded in Germany (7%), ahead of Cyprus (6%), Italy (6%) and Bulgaria (5%).

In the period 2009-2011 this indicator, in Romania had no significant changes.

B. Computer use in education by individuals, 16 to 24 years old

For the population in the age group 16-24 years, the use of computers in education differ significantly (Figure 6). Thus, if in 2004 the percentage of computer use in education by all individuals was only 3%, in the case of individuals, 16 to 24 years old , the percentage of computer use in education was 16% (5 times higher). This is due to the fact that for the population in that age group is almost entirely involved in secondary and tertiary education, the computer use is a necessity.

But unlike tertiary education, in schools in which takes place secondary education, endowment with computers and familiarity of its use has evolved more slowly. Analyzing the evolution of expenditure on education as % of GDP in Romania in Romania (Table 1), but also the evolution of net investment in education (Figure 3) in 2000-2004, we believe that these are the main causes of the low level of percentage of computer use at place of education by individuals, 16 to 24 years old, in 2004.

Policies implemented by governments in Romania since 2004, issues already outlined above, made the situation to be improved so that the percentage of computer use at place of education by Individuals, 16 to 24 years old in Romania to grow almost continuously.

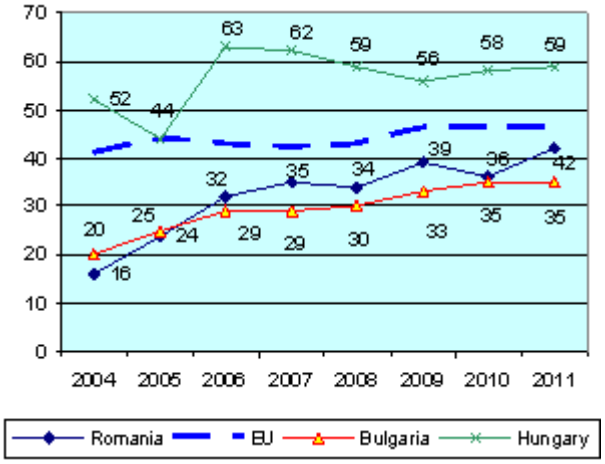


Figure 6 The percentage of computer use at place of education by individuals, 16 to 24 years old in Romania, Bulgaria and Hungary in 2004 – 2011 period

Thus, while the EU Percentage of computer use at place of education by Individuals, 16 to 24 years old ranged between 42% and 46% in Romania has risen from 16% in 2004 to 42% in 2011 (a significant increase of 2.62 times), being only 4 percentage points of EU average.

Following similar policies of Rumania on public expenditure on education (Figure 2), but at a lower average annual rate (6.69% vs. 9.77% in Romania), in Bulgaria the percentage of computer use at the place of education by Individuals, 16 to 24 years old increased from 20% in 2004 to 35% in 2011.

The conclusion on impact of total public expenditure on education on the percentage of computer use at place of education by Individuals, 16 to 24 years old is strengthened by the evolution of these indicators in Hungary, where it recorded values around an average of about 60%. This can be attributed to the fact that, in this country , in the period 2001-2009, the average annual growth rate of public expenditure on education was 3.62% (but with large fluctuations positive and negative).

C. Computer use in education by unemployed

The mutations occurring in the structure of the main occupations required within national economies of most developed countries, and not only had and have the effect of reducing or even disappearance of some jobs and emergence of new ones. Under these conditions the pressure on the labor and especially of the active

population who completed his studies is significant. To adapt to new conditions, continuous training and retraining is a necessity. Computer use in such activities is absolutely necessary.

The unemployed attitude, from different countries, on computer training activities differ significantly.

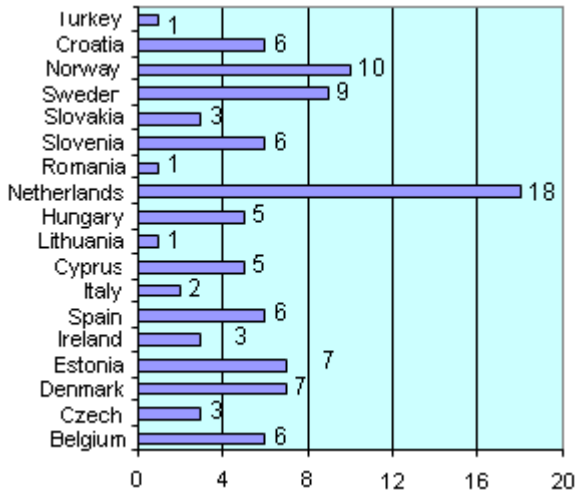


Figure The percentage of computer use at place of education by unemployed in 2010

Among the countries considered in this comparative analysis, Netherlands is on the first place with a percentage of computer use at place of education by unemployed, in 2010, of 18%, followed by Norway (10%) and Sweden (9%).

Also above the EU average (4%) stood Denmark and Estonia (7%), Belgium, Slovenia, Spain and Croatia (6%) and Hungary and Cyprus (5%).

Note that, after the EU enlargement, this indicator decreased significantly from 8% in 2005 (EU-15 countries) to 3% in 2009 (EU-27 countries). Compared to 2009, in 2010 there is a slight reversion to 4%.

Regarding Romania, the percentage of computer use at place of education by unemployed in 2010 was only 1%. This result shows a low interest of unemployed for retraining, and low ability to use a computer, knowing that the highest unemployment rates is recorded among people with a low educational level (primary or secondary).

IV. A Breef Analzsis of Corelattions Between Computer Use and Expenditure on Education

Although the amount of data available is relatively small, in this short chapter is done an assessment of the influence of some indicators used above on the evolution of the percentage of computer use at place of education in Romania in the analyzed period.

To identify the extent to which education expenditures affect the percentage of computer use at place of education had chosen segment of the population consisting of Individuals 16 to 24 years old (CU16_24).

Also to highlight the attention to education in Romania we used two indicators: expenditure on education as% of GDP (EE% GDP) and index of total public expenditure on education in constant prices (2000 = 100), for all levels of education combined (ITPEE).

TABLE II. Linear Corelation Coeficients

	EE%GDP	***CU16_24***	***ITPE***
EE%GDP	1		
CU16_24	0.577084	1	
ITPE	0.893658	0.852277	1

Linear correlation coefficient values are presented in Table II. Of these we are interested in the correlation coefficient between CU16_24 and EE% GDP ($r_{CU16_24/EE\%GDP} = 0.577084$) and that of CU16_24 and ITPEE ($r_{CU16_24/ITPE} = 0.852277$)

To test the significance of determined correlation coefficients, was used "t test" [4] for significance level $\alpha = 0.05$.

The hypotheses for the $r_{CU16_24/EE\%GDP}$ coefficient significance testing are:

$$H_0: \quad r_{CU16_24/EE\%GDF} = 0$$
$$H_1: \quad r_{CU16_24/EE\%GDF} \neq 0$$

For the chosen significance level we obtain $t_c = 1.58 < t_{0.025;5} = 2.57$ and therefore it is accepted the H_0 hypothesis. In conclusion the value of the $r_{CU16_24/EE\%GDP}$ coefficient is not statistically significant and it can not be used.

The hypotheses for the $r_{CU16_24/ITPE}$ coefficient significance testing are:

$$H_0: \quad r_{CU16_24/ITPE} = 0$$
$$H_1: \quad r_{CU16_24/ITPE} \neq 0$$

In this case we obtain $t_c = 3.64 > t_{0.025;5} = 2.57$ and therefore H_0 hypothesis is rejected and H_1 accepted. In conclusion the $r_{CU16_24/ITPE}$ coefficient is statistically significant.

Taking into account that in case of linear regression, correlation coefficient (r) is equal to the ratio of correlation (R), the value of the coefficient of determination is $R^2 = 0.7263$. It follows that the increase of the Percentage of computer use at place of education by Individuals, 16 to 24 years old in Romania was affected in only approximately 72% by the increase in the index of total public expenditure on education, and the rest (28%) by the other factors among which should be mentioned computer skills and quality of education.

V. Conlusion

Economic development and social progress of a country depend on the training and education of all

Citizens. If before 1989 in Romania is given special attention to education, during 1990-2000 the governments concern for education was very low, this has led to the backwardness of this sector. After the 2000, Romania started bridge the gap and to build an education at the European level.

Although Romania is on one the last places on the percentage of expenditure on education as% of GDP, the effort made after 2004 beeing significant. Thus, although the total net investment in the national economy (expressed in 2000 prices) declined from 12.222 billion lei in 2006 to 9.865 billion lei in 2008, the net investments in education increased from 41 million in 2005, to 207 million in 2008.

As a result of the policies implemented by governments in Romania after 2004, the Percentage of computer use at place of education by Individuals, 16 to 24 years old increased almost continuously. If in EU, it has oscillated between 42% and 46%, in Romania has risen from 16% in 2004 to 42% in 2011 (a significant increase of 2.62 times), being only 4 percentage points of the EU average.

On the other hand, we are witnessing yet at a very low interest of unemployed for IT use in professional retraining. The percentage of computer use at place of education by unemployed in 2010 was only 1%. Consequently, are required concrete actions to increase the interest and the skills of computer use in retraining activities.

In Romania, the increase in total public expenditure on education had only a partially influence on percentage of computer use at place of education by Individuals, 16 to 24 years old. To get to the European standards, special attention should be given to education of individuals on the benefits of using IT in continuous learning and increase their level of professional qualification.

REFERENCES

[1] C. Oprea, M. Zaharia**,** "*Using data mining methods in knowledge management in educational field*", in ANNALS of the ORADEA UNIVERSITY. Fascicle of Management and Technological Engineering, Volume X (XX), 2011, NR1, pg. 5311-5320, ISSN: 1583-0691

[2] http://epp.eurostat.ec.europa.eu/portal/page/portal/information_society/data/database.

[3] http://www.insse.ro/cms/rw/pages/index.en.do.

[4] C. Oprea and M. Zaharia, "Elemente de analiza datelor si modelare utilizand Excel" Ed. Universitara, Bucuresti, 2011, pp. 102-106.

Synergy of e-science and libraries as a support to higher education institutions in Croatia

Radovan Vrana
Department for information and communication sciences
Faculty of social sciences and humanities Zagreb
University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
Phone: +385-1-6002365 Fax: +385-1-6156879 E-mail: rvrana@ffzg.hr

Abstract – The paper presents an idea of expansion of the existing modes of collaboration between science with focus on e-science, libraries which services are currently going through a new phase of transformation into services based on use of federated digital information resources and finally, higher education system which seeks quality information resources for their use in the teaching process. All three sides have a historical background of mutual collaboration. Influence of information and communication technology on all three sides has sped up their development and provided them with new tools for creation, organization, storage and exchange of knowledge. All three sides still haven't used a full potential of networked environment and internet services for the benefit of even richer collaboration and designing of new ICT supported services in science, libraries and teaching at higher education institutions that would enable richer scientific, library and educational experiences. The results of this collaboration should be network intensive collaboration based on selected and verified digital content of scientific provenance; user tailored networked services in libraries and innovative forms of teaching at universities.

I. INTRODUCTION

Science and libraries have always been serving as foundations for the teaching process at all levels of education. Science has provided fundamental truths about the world and libraries have supplied critical knowledge in a highly organized and structured manner necessary for functioning of a system of education. In recent years scientific research has become very complex (as it assumes collaboration of large teams of scientists and use of newest information and communication technology and laboratory equipment), networked based (it includes communication of dislocated research team members), collaborative, information resources intensive (it assumes use of published scientific works as well as re-use of data from previous scientific research) endeavor. Integration of vast amounts scientific content (results of all kinds of scientific endeavor) into the teaching process has also become a very challenging task because scientific content is not always well formatted and prepared for direct use in the teaching process. Scientific content requires additional processing and preparation as well as dissemination by use of adequate internet services or proper institutions which would facilitate transfer of knowledge between science and system of education or general population. For centuries libraries have provided such services and helped in educating generations of students by offering access to different types of content including scientifically verified knowledge. The development of the internet helped additionally expanding dissemination of knowledge by offering a myriad of content types accessible almost instantly on this global computer network. According to Shank and Bell [1] "(…) the digital computer revolution has changed the paradigm by which society produces and consumes information, moving from an information model of scarcity and limited access, to an overwhelming abundance of both the quantity and formats of information available." Because of the great quantity and versatility of digital content available on the internet, teachers and students need someone to make the selection and transfer of knowledge between science and them more successfully and easily than they can do it by themselves. That intermediary is library. Library as an institution has changed significantly since computers were integrated into its operation. It changed from offering predominantly paper based information products and services to offering paper and digital information products and related services tailored to the specific needs of users. While transformation in this segment of library services is still going on, libraries are very interested in establishing a more direct bridge between science and their users at least this is true for academic libraries. Bell [2] claims that how well or poorly libraries have accomplish the task of being knowledge gatekeepers determines end users' success or failure in achieving their learning or research outcomes. That is why libraries still play a crucial role in being a bridge between knowledge, scientists and higher education in spite problems they encounter while providing their services.

II. E-SCIENCE

In the past, science underwent many changes of its research methodologies and ways in which results of the research were published. "Dramatic advances in digital technologies in the late 20th and early 21st centuries have created a ripple effect in the ways that scientists conduct research, collect and analyze data, and communicate their findings to colleagues and scholarly communities." [3]. This new phase in development of science is called e-science. E-science is "(…) science supported to a significant degree by digital information-processing and/or computational technologies, or wholly based on

these." [4]. E-science should „(…) enhance information exchange and intensify communication between researchers, improving their competitiveness as well as their ability to cooperate, particularly on the international level." [5].

One of the biggest benefits of e-science is the fact that "The availability of the internet and communication technologies has lowered geographical barriers and fostered virtual collaboration and team science." [3]. These technologies construct a research infrastructure called e-infrastructure which is here to "(…) identify the key characteristics as a mix of hardware, software, instrumentation, digital content, data, and archives, together with human resources, knowledge and expertise that are to be shared among communities of practice, and that are essentially collaborative in nature and form." [6]. "E-infrastructure should allow researchers, teachers, students and other stakeholders and also machines to find and access a rich range of data repositories and their contents determined by their research interests and their associated privileges." [4].

The practice of e-science is not completely defined and authors who are writing about e-science are in many cases indicating what e-science might be or is going to be by pointing to individual examples and case studies, usually in the form of pilot projects [7]. As a result of this situation, higher education system can benefit only partially from e-science results. In addition to the term e-science, literature offers another term: science 2.0. The practice of science 2.0 is based on use of information and communication technology which facilitates exchange of knowledge construction and knowledge sharing between researchers [8]. Use of Web 2.0 technologies in Science 2.0 is a mutual characteristic with e-learning and it represent a link between science and education because they use the same tools (blogs, video casts, podcasts, webinars, videoconferences etc.) for communication and they also understand how these tools are used to create, access and use knowledge. E-learning represents "teaching and learning resources or experiences that are, in some way, delivered electronically" and it implies "(…) that the learner is at a distance from the tutor or instructor, that the learner uses some form of technology (usually a computer) to access the learning materials, that the learner uses technology to interact with the tutor or instructor and other learners, and that some form of support is provided to learners." [9]. In modern learning environment the computer and the internet as a networked environment have become very important elements in life of a learner, so "(…) methods of teaching and learning should incorporate components of e-learning that are based on this environment [10]. "Scientific mash-ups composed out of Web 2.0 services should be an important way to involve school children, educators, and enthusiasts in scientific endeavors." [11]. a Use of Web 2.0 technology enables faster knowledge transfer between those who create it and those who use it and their active participation in evaluation of that knowledge. This technology helps in preparation of future workforce by integrating results of scientific endeavor in all levels of education including higher education which essential function is to impart research skills within the context of a student's education [12]. Scientists also participate actively in education by publishing results of their work in part of e-science called educational e-science. For Underwood, Smith, Luckin and Fitzpatrick [13] "Educational e-science is primarily about enabling scientists, teachers and learners to create collaborative, engaging and realistic science activities, and furthermore about defining and building the infrastructure that will make this easier and practical on a larger scale and in everyday contexts."

Although one could argue that science is interesting only to scientists and therefore to a limited number of people, the results of scientific research are often popularized on television, on radio in newspapers, in popular magazines and on the internet, all available to students. In recent years, science has intensified its presence in general population by publishing its results on the internet in open access information resources huts promoting the concept of open science. Open science designates "(…) research efforts that aim to contribute to a resource that is shared among scientists or researchers" [14]. "Open science calls for liberal dissemination of new information, it is more conducive to both the maximization of the rate of growth of society's stocks of reliable knowledge and to raising the marginal social rate of return from research expenditures." [15]. Open access information resources already have an important role in providing access to scientific knowledge to all members of the global society and they change ways in which students obtain scientific information for their education on the internet on daily basis. The growing number of open access information resources available on the internet proves this point.

III. LIBRARIES AND LIBRARIANS IN THE ERA OF THE INTERNET AND E-SCIENCE

In his article on future of academic libraries, Akeroyd defined libraries in four ways: library as a collection (of resources); library as a building or space; library as function that is organization of information; library as a service, a supporting role in the context of the organization. In the context of this paper, the last way is most relevant for understanding role of libraries in academic community [16]. He points out that "(…) libraries will need to seek relevant partnerships so as to ensure a supportive place in e-learning environments." Libraries will organize digital content and ensure rights management, [16] offer immediate access to information [17], opportunity to transform information into knowledge that can be later applied in the learning process [18]. For Neal [10] core responsibilities of academic research libraries will shift from information selection, acquisition, synthesis, navigation, distribution, interpretation, education, application and preservation to roles of libraries of aggregators, publishers, teachers, research and development agencies, entrepreneurs and information policy advocates. For Lougee [20], libraries are becoming more deeply engaged in the creation and dissemination of knowledge and are becoming essential collaborators with the other stakeholders in these activities. For Bargellini and Bordoni [21], libraries can "(…) confirm and evolve their role as reference points for technological and social

education in a learning society requiring new forms and new places of learning." On the other hand, Beagle discussed the role of libraries in Web based learning environments and he concludes that in the developing world of non sequential learning in which learning will take place in chunks throughout life, at the learner's convenience (time, place, style etc.), libraries could be marginalized as university administrations are not taking into account librarians' knowledge and experience in knowledge organization and delivery [22].

Libraries and their services intended for use in higher education have been deeply affected by use of information and communication technology in the same amount as science. Bazillion [12] analyzed influence of electronic publication and digital networking as equivalent to the printing press in earlier period of human history and concluded that this conclusion can be drawn from several recent developments in higher education:

1. The advent of Web-enhanced teaching and learning, which has rejuvenated older faculty and improved student participation in their own learning;
2. The emergence of librarians as information-technology educators;
3. Basic alterations to the fabric and internal geography of academic libraries constructed within the past decade; and
4. The adoption, by a growing number of institutions, of universal computer-access programs, whether through laptop leasing or by other means.

Libraries are known as institutions with multifaceted informational, educational, social and cultural roles that provide excellent learning opportunities for the diverse user population [23]. "Traditionally, research libraries serve as custodians of "downstream knowledge" - organizing and maintaining collections of post-research publications such as conference proceedings, journal articles, and books." [3]. Librarians help students find information, conduct research and write reports, and use information for personal and academic purposes. Libraries also assist learners "(…) to become critical thinkers by helping them to develop their research skills and the ability to critically analyze information they retrieve, even to the extent of making available" [21]. Furthermore, by use of ICT "(…) librarians have the opportunity to expand their traditional roles and work more closely with instructors and students to provide suitable tools, materials and references as well as instructional services." [21]. Bell suggests that libraries should deliver different types of value added outcomes:

- Accomplishment. Library workers help students and others achieve academic success, they help community members develop new skills and talents, and the act of reading a book is itself an accomplishment.
- Beauty. Libraries are places where community members can indulge in the appreciation of the arts.
- Creation. Libraries provide the raw materials that stimulate creativity, hut unlike other information providers it offers real people with whom creative individuals can establish relationships." [2]

However, position of libraries in the society has shifted from primary or most valued resource of information to an institution that is still used and consulted but it is not a primary resource of information anymore. In his article on position of libraries in digital environment, Bazillion [12] points out that "Technological obsolescence overtook many library buildings less than a decade ago, as campuses created their own networked environments and aspired to give their students ubiquitous access to computers."

It has become a fact that libraries are no longer the most important information resource for students, instead, they are "(…) something used at the end, or at best the middle, of their information search." [9]. Mi and Nesta [24] think similarly: "Libraries are losing their role as the "primary information provider" as users turn to search engines as their starting point to do research." Librarians and libraries "(…) face a value deficit in terms of the perception and acknowledgment of their intrinsic worth compared to other sources of information as well as gathering places where that information is used." [25]. Additional form of pressure to change themselves libraries received from miniaturization of computing devices including mobile phones which enabled ubiquitous access to knowledge in digital format. Mobile phones introduced another type of learning - m-learning which enabled "(…) anywhere and anytime learning" [26] and which depends on availability of educational content in digital format. With the help of mobile technology students can access results of scientific discovery before educators and "(…) they learn through new technology before educators have inquired to their attributes." [26]. In this way, educators started to play catch up with the students and they need help in this game. Libraries can provide "(…) any time anywhere information - structured information that is organized and accessible" [26].

Digital libraries are in slightly better position than traditional libraries because they can offer readymade digital content in form of e-books, e-journals, e-textbooks, and different additional content relevant to the learning process: curricula and courseware materials, lectures, lesson plans, computer programs, modeling and simulation, intelligent tutoring systems, access to remote scientific instruments, project-based learning, tools, the results of educational research, scientific research reported both formally in journals and informally in web sites, raw data for student activities, and multimedia image banks [9]. They also offer "(…) capability to search for desired information by subject area, to have access to scientific data sets, to interact with peers, to provide archiving, location-independent naming, recommender systems, selective dissemination of information, copyright management." [9]. As a result, students started to "(…) rate convenience over quality. They want to get information fast." [24]. In his article on library value deficit Germano [25] discussed actual values of a library and came to conclusion that "(..) the true value is the

accessibility of content along with knowledgeable guides in the form of librarians that can facilitate the best use of that content." Jeong researched user perceptions and behavioral intension towards e-library (another name for digital library) and synthesizes advantages of e-library in comparison to the traditional library: "(1) resources are stored in a digital form are, therefore, easier to track; (2) the access to e-library collections is remote, fast, and fair; and (3) searching techniques offer increased flexibility and power to users." [27]. We can conclude that digital libraries can contribute significantly to availability of scientific information for use in education by offering a selected information resources based on their quality and by enabling easier access to these resources (without administrative limits or fees).

IV. SYNERGY OF E-SCIENCE AND LIBRARIES AS A SUPPORT TO HIGHER EDUCATION INSTITUTIONS

Modern science and higher education cannot easily function without assistance of libraries and librarians. The same applies to higher education system. „Among the many participants in scholarly communication, librarians have been recognized for their expertise in organizing, enhancing and disseminating information, and have been identified as logical partners in the stewardship of digital data", and other type of scientific content [3]. In the new context of digital information resources librarians are also called "data librarians", "subject librarians", e-science librarians" [28] "information architects," "digital asset managers," or "content specialists" and "they are the professionals behind the scene who are selecting, collecting, organizing, preserving, and providing access to digital libraries [29]. Librarians' job also changed and partially merged with teachers' job as they have become educators to students, teachers and also to scientists. They "(…) have had not only to master information technology but also to assume a role as educators who teach the skills of using information technology to obtain access to information sources." [30]. Librarians are also teaching scientists how to describe content they create during scientific research by using appropriate type of metadata which are crucial for later discovery of scientific knowledge in networked environment. For Shank and Bell [1] "Librarians have a historically, unprecedented opportunity to increase their relevance and participation through assisting faculty and other academic staff with student instruction in the various types of literacy (i.e., computer, media) or information fluency." Librarians must become experts in storing, sorting and accessing electronic records, including accessing databases and using search engines; they must acquire expertise in providing online tuition and to collaborate with faculty to incorporate aspects of library use into coursework. [30]. Libraries are excellent intermediaries between science and students. Libraries "(…) improve the quality of life of its constituents by being a center of knowledge and learning." [31]. In addition to traditional activities such as helping students to write term papers and other written assignments, independent research undertaken by students requires different approach to collaboration between students and librarians. "Independent research experiences entail real hands-on experience in research conception, design, conduct, and dissemination and make inherent contributions to a discipline." [32]. This is another opportunity for libraries to promote and extend their services, particularly information literacy programs, into the undergraduate research community [32]. As a result, librarians and „Academic libraries can become a value-added component of the educational process, to which they will contribute in several ways: [12]

- Broadening and improving access to electronic information sources,
- Providing special facilities and equipment such as high-end workstations and interactive TV connections,
- Assisting faculty in integrating technology into their teaching and research activities,
- Maintaining a high level of staff expertise concerning developing information systems,
- Helping students distinguish between Web-surfing and electronic research, and
- Providing an atmosphere in which technology enhances education."

As it was previously stated, scientists and students need to interact with digital information resources. If they didn't like paper based information resources because of their limitations typical for paper as a medium, students would achieve greater familiarity with digital information resources in digital libraries and information and communication technology. By interacting with digital libraries students' negative attitudes towards the information and communication technologies could be decreased while another category of users of digital libraries - researchers - are provided with the opportunity to access information in any format and from any place [33]. In this way both categories of libraries users become more familiar with the technology (not all scientific disciplines use ICT intensively), with digital information resources and software tools which facilitate communication between all three sides.

Changes in science and libraries have introduced a new type of librarians' job which incorporates knowledge of librarianship, newest communication technologies, interpersonal skills, project management and education. Alvaro, Brooks, Ham, Poegel and Rosencrans [28] wrote about such an emerging job of e-science librarians. E-science librarians would use the Internet and other digital tools to facilitate scientific data collection, management, and sharing. In addition to traditional library related skills (developing collections, instructing users, working with copyright issues, and acting as liaison between different user groups.), he or she would also have strong technology skills and have responsibilities related to e-science in the form of liaison or research project management duties. Librarian's role in e-science is yet uncertain as future and all operational aspects of e-science are still not clear. We mustn't forget their role in higher education in which they would help facilitate access of teachers and students to results of the latest scientific achievements. If libraries fail to do so, society

will create a new type of institution to do the same job and society would be limited in use of its own intellectual output [31].

Another form of convergence of librarianship, science and teaching is a concept of embedded librarianship. "Embedded librarianship takes a librarian out of the context of the traditional library and places him or her in an "on-site" setting or situation that enables close coordination and collaboration with researchers or teaching faculty." [34] Librarians are given an opportunity "(…) to demonstrate their expertise as information specialists and to apply this expertise in ways that will have a direct and deep impact on the research, teaching, or other work being done." [34]. Librarians become partners with researchers and teachers and help them develop tighter bonds with those to whom they provide information products and services. Librarians work more closely with researchers and teaching staff participating in research projects or they are part of such projects. In this way librarians are directly involved in the process of scientific discovery and creation and curation of research data and transfer of these materials to other researchers or transfer of the final products such as journal articles and books based on this research the participated in to teachers and students at higher education institutions.

V. CONCLUSION

Implementation of information and communication technology in different areas of life has created new opportunities for collaboration and exchange of knowledge for all using it. Some professions, institutions and areas of human endeavor benefited from this fact more than other. This paper focused on science and its current development phase called e-science which benefited significantly from use of information and communication technology, electronic equipment in laboratories and networked information resources created by libraries. In addition to digital information resources, libraries created new library services tailored to user needs thus creating new types of relationships with science and higher education institutions. This development has inspired Bell, who sees libraries as knowledge gate openers. "In that role we shift from a focus on creating access to resources to creating meaningful relationships with community members – both those who use and those who don't use our libraries. One way to differentiate ourselves while building these relationships is by designing great library user experiences." [2]. Another important aspect of collaboration between researchers, librarians and teaching staff is to build trusted relationships. Research practices are changing all the time and researchers and "(…) confronting new and unfamiliar situations themselves." [34]. Since teaching process and libraries are also changing to accommodate changes in their environments, all three sides should collaborate more closely to use available technology and information resources for the benefit of scientists, librarians, library users, teachers and students. By demonstrating the ability to develop modes of collaboration these three domains will show their knowledge, skills, and expertise to others in ways that are relevant to stakeholders that decide about financing all three domains.

REFERENCES

[1] J. D. Shank and S. Bell, "Blended Librarianship: [Re]Envisioning the Role of Librarian as Educator in the Digital Information Age", Reference & User Services Quarterly , vol. 51, no. 2, pp. 105-110, 2011.

[2] S. Bell, "From gatekeepers to gate-openers", American Libraries, vol. 40, no. 8/9, pp. 50-53, 2009.

[3] D. Kafel, Activities of Regional Consortia in Planning e-Science Continuing Education Programs for Librarians in New England, 2012, Retrieved December 12, 2012, from http://escholarship.umassmed.edu/lib_articles/140

[4] Towards a European e-Infrastructure for e-Science Digital Repositories a report for the European Commission. Twickenham : The Digital Archiving Consultancy Limited, 2008.

[5] A. Osswald, "E-science and information services: a missing link in the context of digital libraries", Online Information Review, vol. 32 no. 4, pp. 516 – 523, 2008.

[6] S. Anderson and T. Blanke, "Taking the Long View: From e-Science Humanities to Humanities Digital Ecosystems" Historical Social Research, vol. 37, no. 3, pp. 147-164, 2012.

[7] New Infrastructures for Knowledge Production: Understanding E-science. Hershey: Information Science Publishing, 2006.

[8] T. Honkela, Directions for E-science and Science 2.0 in Human and Social Sciences. Retrieved January 5, 2013, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.174.7061&rep=rep1&type=pdf

[9] S.R. Sharifabadi, "How digital libraries can support e-learning" The Electronic Library, vol. 24, no. 3, pp. 389-401, 2006.

[10] K. Shulamit and E. Yossi, "Development of E-Learning environments combining learning skills and science and technology content for junior high school", Procedia Social and Behavioral Sciences, vol. 11, pp. 175–179, 2011.

[11] G. C. Fox, M. E. Pierce, A. F. Mustacoglu, and A. E. Topcu. "Web 2.0 for E-Science Environments" in Proceedings of the Third International Conference on Semantics, Knowledge and Grid, Washington, DC, USA : IEEE Computer Society, 2007, pp. 1-6.

[12] R.J. Bazillion, "Academic Libraries in the Digital Revolution", Educause quarterly, vol. 1, pp. 51-55, 2001.

[13] J. Underwood, H. Smith, R. Luckin and G. Fitzpatrick, "E-Science in the classroom – Towards viability", Computers & Education, vol. 50, no. 2, pp. 535–546, 2008.

[14] R. Schroeder, "e-Research Infrastructures and Open Science: Towards a New System of Knowledge Production?", Prometheus, vol. 25, no. 1, pp. 1-17, 2007.

[15] P.A. David, M. den Besten and R. Schroeder, Collaborative Research in e-Science and Open Access to Information, 2009. Retrieved December 12, 2012, from http://www-siepr.stanford.edu/papers/pdf/08-21.pdf

[16] J. Akeroyd, "The future of academic libraries", Aslib Proceedings, vol. 53, no. 3, pp. 79-84, 2001.

[17] D. Pietruch-Reizes, "A new look at the university libraries in context: European Research Area", Library Management, vol. 31, no. 1, pp. 36-41, 2010.

[18] E.J. Suaiden, "The social impact of public libraries", Library review, vol. 52, no. 8, pp. 379-387, 2003.

[19] J.G. Neal, The Future of the Academic Research Library: Hope/Power/Action Through Primal Innovation and Radical Collaboration, 2010, Retrieved December 12, 2012, from http://www.neal-schuman.com/academic/Neal2010.pdf

[20] W.P. Lougee, Diffuse Libraries: Emergent Roles for the Research Library in the Digital Age, Washington, DC. : Council on Library and Information Resources, 2002.

[21] M.L. Bargellini and L. Bordoni, “The role of the library in a new learning scenario”, The Electronic Library, vol. 19, no. 3, pp. 153-157, 2001.

[22] D. Beagle, “Web-based Learning Environments: Do Libraries Matter?” College & Research Libraries, vol. 61, no. 4, pp. 367-379, 2000.

[23] H. Miao and M.W. Bassham, “Embracing customer service in libraries”, Library Management, vol. 28, no. 1/2, pp. 53-61, 2007.

[24] J. Mi and F. Nesta, “Marketing library services to the Net Generation”, Library Management, vol. 27, no. 6/7, pp. 411-422, 2006.

[25] M. Germano, “The library value deficit”, The Bottom Line: Managing Library Finances, vol. 24, no. 2, 100-106, 2011.

[26] J. Hahn, "Mobile learning for the twenty-first century librarian", Reference Services Review, vol. 36, no. 3, pp. 272 – 288, 2008.

[27] H. Jeong, “An investigation of user perceptions and behavioral intentions towards the e-library”, Library Collections, Acquisitions, & Technical Services, vol. 35, np. 2-3, 45–60, 2011.

[28] E. Alvaro, H. Brooks, M. Ham, S. Poegel and S. Rosencrans, “E-Science Librarianship: Field Undefined”, Issues in Science and Technology Librarianship, 2011 Retrieved December 12, 2012, from http://www.istl.org/11-summer/article1.html

[29] C.L. Borgman, “Where is the librarian in the digital library?”, Communications of the ACM, vol. 44, no. 5, pp. 66-67, 2001.

[30] H.-C. Sun, K.-N. Chen, C. Tseng and W.-H. Tsai, “Role changing for librarians in the new information technology era”, New Library World , vol. 112, no. 7, pp. 321-333, 2011.

[31] B.P. Heidorn, “The Emerging Role of Libraries in Data Curation and E-science”, Journal of Library Administration, vol. 51, no. 7-8, pp. 662–672, 2011.

[32] A. Stamatoplos, “The Role of Academic Libraries in Mentored Undergraduate Research: A Model of Engagement in the Academic Community”, College & Research Libraries, vol. 70, no. 3, pp. 235-249, 2009.

[33] M. Yalmana and T. Kutlucaa, “Future of e-libraries in universities”, Procedia - Social and Behavioral Sciences, vol. 47, pp. 2225–2228, 2012.

[34] J. Carlson and R. Kneale, “Embedded librarianship in the research context: navigating new waters”, College & Research Libraries News”, vol. 72, no. 3, pp. 167-170, 2011.

Modes of use and preservation of digital information resources as a support to the learning process at the Faculty of humanities and social sciences in Zagreb

Radovan Vrana
Department for information sciences
Faculty of social sciences and humanities Zagreb
University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
Phone: +385-1-6002365 Fax: +385-1-6156879 E-mail: rvrana@ffzg.hr

Abstract – The aim of this paper is to present results of the research among students at the Faculty of humanities and social sciences in Zagreb (FHSS) on two important topics: students' use of digital information resources and preservation of these digital information resources. To find out more about these two topics, a research study was designed. The results confirmed that digital information resources are very popular among students for use in their education. The results also indicated the need of students to have constant access to digital information resources and to be able to store a selection of resources on their computers. The results also showed that students want digital information resources used in education to be available indefinitely and that their teachers should be responsible for care for digital information resources used in education, which is rather peculiar. Students think that digital information resources still cannot replace printed information resources. The results indicate lack of information about students' needs and attitudes about access and use of digital information resources in education for the purpose of development of university policies on organization and preservation of digital information resources for use in higher education.

I. INTRODUCTION

It is well established fact that teaching in the era of the Internet relies increasingly on use of digital information resources that offer all the benefits that can be attributed to the nature of digital environment (searching, copying, modifying, sharing, tagging etc.). Easy access to the existing knowledge in digital format on the Internet has popularized and sped up the process of knowledge discovery and knowledge acquisition. Popularity of digital information resources is growing as "Teaching and research in higher education are increasingly dependent on primary and secondary sources that are available in forms other than traditional paper-based books, journals, and manuscripts."[1] In its document about use of the internet in education, UNESCO points out that "It is necessary to prepare the students beforehand to work with information or to provide those who use the distance form of education with special tasks destined to develop intellectual skills of critical thinking, working with verbal texts, multimedia environment, to create all kinds of so-called secondary texts (abstracts, summaries, essays, etc.), to be able to work with information.".[2] Until recently, students have been prepared for future work by using paper based information resources such as books, textbooks, journals, and theses etc., which are still widely available in academic community worldwide; however, the demand for digital content is on the constant rise and focus of educators is shifting towards creation, use and preservation of information resources in digital format. Current mixed use of printed and digital information resources is still present and desirable form of students' behavior at the moment mostly due to questionable quality of digital content found on the internet. Hwang, Wu and Ke [3] suggest that "(…) it is important to situate students in a learning scenario that combines both real and digital learning resources." Preferred use of printed or digital information resources depends on study program requirements, recommendations given by teachers, libraries and their offer of textbooks and other types of educational materials and motivation of students to use printed or digital information resources. To find out more about use and preservation of digital information resources for the inclusion in the learning process, a research study was initiated. This paper presents results of the research of students at the Faculty of Humanities and Social Sciences in Zagreb (FHSS) about use digital information resources in their education and their perceptions on possibilities for preservation of digital information resources for its inclusion in learning process.

II. AVAILABILITY AND USE OF DIGITAL INFORMATION RESOURCES FOR EDUCATION

Availability of information resources is one of the most important prerequisites for functioning of the education process in general. In the higher education system worldwide, a significant amount of information resources are published by commercial publishers and they are inaccessible unless universities pay license fees for access to these resources in digital format or buy printed copies. Until recently libraries have provided access to most of information resources intended for use

in higher education. Today teachers, scholars and students seek information on the internet more frequently than ever. For Lynch [4] "(…) a new generation of scholars and students is demanding a new kind of workspace and employing new methods of seeking and finding information.". MacKenzie [5] suggested that "The modes of scholarship - research, teaching, and communication - continue to evolve toward online digital content that supports critical innovations."

As a consequence, digital information resources could be found more and more on the Web because of "(…) the ubiquity of the Web means that any educational institution can exploit and share its teaching expertise and learning resources through a global online presence" [6]. According to the same group of authors, "Educational content on the Web frequently consists of digital or digitised versions of traditional learning materials such as lecture notes, textbooks, tutorial sheets, and so on. This content must be managed and content must be available to students and other colleagues teachers, but not to commercial enterprises without explicit author's consent [6].

Digital information resources (generally speaking) have become more acceptable for integration in education as their availability and quality increased over time as well as number of different electronic devices on which digital content can be used in education. However, in countries where educational institutions cannot provide access to large quantities of different digital resources (books, journals, magazines, newspapers, reports, theses etc.), paper based information resources are still the most important sources of information. They are perceived as more trustworthy and more easily accessible in academic libraries. Furthermore, in Croatia commercial publishers are unenthusiastic about publishing digital books as they are afraid of copyright infringement.

III. CARE AND PRESERVATION OF INFORMATION RESOURCES IN DIGITAL FORM

Care for digital information resources used in education is left mostly to teachers, students and scientists while libraries provide care for printed materials and provide access to subscribed digital information resources. Additional attention should be given to digital information resources used in science and education because of their short lifespan which is nowadays measured by a few years [5]. To solve this problem, universities around the world started to develop open access digital repositories in order to preserve and facilitate use of digital materials among the population of students, their teachers and scientists. According to [7], "Most digital repositories in academic institutions are focused on "traditional" materials such as theses, dissertations, images, video, or data sets. While these materials are useful to the researcher and in the classroom, faculty teaching in the online environment often express an interest in developing a common set of learning materials that can be reused in across courses."

Digital repositories storing education materials are often called educational repositories or instructional repositories. An instructional repository is "an organized collection of online teaching materials" [8]. "The best instructional repository is the one that contains what an instructor is looking for; next best repository that leads an instructor to consider something of value that she wasn't looking for."[8]. Instructional repositories also contain software, multimedia packages, textbooks, syllabi, quizzes, simulations, lesson plans, lecture notes, audio/video lecture archives etc. [8]. The purpose of instructional repositories is therefore the possibility to store, organize and distribute educational materials to students.

While many existing digital repositories offer open access to the content they store, instructional repositories are perhaps more restrictive to sharing content thus protecting their content from access of the people who are not members of their or any other academic institution (university). Instructional repositories are often part of academic libraries which are in charge for preservation of education materials at universities. This job of taking a long term care of digital material is called digital preservation or digital stewardship. It "(…) encompasses the creation, maintenance, preservation, dissemination, and exhibition of trusted bodies of digital information for current and future use. Our current knowledge of digital preservation still concentrates largely on the creation of durable digital objects and on their maintenance over time." [9]. For this job to be successful, libraries need an infrastructure that will enable preservation and access to digital material in an academic institution and digital material available in other academic institutions or on the Internet. Libraries work in a rapidly changing environment. Changes that influence care for digital content happened in "(…) industries, such as publishing, in which many paper-based publications are giving way to digital ones; a rise in online education; the use of handheld devices for an expanding variety of information needs; and a rapid evolution in social networking are resulting in changes in cultural heritage institutions. Consequently, preservation is evolving too."[9].

IV. RESEARCH

To find out more about problems of selected aspects of use of digital educational materials and their preservation, a research study was initiated. The purpose of this study is to understand how students at the Faculty of Humanities and Social Sciences in Zagreb (FHSS) use digital information resources in their education and to understand their perceptions on possibilities for preservation of digital information resources for its inclusion in educational process. The aim of this research study is to collect data about use and preservation of digital information resources in higher education. The hypotheses of this research are: (1) students at the FHSS still do not use digital information resources in their education in sufficient amount; (2) issues related to systematic preservation of digital information resources for education are still unresolved. An online survey consisting of 11 closed type questions was chosen as a method of research. was developed to collect data in this research. The survey was divided into two parts. The first part of the survey contained questions about use of digital information resorces in education. The second part contained questions about preservation of digital information resources needed

in education. An e-mail message with an invitation was sent to all students using mailing list at the FHSS on December 6, 2012. The survey was closed on January, 15 2013. The total of 182 students participated in the survey. The number of students was limited by the fact that not all students are subscribed to the students' mailing list at the FHSS because they were given a choice to decide whether or not they wanted to be on that mailing list.

V. FINDINGS

This part of the paper will present findings from the research. Hundred and fifty female and 32 male students participated in this research (some students didn't indicate their sex). Most respondents in the research were students from the second year of the graduate study, 95 of them (52,2%), 26 students from the first year of the graduate study (14,3%), 25 students from the third year of the undergraduate study (13,7%), 14 students from the postgraduate study (7,7%), 8 students from the old unreformed study (4,4%), 6 students from second year of the undergraduate study (3,3%), 4 students from the fourth year of the undergraduate study (2,2%), and 4 students from the first year of the undergraduate study (2.2%). More participants were expected from the group of students at the undergraduate level of the study.

The next part of the survey aimed at finding out facts about use of printed and digital information resources in students' education.

TABLE I.
USE OF PRINTED AND DIGITAL MATERIALS FOR STUDYING (MULTIPLE ANSWERS)

	Printed content	Digital content	N
Which of the two types of content do you use more frequently for studying?	50,5% 92	49,5% 90	182
Which of the two types of content is more accessible to you for studying?	35,9% 65	64,1% 116	181
Which of the two types of content is more usable to you for studying?	40,2% 72	59,8% 107	179

TABLE II.
WHAT ELECTRONIC DEVICES DO YOU USE FOR ACCESS TO EDUCATIONAL CONTENT ON THE INTERNET (N=180) (MULTIPLE ANSWERS)

	N (%)
Personal computer	68 (37,8%)
Portable computer	150 (83,3%)
Family computer	71 (39,4%)
Computer owned by a friend	23 (12,8%)
Computer in some public institution	156 (86,7%)
Computer in some public space	3 (1,7%)
Tablet	9 (5,0%)
Older generation mobile phone	12 (6,7%)
Smartphone	73 (40,6%)
TV with Ethernet	3 (1,7%)
DVB-T or SAT receiver with Ethernet	2 (1,1%)
Some other device	5 (2,8%)

TABLE III.
WHAT CHARACTERISTICS OF DIGITAL CONTENT DO YOU CONSIDER IMPORTANT (N=181) (MULTIPLE ANSWERS)

	N (%)
Possibility of information retrieval	177 (97,8%)
Possibility of content printing	140 (77,4%)
Unlimited access to content	151 (83,4%)
Copy / paste capability	134 (74,0%)
Digital content highlighting	67 (37,0%)
Content sharing	138 (76,2%)
Content storing (locally)	169 (93,4%)
Some other characteristics	12 (6,6%)
I don't use digital content	0 (0,0%)

TABLE IV.
WHERE DO YOU SEARCH FOR EDUCATION MATERIALS IN DIGITAL FORMAT FOR STUDYING (N=181) (MULTIPLE ANSWERS)

	N (%)
Google	166 (91,7%)
Google Scholar	69 (38,1%)
Full text databases	136 (75,1%)
FHSS library	143 (79,0%)
Library outside FHSS	39 (21,6%)
Bookstore	8 (4,4%)
Friend	17 (9,4%)
Students of senior years	29 (16,0%)
Personal Web pages of professors	14 (7,7%)
Web pages of courses	52 (28,7%)
LMS Omega	117 (64,6%)
FHSS digital repository	29 (16,0%)
Somewhere else	5 (2,8%)

Somewhere else: "illegal web sites with books", "Mendeley, CiteULike, Zotero", "private library", "at home", "references that are of interest to me".

Next section of the research is focused on preservation of digital content used for studying.

TABLE V.
WHERE DO YOU STORE OUTPUT OF YOUR WORK? (N=182) (MULTIPLE ANSWERS)

	N (%)
Personal computer	171 (94,0%)
Computer shared with someone	40 (22,0%)
On the Internet (Dropbox, Skydrive Google Drive etc.)	74 (40,7%)
LMS Omega	27 (14,8%)
Mobile phone	19 (10,4%)
Somewhere else	29 (15,9%)

Somewhere else: "USB memory key" (12 times), "USB and external HDD" (5 times), "e-mail" (5 times), "external HDD" (4 times), "Mendeley internet service", "E-reader", "at home – paper documents"-

TABLE VI.
FOR HOW LONG SHOULD COURSE RELATED DIGITAL CONTENT BE AVAILABLE? (N=181)

	N (%)
Only during the course	12 (6,6%)
3 months after the course finishes	16 (8,8%)
6 months after the course finishes	7 (3,9%)
9 months after the course finishes	3 (1,7%)
12 months after the course finishes	11 (6,1%)
24 months after the course finishes	11 (6,1%)
Indefinitely	121 (66,9%)

TABLE VII.
IN YOUR OPINION, WHO SHOULD PRESERVE AND CARE ABOUT EDUCATIONAL MATERIALS IN DIGITAL FORMAT? (N=182)

	N (%)
Students themselves	8 (4,4%)
Professors themselves	106 (58,2%)
FHSS library	41 (22,5%)

National and university library	6 (3,3%)
University of Zagreb	14 (7,7%)
Someone else	7 (3,9%)

TABLE VIII.
IN YOUR ESTIMATION, COULD DIGITAL CONTENT REPLACE COMPLETELY PRINTED CONTENT IN YOUR STUDYING (N=180)

	N (%)
Yes	23 (12,8%)
No	136 (75,6%)
Can't estimate	21 (11,7%)

TABLE IX.
IN YOUR OPINION, IS THE CONTENT AVAILABLE FREE ON THE INTERNET OF LOWER QUALITY THAN CONTENT AVAILABLE FOR FEE (ON OR OFF THE INTERNET) (N=182)

	N (%)
Yes	38 (20,9%)
No	144 (79,1%)

VI. DISCUSSION

Students at the FHSS who participated in this research are more and more relying on use of information and communication technology (ICT) during the period of their study as ICT becomes more available and accessible to the growing number of students. Wider availability of ICT and its use in daily activities at the FHSS reflected on the results of this research. According to the results about use of printed and digital content in education ("Table I"), a little more than a half of students (50.5%) use digital content more frequently than printed counterpart; the same type of content is more accessible in daily activities of two thirds of students in this research (64.1%) and the same type of content is more usable for over one half of the same group of students (59.8%). These results indicate that digital content has become a very significant factor for achieving success during the study at the FHSS. It should be noted that the FHSS is not a technically oriented Faculty and that these research results indicate increased level of penetration of use of digital content among the students at the FHSS. Digital content has become more available in academic communities around the world and the same applies to the students at the FHSS in Zagreb, Croatia. Actual use of digital content can be observed and evaluated by doing an bibliometric analysis of students' written works in which they use digital information resources more and more intensively.

Digital content cannot be used without some kind of an electronic device. Findings in the fourth question ("Table II") suggest growing importance and actual use of portable devices as well as use of desktop computers in public institutions such as libraries. These two categories of electronic devices are followed by use of smartphones and desktop computers shared with other persons while use of desktop computers used exclusively by a student is declining. Other electronic devices such as mobile phones of an older generation, or TV sets and TV receivers with network connection capabilities are used significantly less. Generally speaking, increased use of portable electronic device indicates the existence of need for the education related digital content tailored for use on mobile phones, portable computers and tablets. The demand for digital content adapted for use on portable devices will have influence on commercial and academic publishers publishing educational materials for use in higher education. Scientific information resources have been accessible in digital format for many years; however, in educational practice there is an evidence lack of educational materials in the same format, at least in Croatia and in the Croatian language.

The results in the next question ("Table III") indicate the significance of a selected number of characteristics of digital content used in the learning process. Students were given an opportunity to select most important characteristics of the digital content they use in the process of studying. Their choice shows that possibilities of information retrieval, storing on an electronic device and unlimited access are three most important characteristics that drive their use of this type of the content. These top three characteristics are followed by content printing, content sharing and copy / paste possibilities while content highlighting and some other characteristics are less represented. Students in this research have chosen those characteristics which provide them with direct access to the wanted information which they use in the learning process at the FHSS. As digital scientific and educational materials are replacing their printed counterparts, these characteristics are becoming more important. They enable quicker access, selection, acquisition, storage and use of information needed for academic work at universities.

Searching for information has usually been made possible mostly by libraries which jobs was to provide access to printed information resources to students, professors and scientists. The Internet with its search engines and full text databases of scientific and professional content changed this situation. Information resources available on the internet have become more important to students than printed resources. Results in this question ("Table IV") show that the Internet search engine Google is the first choice for searching information. The second place is reserved for the FHSS library in which many students search for educational materials. The FHSS library also provides access to full text databases (http://knjiznica.ffzg.unizg.hr/baze) which are third most popular choice of a medium or a place where students search for information. Other places or resources where students search for information are significantly less represented. Among them LMS Omega at the FHSS (http://omega.ffzg.hr/) holds fourth position since students who use this LMS have access to educational materials published there by their professors which is available mostly during the semester. Other places or resources such as widely advertised Google Scholar or libraries outside the FHSS are not of great interest to students in this research. The Internet and search engines such as Google have become most ubiquitous information searching resource available and students accepted them instead of libraries, except for libraries that are closest to the place of study.

The next part of the survey had focus on preservation of digital content used by students during their studying. Almost all students store digital content on their personal computers ("Table V") and on some Internet service that

provides storage space. Both locations can be related to convenient use of digital content that must be available to the student at all times and be portable at the same time. These are possible reasons for the popularity of these two storage options. The remaining options are computers shared with someone, LMS Omega, some other option chosen by the students in research and not listed in this list, and mobile phone. It is expected that students will choose places or devices to store content they use on the basis on accessibility of the content and reliability of that place or device. Furthermore, as mobile devices will develop and offer more storage space, they will be seen as more attractive option too.

The long term availability of education materials is one of the most important questions for students. Students' answers to the question of length of availability of study related materials were expected ("Table VI"). A little more than two thirds of the respondents want materials to be available indefinitely. The next two answers are 3 months after the completion of the course and only during the participation in the course. Course materials are materials selected from a variety of materials available to professors and students might like to use pre-selected reading materials which would save them time for doing other activities. The downside of relying too much on pre-selected materials is that students don't search for additional materials on their own and expect that the teaching staff do the searching instead of them. The good side of the long term availability of the reading material is that students can always come back and consult the material over and over again, as needed. Course materials could include presentations, notes, videos, podcasts etc.

According to the results on the question about responsibility for preservation of the teaching materials ("Table VII"), students think that professors themselves should be responsible for the preservation and after them, the FHSS library and University of Zagreb. It is interesting that only 4.8% of the respondents think that they (students) should be responsible for preservation of the teaching materials. Such a low percentage is rather surprising given the results in the question of storing the teaching materials where students indicated that they store the materials on the personal computers and in the internet services that offer the storage space. That means that they are aware of the problem of the materials preservation, they store them for their own use but think that they are not responsible for the long term preservation. Reasons for that behavior might be that students do not posses enough knowledge about the long term preservation of digital content, that they do not have the necessary computer resources or that they think that someone else should do the job of preserving the teaching materials and keep it online forever. Some additional research would clarify students' position on this issue.

Almost 75% of the respondents think that digital content still cannot replace printed content necessary for studying completely ("Table VIII"). Although they prefer use of digital content, they still give more importance to printed content. This could mean that digital content is related to practical everyday use, while printed content is more durable and reliable. Of course, the use of digital content varies between study programs at the FHSS and this fact should be also taken into consideration.

The final question was related to the quality of digital content freely available on the Internet in comparison to the content available for fee (online or offline) ("Table IX"). Students' answers suggest that they believe in the quality of digital content available for free on the Internet (80% of the respondents). Answers to this question might suggest some discrepancies in attitudes of the respondents towards reliability and quality of printed and digital content. While students believe in accessibility of digital content on the Internet, they still think printed content wouldn't be replaced by digital content (which suggests belief in the quality of printed content) and yet they overwhelmingly believe in the quality of free content on the Internet. It is possible that part of the respondents is not fully aware of the content quality assessment criteria and that they equalize availability and quality. More research in this area should also clarify the issue.

VII. CONCLUSION

Digital content use and long term preservation are gaining momentum at the Croatian universities. Research in this paper demonstrated how students at the FHSS use digital content in their education and collected data about their opinions on its preservation. The results show that students seek unlimited accessibility and availability of digital content that can be used on a number of electronic devices. As much as students would love to use digital content without any obstacles and as much as they store digital content on their personal computers for future use, they would still like someone else to take care of the long term preservation of the digital content. In such scenario, libraries would be the best institutions for this task as they already possess knowledge and skills necessary for long term content preservation. In some cases libraries already have the necessary infrastructure which can be employed to collect and organize digital content of different type. At the present moment, students think that digital content cannot replace printed content completely.

REFERENCES

[1] D.J. Waters, "Managing digital assets in higher education: an overview of strategic issues", ARL Bimonthyly Report, No. 244, 2006.

[2] Internet in education: support material for educators. Moscow : Unesco institute for information technologies in education, 2003. Retrieved December 18, 2012, from http://unesdoc.unesco.org/images/0013/001330/133023e.pdf

[3] G.J Hwang, P.H. Wu and H.R. Ke. "An interactive concept map approach to supporting mobile learning activities for natural science courses", Computers & Education, vol. 57, no. 4, pp. 2272–2280, 2011.

[4] B.P. Lynch et al. "Attitudes of presidents and provosts on the university library", College and research libraries, vol. 68, no. 3, pp. 213-228, 2007.

[5] S. MacKenzie. "Exploring Variety in Digital Collections and the Implications for Digital Preservation", Library Trends, vol. 54, no. 1, pp. 6-15, 2005.

[6] C. Allison et. al. "The Web in education", Computer Networks, vol. 56, no. 18, pp. 3811–3824, 2012.

[7] H.F. Cervone. “Digital learning object repositories”, OCLC Systems & Services: International digital library perspectives, vol. 28, no. 1, pp. 14-16, 2012.

[8] J. Hart and B. Albrecht. "Instructional repositories and referatories", Research Bulletin, vol. 2004, no. 5., pp. 1-12, 2004.

[9] J.A. Bastian, M.V. Cloonan and R. Harvey, Ross. “From Teacher to Learner to User: Developing a Digital Stewardship Pedagogy”, Library trends, vol. 59, no. 4, pp. 607–622, 2011.

Towards a Framework for Usability Evaluation of Educational Artifacts created with Web 2.0 Applications: A Pilot Study

T. Orehovački and N. Žajdela Hrustek
University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
{tihomir.orehovacki, nikolina.zajdela}@foi.hr

Abstract - Web 2.0 refers to the second generation of web applications which encourage different kinds of interaction among users and enable them to create various artifacts. Although Web 2.0 applications are widely used in educational settings, researches on usability of artifacts created by their means are fairly modest. This paper explores attributes which contribute to the usability of educational artifacts. With an aim to develop a comprehensive framework, we conducted a pilot study during which data was collected with the use of an online questionnaire. Participants in the study were students of Information and Communication Sciences. Empirical findings are presented and discussed.

I. INTRODUCTION

Almost a decade ago, Tim O'Reilly and Dale Dougherty [1] coined a term "Web 2.0" which is nowadays commonly used as a synonym for the second generation of web applications. Web 2.0 is considered to be a significant shift from static web sites where users could only browse, view, read, and download published content to dynamic web services which support different types of interaction among users and enable them to share created artifacts. Influenced by Web 2.0 paradigm, the traditional e-learning process where students were passive recipients of educational content solely produced by teacher was transformed into e-learning 2.0 [2] where students are able to interact and actively contribute to the development of educational artifacts. Standalone learning platforms aimed for publishing learning content are replaced with decentralized and interoperable learning communities where students can remix and use educational artifacts in accordance with their own preferences. One of the core elements underlying the concept of e-learning 2.0 is connectivism [3] according to which learning is no longer an individualistic but collaborative activity.

Usability assessment plays a fundamental role in the user-centred design of web applications. According to the recent standard on software quality evaluation [4], usability is defined as "degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". Current research on evaluation in the e-learning context deals with usability assessment of learning environments [5] or measuring technical and pedagogical criteria of static learning material [6]. Latest research effort on assessment of Web 2.0 applications is focused on measuring quality of mashups [7], and development of a methodology for evaluating the quality in use of Web 2.0 applications [8][9] whose adequacy has been empirically validated in the context of services for collaborative writing [10] and mind mapping [11]. However, the existing research on evaluation of artifacts created with Web 2.0 applications is rather rare. The objective of this paper is to present results of the pilot study where usability of educational artifacts created by means of various Web 2.0 applications was investigated by means of an online questionnaire.

The remainder of the paper is structured as follows. Section 2 describes Web 2.0 applications that were used in our research for completing educational activities. Details on employed research methodology are contained in the fourth section. Discussion on empirical results of conducted pilot study is provided in the fourth section. Concluding remarks and future work directions are offered in the last section.

II. WEB 2.0 APPLICATIONS

There are a number of web applications that demonstrate the foundations of the Web 2.0 phenomena and are widely applied in educational context. More recently, taxonomy of Web 2.0 applications whose features can be used in educational settings has been proposed [12]. This section offers a brief overview of the most prominent representatives of the set forth taxonomy that are tailored to the context of our study and supplemented with practical examples of their implementation in educational environment.

Wiki refers to a web service such as *MediaWiki* that enables users to create a knowledge repository in a form of interlinked web pages. It can be effectively used for gaining pedagogical benefits by both students and teachers. Students can employ a wiki for writing, editing, and sharing their own thoughts during classes, creating content that supplements and extends delivered course materials, and asynchronously collaborate with other peers on joint projects. On the other hand, teachers can apply a wiki for publishing lectures, providing feedback to

students, etc. In that respect, wiki can serve as a central place for the implementation of educational activities as well as for the integration of educational artifacts created by means of other Web 2.0 applications [13].

Cloud Suite Services such as *Google Drive* represent a free alternative to commercial office applications. They can be used instead of the paper and pencil for taking notes during lectures as well as for generating different types of educational artifacts thereby creating collection of course study materials. Being collaborative in nature, cloud suite services enable students to work synchronously on the same document, participate in brainwriting sessions, share and organize created documents, etc.

Social Bookmarking Sites such as *StumbleUpon* are aimed for discovering, saving, sharing, managing, and referencing various artifacts. They allow users to associate flat, unstructured, and individually selected keywords known as tags with a particular artifact. The emergent collection of all the users' tags and artifacts is commonly referred to as folksonomy. Given that annotation of web artifacts has proven to have effect on problem solving ability [14], social bookmarking sites represent a valuable source of artifact needed for completing educational activities (e.g. writing a seminar paper).

Web applications for knowledge organization such as mind mapping and diagramming services are meant for visual representation of information. *Mindomo* as a representative of mind mapping services helps students to decompose and understand complex problems thus enabling them to arrange, memorize, and learn information in highly sophisticated way. *Gliffy* as a member of the diagramming family allows students to generate a graphical illustration of an algorithm or process thereby assisting them in the analysis of given assignment and design of its solution. Both mind maps and diagrams can be created collaboratively, shared publicly or privately, and embedded into other Web 2.0 applications.

Screencasting services such as *Screencast-O-Matic* enable users to record screen output together with audio narration. Teachers can use them for recording lectures or tutorials and accordingly make screencasts available to all students for future reference. Attendees who were absent from a particular lecture can use screencasts to catch up on the content they missed. In addition, screencasts are adaptable to students' learning style and pace. Namely, students can view screencasts whenever they want and as often as they need. Finally, students can record a screencast to demonstrate a procedure of solving a given assignment in specific environment.

Social Networks are structure of personalized web pages that represent individuals and relationships among them within a particular domain. They allow individuals to create public or semi-public profile in order to establish a connection, interact, and share artifacts with others who have similar interests. Depending on their primary purpose, Orehovački et al. [12] distinguish three types of social networks: general, professional, and specialized. Social networks meant for general use such as *Facebook* can be utilized by students for acquiring new knowledge through discussion with their peers, getting feedback from teacher, creating a repository of diverse artifacts, etc. Professional social networks like *Linkedln* enable students to create portfolios of their achievements over a certain period of time. Moreover, they allow students to connect with experts in a certain field and through interaction with them widen theoretical knowledge gained during the course. Specialized social networks such as *Snipplr* enable students to obtain feedback from teacher, professionals, and other peers on correctness of published snippets of the source code. In that manner, social networks encourage collaborative learning while their features motivate students to become more skillful in solving programming assignments.

Web 2.0 applications that merge data from multiple sources into one integrated interface are referred to as mashups. Popular representatives of mashup applications are personalized portals like *Netvibes* which provide their users with a feature of subscribing to the Rich Site Summary (RSS) or Atom Syndication Format (Atom) feeds of interest. Feeds are XML formats that provide users with concise and frequently updated content. Individuals are able to use mashups for following updates from multitude web resources. By employing services such as *Dapper*, both students and teachers can create feeds thus enabling all interested to track changes on artifacts generated by means of other Web 2.0 applications. Finally, students have the ability to subscribe to different educational feeds and thereby design Personal Learning Environment (PLE).

III. RESEARCH DESIGN

With an aim to evaluate educational artifacts created by means of Web 2.0 applications, a questionnaire with 46 items related to various facets of usability was designed. These items were supplemented with demographic questions regarding participants' age and gender, as well as with questions on frequency of using most popular representatives of different types of Web 2.0 applications. Responses were modulated on a five point Likert scale (1-strongly disagree, 5-strongly agree). In addition, overall usability of educational artifacts was assessed by means of five point semantic differential item (1-low, 5-high). The pilot study lasted for four months (one semester) and was comprised of several parts. As an integral part of each of four different assignments, students had to create educational artifacts by means of Web 2.0 applications described in previous section. For each assignment they were asked to use different representative of particular type of Web 2.0 applications. All created artifacts were integrated with wiki by means of links or embedded images. Thereafter, students were asked to use integrated artifacts when learning for midterms. At the end of the semester, an online questionnaire was employed for collecting data on perceived usability of educational artifacts. A total of 92 subjects (63% male, 37% female) aged from 19 to 22 (M = 20.26, SD = 0.800) took part in the pilot study. The subjects in the pilot study were students of Information and Communication Sciences. At the time when the pilot study was conducted, subjects were in the second year of study. The pilot study was conducted within the Data Structures course.

IV. PILOT STUDY RESULTS

The data on use frequencies of most popular representatives of Web 2.0 applications with educational potential (blogs, social bookmarking sites, social networks, image sharing services, cloud suite services, mashups, instant messengers, virtual worlds, microblogs, wikis, and video podcasting services) is presented in Figure 1.

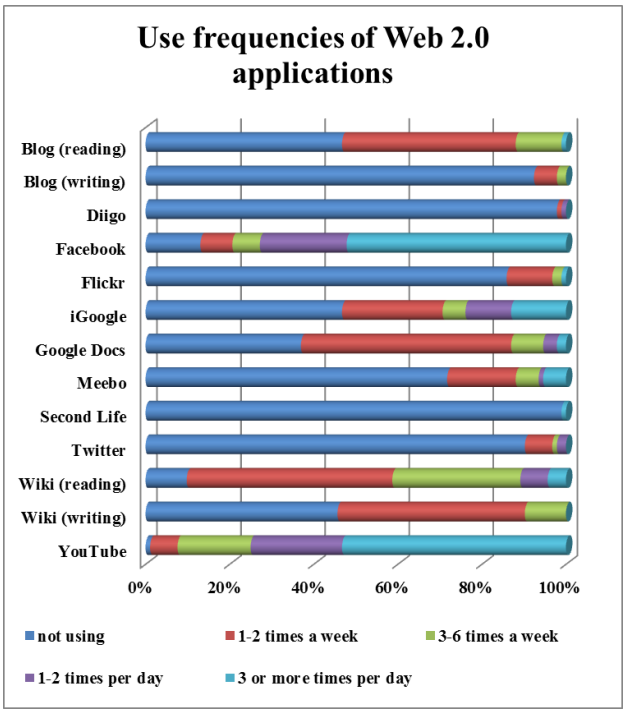


Figure 1. Use frequencies of Web 2.0 applications

Among eleven different types of Web 2.0 applications, participants in the pilot study most commonly use video podcasting services and social networking sites. Namely, the data in Figure 1 indicate that YouTube and Facebook are used at least once a day by 75% and 72.82% of pilot study participants, respectively. On the other hand, both virtual worlds and social bookmarking sites are used quite rare. More specifically, 98.91% and 97.83% of pilot study participants reported that they do not use Second Life and Diigo, respectively. When contrasting active and passive use of wikis, majority of pilot study participants more often read wiki entries (90.22% at least once a week) than contribute to its body of knowledge (54.35% at least once a week). Pattern of using blogs is very similar. As many as 53.26% of pilot study subjects read blogs at least once a week while only 7.61 of them writes their own blog posts. Use frequencies of remaining five Web 2.0 applications are as follows: Google Docs, iGoogle, Meebo, Flickr, and Twitter are used at least once a week by 63.04%, 53.26%, 28.26%, 14.13%, and 9.78% of study subjects, respectively. The aforementioned findings indicate that students of Information and Communication Sciences do not use frequently enough various types of Web 2.0 applications which are in accordance with findings discussed in [15] and [16].

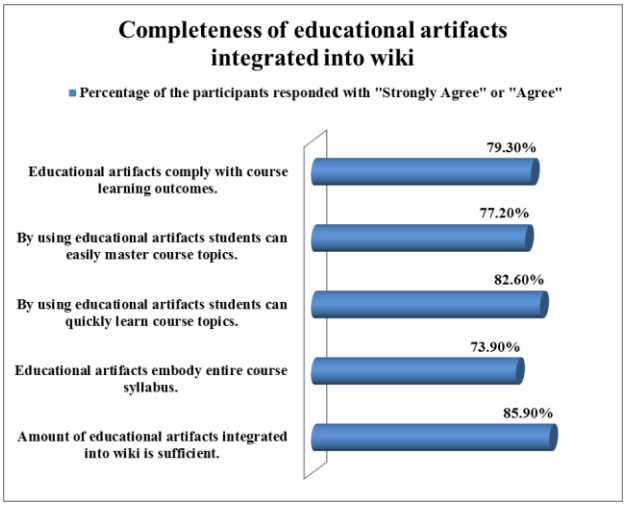


Figure 2. Completeness of educational artifacts integrated into wiki

As can be observed from the responses to questionnaire items which are presented in Figure 2, pilot study participants believe that educational artifacts integrated into wiki cover all course topics. In particular, majority of participants assume that educational artifacts comply with course learning outcome (M = 3.61, SD = 0.877) and embody entire course syllabus (M = 3.53, SD = 1.133). Moreover, most of pilot study subjects agree that use of educational artifacts leads to increased learning efficiency (M = 3.72, SD = 0.869) and facilitates knowledge acquisition (M = 3.52, SD = 1.032). Finally, 85.9% of students support the statement that the amount of educational artifacts integrated into wiki is sufficient (M = 3.79, SD = 0.819).

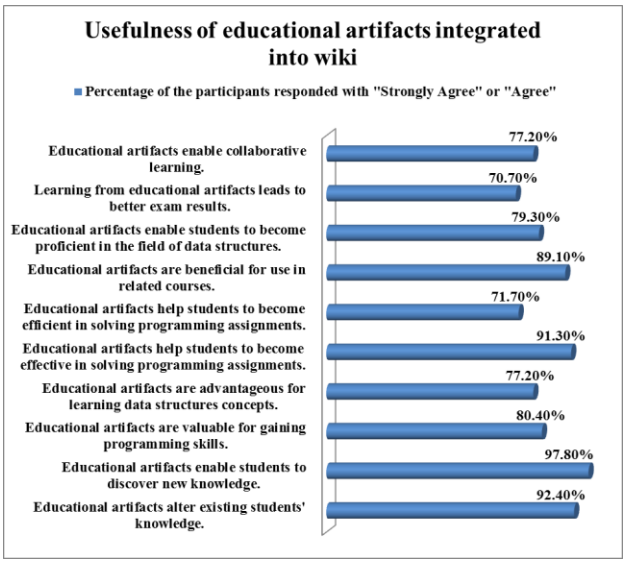


Figure 3. Usefulness of educational artifacts integrated into wiki

According to the data presented in Figure 3, pilot study subjects perceive educational artifacts integrated into wiki as useful. For instance, predominant number of respondents strongly agree or agree that educational artifacts alters existing students' knowledge (M = 4.00,

SD = 0.679) and enables them to discover new knowledge (M = 4.22, SD = 0.551). Furthermore, most of the pilot study participants strongly agree or agree that educational artifacts help students to become effective (M = 4.01, SD = 0.734) and efficient (M = 3.49, SD = 1.094) in solving programming assignment as well as proficient in the field of data structures (M = 3.68, SD = 0.983). Also, majority of students responded that they strongly agree or agree that educational artifacts integrated into wiki are valuable for learning data structures concepts (M = 3.55, SD = 1.152), and gaining programming skills (M = 3.72, SD = 0.941) which leads to better exam results (M = 3.41, SD = 1.111). Majority (89.1%) of pilot study subjects used the same responses to express their agreement with the statement that educational artifacts are advantageous for use in data structures related courses (M = 4.04, SD = 0.876). Finally, as many as 77.20% of students strongly agree or agree that educational artifacts enable collaborative learning (M = 3.60, SD = 1.070).

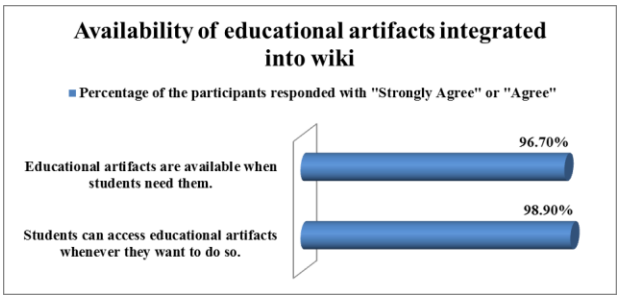


Figure 4. Availability of educational artifacts integrated into wiki

The data presented in Figure 4 illustrate that students perceive educational artifacts integrated into wiki as available. As many as 96.70% responded with "Strongly Agree" or "Agree" to the questionnaire item "Educational artifacts are available when students need them." (M = 4.55, SD = 0.618). Also, 98.90% of them gave the same responses to the statement "Students can access educational artifacts whenever they want to do so." (M = 4.51, SD = 0.671).

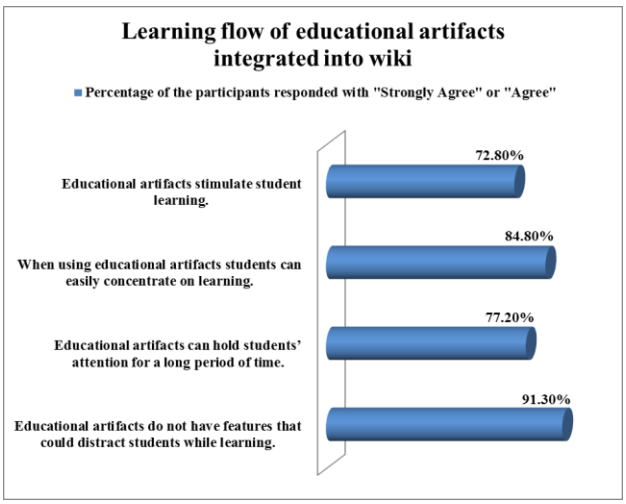


Figure 5. Learning flow of educational artifacts integrated into wiki

According to the results presented in Figure 5, educational artifacts integrated into wiki positively affect students' learning flow. Namely, most of pilot study participants strongly agree or agree that educational artifacts do not have features that could distract students while learning (M = 4.04, SD = 0.797) and accordingly can hold their attention for a long period of time (M = 3.64, SD = 0.979) and stimulate their learning (M = 3.48, SD = 0.955). Also, predominant number of them (84.80%) responded affirmatively to the item "When using educational artifacts students can easily concentrate on learning." (M = 3.78, SD = 0.849).

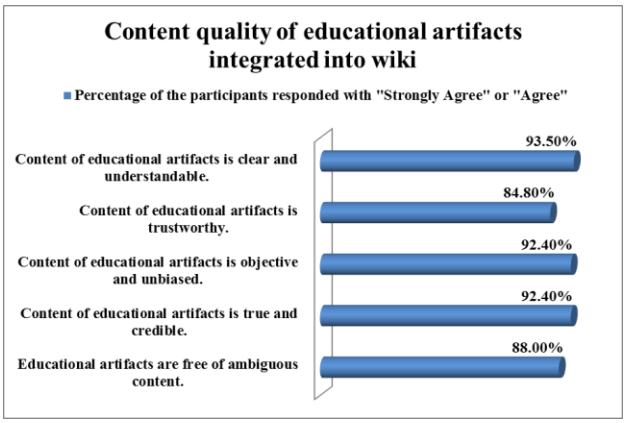


Figure 6. Content quality of educational artifacts integrated into wiki

The data in Figure 6 indicate that content quality of educational artifacts integrated into wiki is acceptable. More specifically, most of the pilot study subjects strongly agree or agree that content of educational artifacts is clear and understandable (M = 3.97, SD = 0.601), trustworthy (M = 3.79, SD = 0.896), objective and unbiased (M = 4.20, SD = 0.905), as well as true and credible (M = 4.02, SD = 0.695). In addition, majority of them (88%) support the statement that educational artifacts are free of ambiguous content (M = 3.87, SD = 0.801).

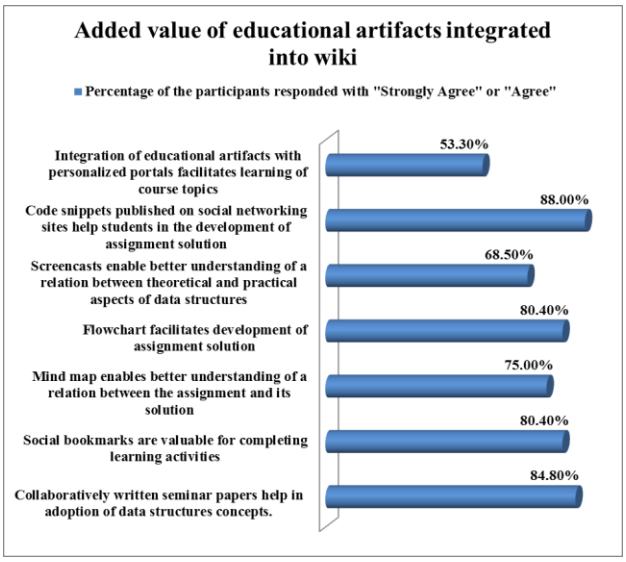


Figure 7. Added value of educational artifacts integrated into wiki

According to the data presented in Figure 7, pilot study subjects believe that educational artifacts integrated into wiki add value to the learning process. More specifically, the predominant number of respondents stated that they strongly agree or agree that a) collaboratively written seminar papers help in adoption of data structures concepts (M = 3.85, SD = 1.005), b) social bookmarks are advantageous for completing learning activities (M = 3.59, SD = 1.060), c) mind map enables better understanding of a relation between the assignment and its solution (M = 3.51, SD = 1.254), d) flowchart facilitates development of assignment solution (M = 3.77, SD = 1.196), e) screencasts enable better understanding of a relation between theoretical and practical aspects of data structures (M = 3.47, SD = 1.321), f) code snippets published on social networking sites help students in the development of assignment solution (M = 4.08, SD = 1.030), and g) integration of educational artifacts with personalized portals facilitates learning of course topics (M = 3.58, SD = 1.170).

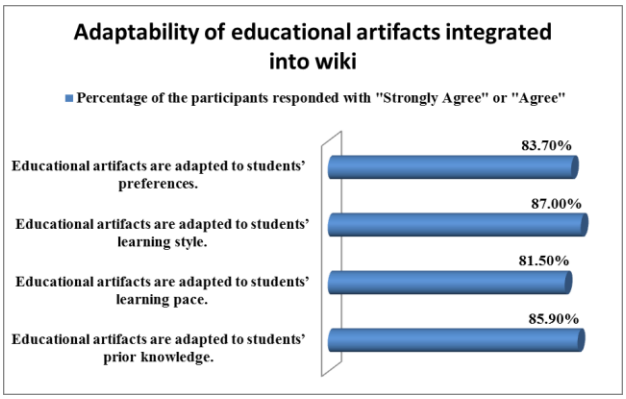


Figure 8. Adaptability of educational artifacts integrated into wiki

The data illustrated in Figure 8 indicate that educational artifacts integrated into wiki are adapted to students' needs. Namely, most of pilot study participants strongly agree or agree that educational artifacts are adapted to students' prior knowledge (M = 3.84, SD = 0.893), learning pace (M = 3.68, SD = 0.983), learning style (M = 3.82, SD = 0.838), and preferences (M = 3.70, SD = 0.848).

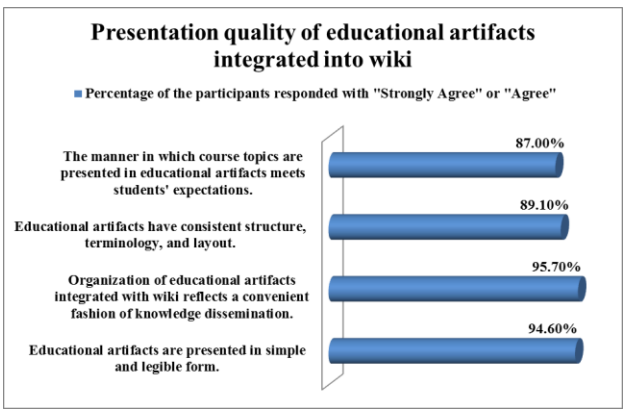


Figure 9. Presentation quality of educational artifacts integrated with wiki

According to the data shown in Figure 9, presentation quality of educational artifacts integrated with wiki is satisfactory. For instance, majority of students strongly agree or agree that educational artifacts are presented in simple and legible form (M = 4.08, SD = 0.633), and have consistent structure, terminology, and layout (M = 3.84, SD = 0.774). In addition, most of students (95.70%) believe that organization of educational artifacts integrated with wiki reflects a convenient fashion of knowledge dissemination (M = 4.07, SD = 0.570). Finally, 87% of them strongly agree or agree that the manner in which course topics are presented in educational artifacts meets students' expectations (M = 3.75, SD = 0.689).

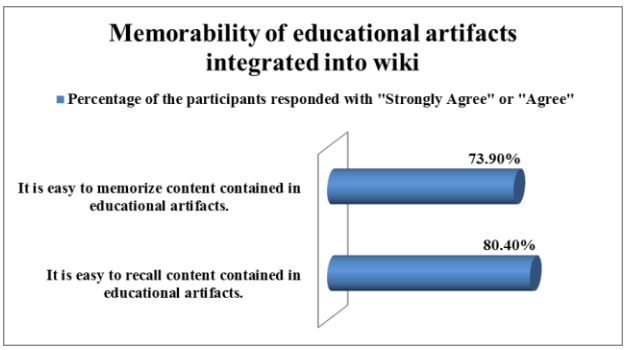


Figure 10. Memorability of educational artifacts on the wiki

The findings shown in Figure 10 illustrate the high level of educational artifacts' memorability. In particular, majority of pilot study participants strongly agree or agree that it is easy to both memorize (M = 3.49, SD = 1.094) and recall (M = 3.66, SD = 0.855) content contained in educational artifacts.

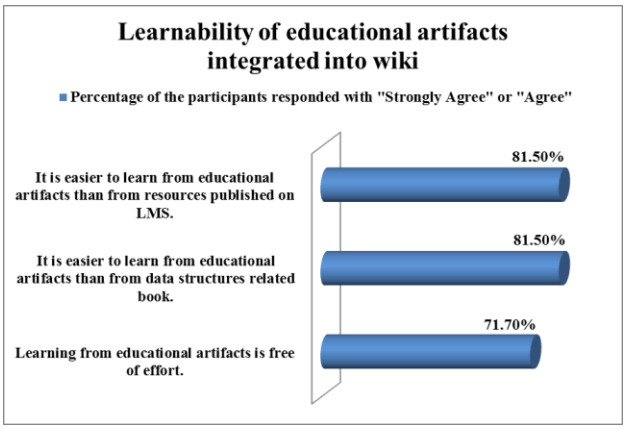


Figure 11. Learnability of educational artifacts integrated into wiki

The data presented in Figure 11 indicate that it is easy for students to learn course topics from educational artifacts integrated into wiki. Namely, most of students (71.70%) strongly agree or agree that learning from educational artifacts is free of effort (M = 3.42, SD = 0.997). Moreover, majority of them believe that it is easier to learn from educational artifacts than from resources

published on LMS (M = 3.70, SD = 0.958) or from data structures related books (M = 3.85, SD = 1.157).

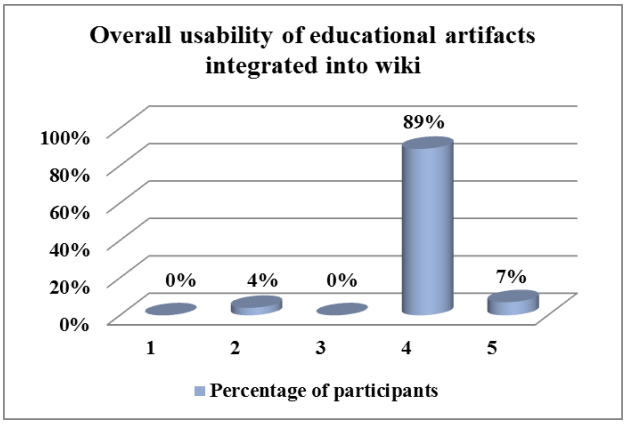


Figure 12. Overall usability of educational artifacts integrated into wiki

According to the results presented in Figure 12, the overall usability of educational artifacts integrated into wiki is perceived as fairly high (M = 3.98, SD = 0.491). The set forth indicate that students support the use of artifacts created by means of Web 2.0 applications for educational purposes.

V. CONCLUSION

Implementation of Web 2.0 applications in e-learning environment brings numerous benefits to students. The concept of e-learning 2.0 enables students to solely or collaboratively create, remix, and use educational artifacts according to their own needs. Pilot study results indicate that completeness, usefulness, availability, learning flow, content quality, added value, adaptability, presentation quality, memorability, and learnability are distinguishable facets that determine perceived usability of educational artifacts. In order to examine the robustness of findings presented in this paper, more studies should be conducted. Therefore, our future work will be focused on revision and validation of proposed attributes as well as development of a comprehensive framework meant for usability evaluation of educational artifacts created with Web 2.0 applications.

REFERENCES

[1] T. O'Reilly, "What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software", 2005, http://oreilly.com/web2/archive/what-is-web-20.html

[2] S. Downes, "E-learning 2.0", eLearn Magazine, 2005, http://elearnmag.acm.org/featured.cfm?aid=1104968

[3] G. Siemens, "Connectivism: A Learning Theory for the Digital Age", 2004, http://www.elearnspace.org/Articles/connectivism.htm .

[4] ISO/IEC 25010, "Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) -System and software quality models", 2011.

[5] C. Ardit, M. F. Costabile, M. De Marsico, R. Lanzilotti, S. Levialdi, T. Roselli, V. Rossano, "An approach to usability evaluation of e-learning applications", Universal Access in the Information Society, vol. 4, no. 3, pp. 270-283, 2006.

[6] P. Nokelainen, "Conceptual definition of the technical and pedagogical usability criteria for digital learning material", Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications. Chesapeake: AACE, 2004, pp. 4249-4254.

[7] C. Cappiello, F. Daniel, A. Koschmider, M. Matera, and M. Picozzi, "A Quality Model for Mashups", Lecture Notes in Computer Science, vol. 6757, pp. 137–151, 2011.

[8] T. Orehovački, "Development of a Methodology for Evaluating the Quality in Use of Web 2.0 Applications", Lecture Notes in Computer Science, vol. 6949, pp. 382-385, 2011.

[9] T. Orehovački, "Proposal for a Set of Quality Attributes Relevant for Web 2.0 Application Success", Proceedings of the 32nd International Conference on Information Technology Interfaces. Cavtat: IEEE Press, 2010, pp. 319-326.

[10] T. Orehovački, "Perceived Quality of Cloud Based Applications for Collaborative Writing", Information Systems Development – Business Systems and Services: Modeling and Development, J. Pokorny, V. Repa, K. Richta, W. Wojtkowski, H. Linger, C. Barry, M. Lang, Eds. Heidelberg, Springer, 2011, pp. 575–586.

[11] T. Orehovački, A. Granić, and D. Kermek, "Exploring the Quality in Use of Web 2.0 Applications: The Case of Mind Mapping Services", Lecture Notes in Computer Science, vol. 7059, pp. 266–277, 2012.

[12] T. Orehovački, G. Bubaš, and A. Kovačić, "Taxonomy of Web 2.0 Applications with Educational Potential", Transformation in Teaching: Social Media Strategies in Higher Education, C. Cheal, J. Coughlin, and S. Moore, Eds. Santa Rosa: Informing Science Press, 2012, pp. 43–72. (12,15)

[13] G. Bubaš, A. Ćorić, and T. Orehovački, "The integration of students' artifacts created with Web 2.0 tools into Moodle, blog, wiki, e-portfolio and Ning", Proceedings of the 34th International Convention MIPRO. Opatija: IEEE Press, 2011, pp. 1084-1089.

[14] C. Hsu, H. Wang., C. Xie, and H. Tzeng, H., "Using classified Social Bookmarking of network knowledge to improve elementary students' science problem-solving ability", Proceedings of 3rd IEEE International Conference on Computer Science and Information Technology. Chengdu: IEEE Press, 2010, pp. 72-77.

[15] T. Orehovački, G. Bubaš, and M. Konecki, "Web 2.0 in Education and Potential Factors of Web 2.0 Use by Students of Information Systems", Proceedings of the 31st International Conference on Information Technology Interfaces. Cavtat: IEEE Press, 2009, pp. 443-448.

[16] T. Orehovački, M. Konecki, and D. Radošević, "Web 2.0 technologies in university education", Proceedings of the 31st MIPRO International Convention on Computers in Education. Opatija: MIPRO, 2008, pp. 269-273.

Parameter Identification of Transfer Functions Using MATLAB

Mato Fruk, Goran Vujisić, Tomislav Špoljarić
Department of Electrical Engineering
The Polytechnic of Zagreb
Konavoska 2, Zagreb, Croatia
mato.fruk@tvz.hr, **gvujisic@tvz.hr**, **tomislav.spoljaric@tvz.hr**

ABSTRACT - This paper describes a procedure for identifying the transfer function parameters by using Matlab's System Identification Toolbox (SIT). The procedure is explained for obtaining numerical data using a digital oscilloscope in a format suitable for loading the data into Matlab and processing within the SIT.

The measured data series connected RLC circuits were analyzed. Also, using the speed response of separately excited DC motor with and without smoothing coil was measured, and transfer functions were obtained.

I. INTRODUCTION

Sometimes, in process analysis, there is a lack of possibility to access certain parts of the examined system, whether the structure of the system is not known, or the knowledge of mathematical model needed for system analysis is insufficient. In that case, the process needs to be tested with known function, and time response needs to be obtained and handled with either conventional or modern techniques.

This article describes the way in which Matlab® and Microsoft Excell® are connected. Measurements are made with digital oscilloscope. Input and output signal processing is examined before testing and parameter identification of transfer functions of the system is calculated via Matlab's System Identification Toolbox. A set of examples including voltage response in RLC circuit and speed response of separately excited DC motor is being used.

II. PROCESS ANALYSIS

Process analysis is defined as an iterative procedure that combines theoretical analysis and experimental examination.

Mathematical model of the process is obtained by means of input-output process signals. The advantage in using experiments is that they do not require detailed specific knowledge of researched system, and simple mathematical models that describe the system can be given as a result. The disadvantage of this analysis is that an earlier research system must exist as well as the fact that the results have limited application for similar systems.

Input signal is chosen on the basis of knowing the process *a priori*. When using parameter procedures (model structure is unknown), iterative calculation is used for obtaining the element order and its dead time (transportation lag). Parameter estimation is being repeated until process and model are successfully matched. For identification of continuous processes the signal must be discretized, so it could be processed through various numerical methods. When using discrete signals, one must take into consideration elements such as sampling frequency, number of bits for coding the received signal and quantization noise.

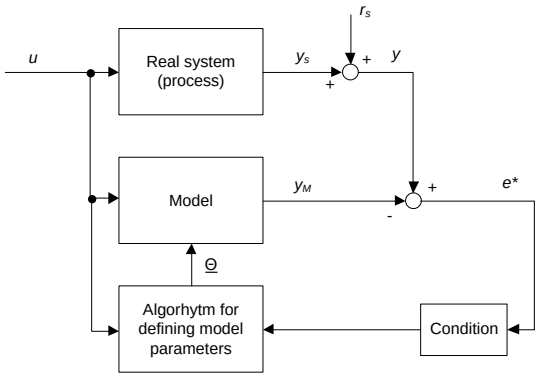


Fig. 1 Identification based on parameter estimation

III. PARAMETER IDENTIFICATION IN MATLAB

Parameter identification can be calculated by using System identification toolbox (SIT), a tool that exists under Matlab since version 6.5. SIT is a tool for creating mathematical models, and is based on collecting input/output data of the system. It has a GUI that is user friendly in model building. Certain knowledge in system modelling is needed when choosing the model in SIT, because mathematical models process the measured data.

Model can be defined by adjusting settings of the assumed process model, until modelling is satisfactory accurate according to the input data in SIT package.

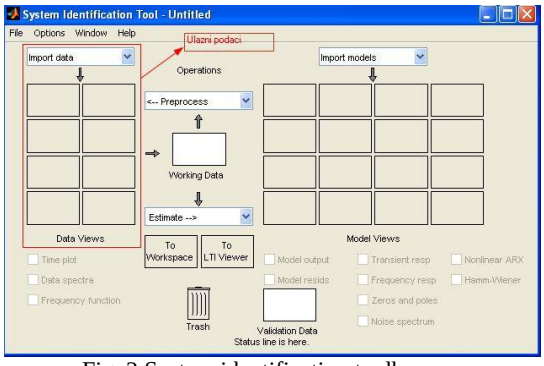


Fig. 2 System identification toolbox

Fig. 2 shows a graphical interface. On the left side of the interface there is a space for entering the import data which need to be entered in associated windows. This can be done by using the following command:

data = iddata (output,input,T_s)

In the above mentioned case import data (Fig. 9) will be equal to *data*, and *iddata* is the command which enters data in System Identification Toolbox. In parentheses, *output* defines measured output data, *input* defines measured input data, while T_s defines sampling time (time base for collecting samples by means of the measuring instrument used). One must take into consideration that matrices containing input and output data must have the exact number of rows. This means that the quantity of output data must be the same as the quantity of input data. It is not necessary that one set of input data gives one set of output data, but it is necessary that matrices have the following form:

$$[output] = [n]\text{x}[k]$$
$$[input] = [n]\text{x}[1]$$

A. CONNECTION BETWEEN MATLAB AND MICROSOFT EXCELL

Connecting of the two packages mentioned above takes place when measured data are acquired in the format that is not compatible with Matlab based processing. Most digital oscilloscopes have the ability to save measured data as documents in formats such as text or CSV (comma separated values). These formats are compatible with Microsoft Excell, and Microsoft Excell has the ability (in form of add-in) to forward this data to Matlab. In Excell, when „Excell options“ window is opened, mentioned add-in is selected and then button „Go...“ needs to be clicked (Fig. 3).

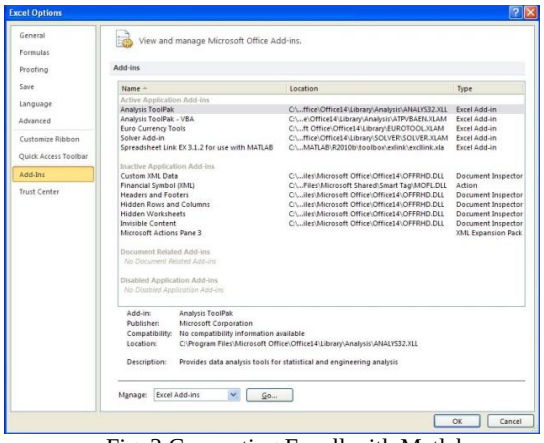


Fig. 3 Connecting Excell with Matlab

Then, on the next window click „Browse...“ and find the directory containing installed Matlab package. Inside that directory there is a „toolbox“ directory which contains „exlink“ directory. Inside „exlink“ directory there is a file named „exclink.xla“ and this file is a software expansion used for connecting Excell with Matlab (Fig. 4).

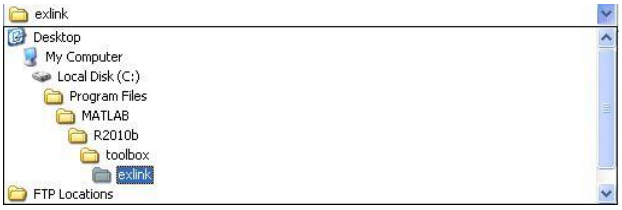


Fig. 4 Path to „exclink.xla“

When add-in is installed, in Excell options window, under „Add-Ins“ options, a package for connecting with Matlab will appear, which gives the possibility to convert matrices from Excell format to Matlab format and vice versa (Fig. 5).

Fig. 5 Access toolbar in Excell for *exclink* add-in

Furthermore, an explanation for loading data from Excell to Matlab's workspace will be given, based on the example of measured data on real system (RLC circuit)

B. PARAMETER IDENTIFICATION BASED ON REAL SYSTEM (RLC CIRCUIT)

Input and output data for RLC series circuit have been acquired by using digital oscilloscope, where output data are observed as voltage on capacitor. Numerical data are saved in *.txt* format. On the basis of time response of voltage drop on capacitor parameters have been defined using conventional methods, and then given as data to System Identification Toolbox. The results have been compared. Transfer function of the RLC series circuit is as follows:

$$G_{RLC}(s) = \frac{K}{1+2\zeta\, T_w s + T_w{}^2 s^2} = \frac{K}{1+\frac{2\zeta}{\omega_n}s+\frac{s^2}{\omega_n^2}} = \frac{K\omega_n^2}{\omega_n^2+2\zeta\,\omega_n s+s^2} \qquad (1)$$

where:

$$\omega_n = 1/T_w = 1/\sqrt{LC}\ ,\ 2\zeta\omega_n = R/L$$

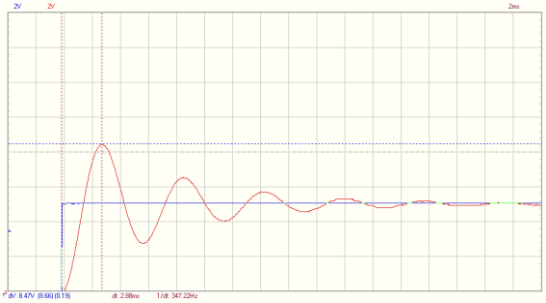


Fig. 6 Time response of voltage drop on capacitor of the RLC series circuit

Fig. 6 shows that parameters of the second order transfer function are:

- K = 1 (gain),
- t_p = 3 ms (moment of peak overshoot),
- y_m = 8,5 V (peak overshoot),
- y_{st} = 5 V (steady state value).

From these values parameters of the second order transfer function can be defined by these formulas:

$$\sigma_m = \frac{y_m - y_{st}}{y_{st}} = e^{-\frac{\pi\zeta}{\sqrt{1-\zeta^2}}};\ \zeta = 0{,}1128;\ \ \omega_n = \frac{\pi}{t_p\sqrt{1+\zeta^{-2}}} = 1074\ \frac{rad}{s} \qquad (2)$$

The second order transfer function is as follows:

$$G_{RLC} = \frac{1074}{1074^2 + 2\cdot 0{,}1128\cdot 1074\cdot s + s^2} = \frac{1}{1+21\cdot 10^{-3}s+86{,}7\cdot 10^{-6}s^2} \qquad (3)$$

By using SIT to obtain transfer function, the procedure is faster and simpler than using conventional methods. This procedure is explained in next few steps. First, all data are being transferred from Excell to Matlab's

workspace. This can be done by selecting data matrix that is to be transferred from Excell, and by clicking the „putmatrix“ button. This command executes the action of transferring data matrix to Matlab, after typing desired name for Matlab variable (Fig. 5, 7).

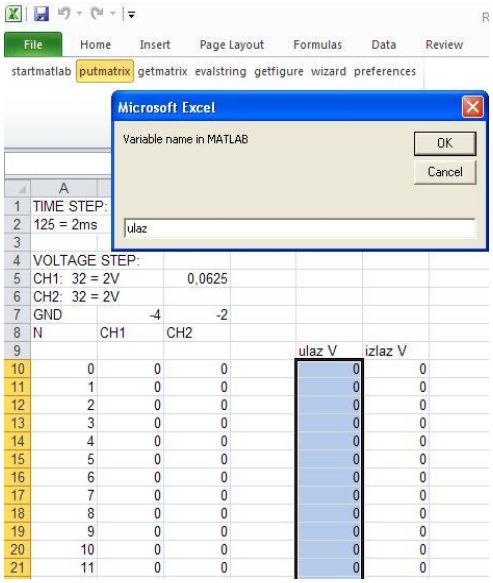


Fig. 7 Transferring data matrix from Excell to Matlab

Output data need to be transferred after transferring the input data. In Fig. 9 window shows that it is needed to define sampling frequency (sampling period) in SIT package. Furthermore, inside command line on Matlab's workspace the command for data entering needs to be typed in, as follows:

RLC = iddata(izlaz,ulaz,0.002/125)

These three variables (*RLC*, *ulaz*, *izlaz*) are being held as data in Matlab's workspace. Inside the workspace, with ***ident*** command, a graphical interface can be opened. Created and adjusted data need to be entered in this interface.

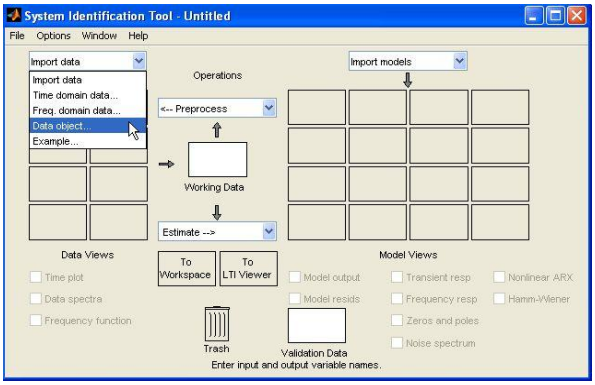


Fig. 8 Entering data in SIT

After pressing „Data object“ in the drop-down menu, a window will appear, in which the name of data set for SIT needs to be entered. In workspace variable the name of a variable for further data processing needs to be typed in. If the variable is accurately structured, SIT will automatically set the „sampling interval“, while for „Starting time“ a value 0 is preferred for RLC series circuit (Fig. 9).

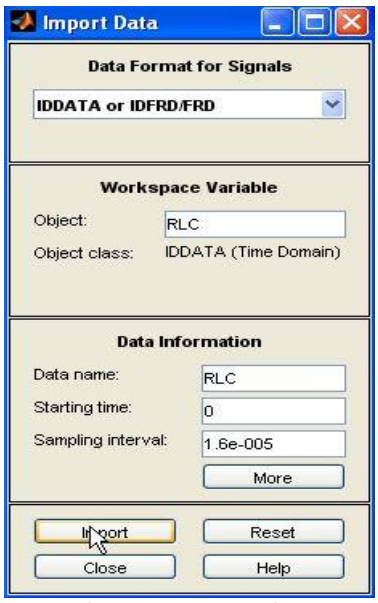


Fig. 9. Data entering in SIT

Now, inside the main menu of SIT, *RLC* variable is entered. *RLC* is also found inside „Working data“ and „Validation data“ windows. However, it is impossible with oscilloscope to save only a useful part of the response, one need to adjust interval for data processing by using preprocesses.

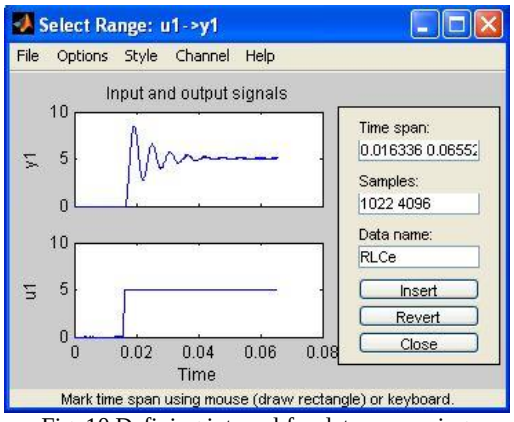


Fig. 10 Defining interval for data processing

After preprocessing, input data for SIT can be mutually compared in three categories: „Time plot“, „Data spectra“ and „Frequency function“. From these graphs a most suitable set of input data can be chosen to process.

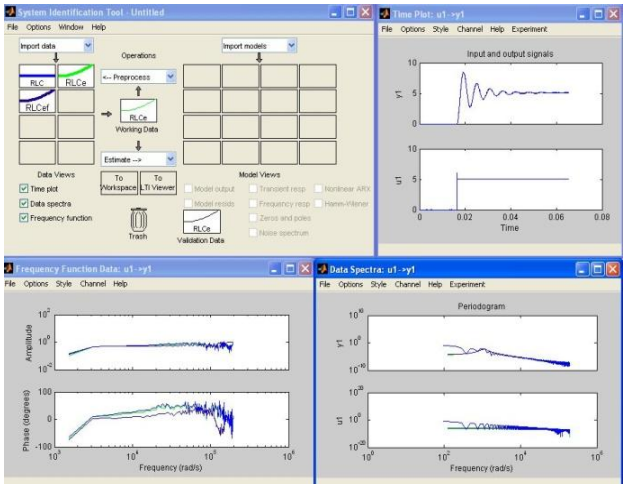


Fig. 11 Possibilities of using preprocess analysis of input data in SIT

Fig. 11 shows that the most suitable set of data is green pattern - *RLCe*. First sample, named *RLC*, is not suitable because it has a number of higher overtones, it does not start with step function and it has a transportation lag. At he beginning *RLCef* has a high overshoot in step function. Due to these criteria the most suitable signal for observation is *RLCe*. Furthermore, it can be concluded that it is not always desirable to go to great lengths when adjusting signal for processing.

The processing itself, in SIT, begins in drop-down menu, when choosing „Working data“ where one can choose process models. A window appears, used for defining process model estimation.

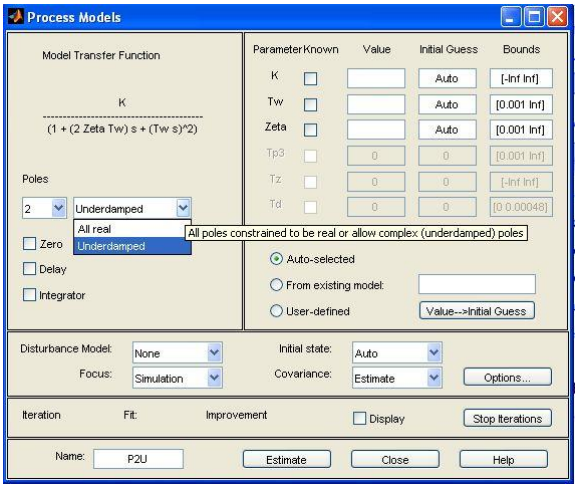


Fig. 12 Process models

Fig. 12 shows a process model used for transfer function parameters calculation. Before using this process model, knowledge of model itself is needed. Fig. 6 shows a model that is similar to the second order element with exponentially damped response. It is known that this element has a pair of complex conjugated roots, and so the parameters are set, as shown in Fig. 12. Precaution needs to be taken by carefully setting „Bounds“. It can happen that SIT calculates coefficients that are set on a boundary. This means that either a model is not chosen properly, or boundaries need to be expanded. After these parameters are set, by pressing „Estimate“ SIT calculates transfer function parameters for defined model. Fig. 13 shows the elements that appear in abovementioned window. This window now contains system identification after model processing.

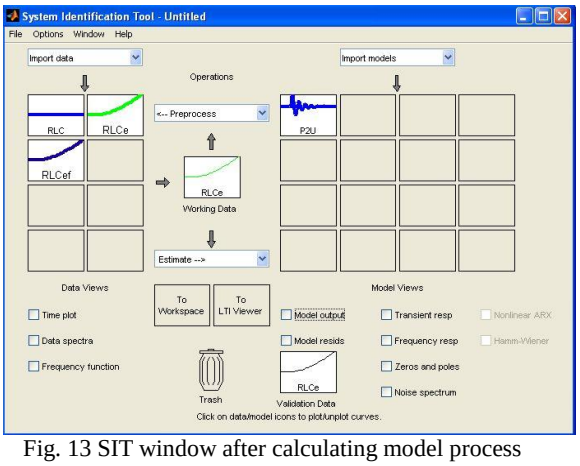


Fig. 13 SIT window after calculating model process

On the right side of the window, it is clear that previously analyzed model still exists. By pressing „Model output“, a comparison between calculated and measured data will be shown, and in the right corner of the window the accuracy of calculated model due to measured data will appear. If accuracy is satisfactory, the graphs can be calculated by using LTI Viewer, or the model itself can be transferred into Matlab's workspace.

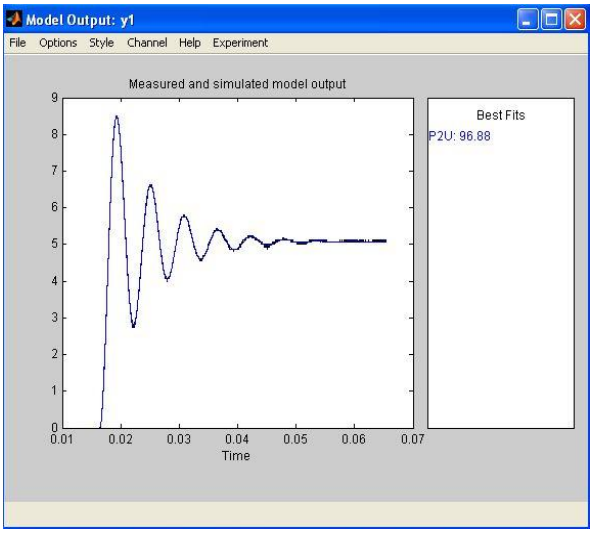


Fig. 14 A comparison between input data and calculated model

Accuracy is measured with relative difference between measured and calculated data, and is equal to 96.88%. By double clicking „P2U“ icon, a window containing parameters K_p, T_w, ζ and a transfer function of the model will appear (Fig. 15).

By comparing parameters gained from conventional method and parameters calculated by SIT, it can be seen that no significant difference between them exists. Final transfer function is as follows:

$$G_{RLC} = \frac{1,0031}{1+\frac{2\cdot 0,12452}{1095}s+\frac{s^2}{1095^2}} = \frac{1,0031}{1+2,27\cdot 10^{-4}s+8,34\cdot 10^{-7}s^2} \qquad (4)$$

Small difference in data exists because a calculation based on graph reading is not completely accurate, a quantization noise exists, and noise present on oscilloscope probe during the measuring also exists. Fig. 7 shows minimum voltage difference between two consequent samples that equals 0.0625 V.

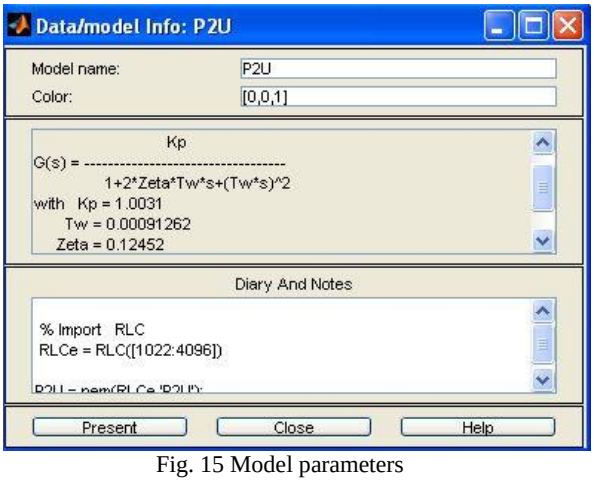


Fig. 15 Model parameters

Because a simple laboratory model is used, capacitor voltage does not show any significant distortion (noise). However, larger and more complex models are more subject to inaccuracies of measuring elements and distortions that result from electromagnetic disturbances in signal cables.

C. PARAMETER IDENTIFICATION BASED ON REAL SYSTEM (SEPARATELY EXCITED DC MOTOR)

As the second example for data processing, a model of separately excited DC motor is chosen. This type of motor is simply regulated and it has linear static characteristics, therefore it is suitable for educating purposes as the second order element with damping time response. Its transfer function is:

$$G_{m1}(s)=\frac{\omega(s)}{U_a(s)}=\frac{1}{K}\cdot\frac{1}{1+T_m s+T_a T_m s^2},(uz\ M_t(s)=0) \quad (5)$$

where:

- $T_a = L_a/R_a$ – armature time constant,
- $T_m = JR_a/K^2$ – electromechanic time constant,
- K – motor constant.

The following relations are also applied:

$$\omega_n=\frac{1}{\sqrt{T_a\cdot T_m}},\ \zeta=\frac{1}{2}\sqrt{\frac{T_m}{T_a}} \quad (6)$$

When experimental methods are used, motor with following characteristics is taken into consideration:

- $P_n = 750$ W,
- $i_u = 0{,}3$ A,
- $U_{an} = 90$ V,
- $I_{an} = 10{,}7$ A,
- $n_n = 1750$ min^{-1}.

Measuring the speed change due to armature voltage step from $U_{a1} = 46$ V to $U_{a2} = 56$ V. DC tacho-generator is used for speed measurement. When stepping the voltage from U_{a1} to U_{a2}, speed increases from $n_1 = 1200$ min^{-1} to $n_2 = 1480$ min^{-1}.

Time response of speed change is observed with RIGOL DS1052D oscilloscope and is shown in Fig. 16.

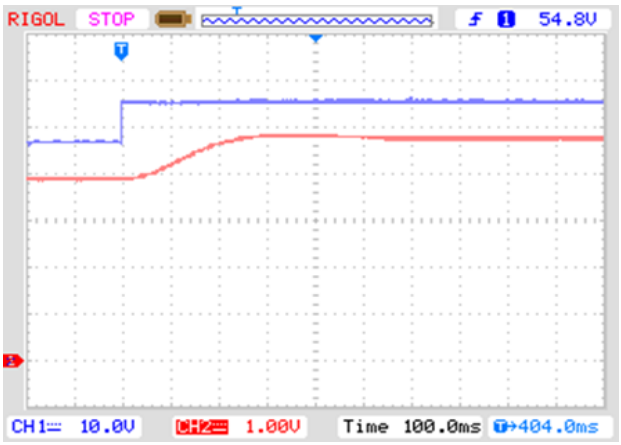


Fig. 16 Speed response due to armature voltage step

Fig. 16 shows second order speed response with strong damping. When calculating in SIT, it is important to select correct model process, one that suits this type of response. Otherwise, incorrect results will appear. Data must be set to be processed from the first moment when voltage step takes place, until steady state value of speed response is given (setting the preprocess). There is also a high frequency noise from DC tacho-generator, so it is necessary to filter input data, and with caution that the shape of voltage step remains authentic.

After analyzing input data, the obtained parameters are shown in Fig. 17.

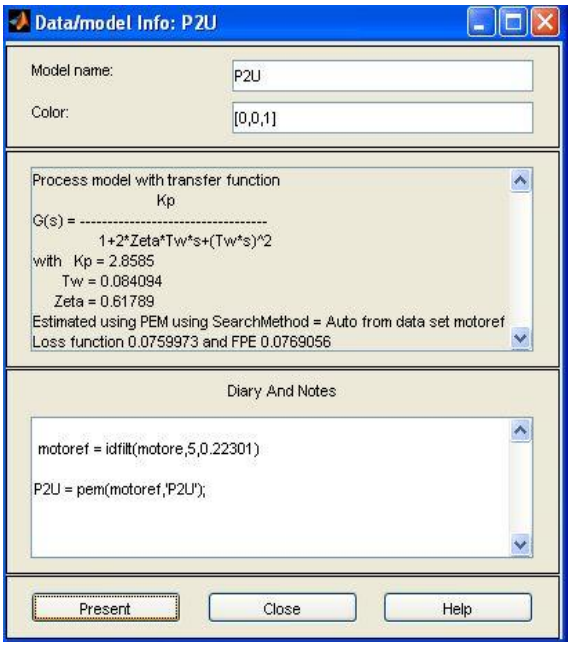


Fig. 17 Parameters of a DC motor with load

Transfer function (6) parameters resulting in SIT calculation are as follows:

$$K=\frac{1}{K_p}=0{,}34983\approx 0{,}35\ Vs;\ T_m=2\zeta T_w=0{,}104s;\ T_a=\frac{T_w^2}{T_m}=0{,}068s$$

This result is consistent with data results from conventional methods (T_a – from the response to the step voltage with locked rotor, T_m – determining the moment of inertia from stopping procedure, K – static equations).

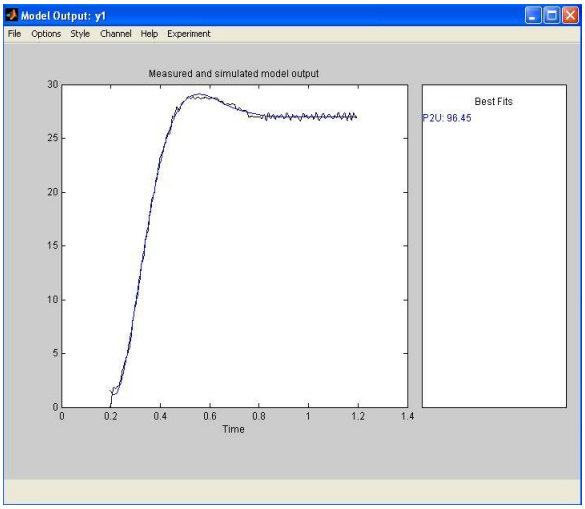


Fig. 18 Time response of speed gained by SIT calculation and input data of observed DC motor

When observing „Model output“ window, it is clear that SIT calculation is consistent with the real system

results with small deviation – the accuracy rate is 96.45%. Since it is impossible to filter high frequency noises completely, so more accurate results cannot be obtained (Fig. 18).

IV. CONCLUSION

This article describes the approach in parameter identification of linear processes with single input and single output (SISO) by using System Identification Toolbox from Matlab software package. In order to identify the model correctly, it is necessary to record results in digital form accurately and precisely. Measured data can be saved as text or in CSV format. Quantization noise of A/D conversion and the noise from measuring equipment also need to be taken into account.

After loading data into Matlab package, it has to be limited to number of samples needed for time response observation, and high frequency noises need to be filtered afterwards. Sampling frequency has to be high enough, but it must not exceed its limits. This means that two consequent samples have to be different enough by quant. Otherwise, a problem with parameter identification occurs. The whole algorithm of identification can be improved by using DAQ cards, because they have the option of signal filtering, thus giving more accuracy.

Bounds need to be taken into account in „Process models“. It could happen that SIT calculates coefficients that are on a boundary, thus meaning that incorrect model is chosen or bounds need to be expanded.

Comparing with conventional methods, this identification process is faster and more accurate. For identification with SIT package, correct type of transfer function needs to be chosen according to the response. Accuracy of chosen transfer function calculation needs to be compared with input data given to Matlab. Also, on the basis of transfer function, Matlab can give graphic results in the form of output response, impulse response and Bode diagrams.

„Process Models“ gives the opportunity to choose 44 different types of transfer functions.

BIBLIOGRAPHY

[1] L. Ljung; System identification toolbox, Mathworks 1997.

[2] http://www.mathworks.com/help/toolbox/ident/ , 10. 6. 2012.

[3] N. Perić, I. Petrović; Identifikacija procesa, predavanja, FER, Zagreb, 2005.

Web Interactive Learning
Case Study: Faculty of Economics of Josip Juraj Strossmayer University in Osijek

Maja Lamza-Maronić*, Jerko Glavaš **
* Faculty of Economics in Osijek, Osijek, Croatia
** Faculty of Economics in Osijek, Osijek, Croatia
maja@efos.hr

Abstract - **Interactive learning as a method can definitely contribute to the quality of higher education by using advanced technologies. It has been proven that interactive learning outperforms traditional ways of teaching. In addition, the time required for participants to adapt to new methods is minimal. Interactive teaching generates skills that are equally important for academic, social and business success. To achieve better connection of academic community with its business environment, it is necessary to continuously, actively and intensively confront the competition and the challenges of the growing national and global market of education. Harmonious use of ICT in tertiary education provides possibilities for efficient communication of the entire university virtual environment for learning and closer cooperation between universities and the business world. The paper will elaborate on possibilities of applying modern web technologies in tertiary education using the example of the Faculty of Economics in Osijek, specifically in the course Urban Economics. The aim of this novel approach is to improve interactive communication, and to make broader use of web technologies in the process of higher education.**

I. INTRODUCTION

In a world of dynamic change, globalization and advanced technology, higher education and education in general, undergo constant change. Entirely new structures and areas of education are created. Completely new methods of education are developed; traditional ones disappear or change, while the environment, working conditions and needs change with each passing day. As a result, the choice of career and education has become increasingly difficult. To know the world of labour means to become familiar with different occupations, their main features, requirements, conditions and employment prospects. Therefore, the current choice of career or education does not necessarily have to be for life but changes during different stages of life.

The Republic of Croatia has begun an ambitious reform of the education system based on the "2005-2010 Education Sector Development Plan" (Government of the Republic of Croatia, 2012) and Croatian National Educational Standard (essential section) adopted by the Government of the Republic of Croatia in 2005. It sets out four key development priorities: (1) improve the quality and efficiency of education, (2) promote continuous professional advancement of teachers and other employees in the education sector, (3) develop strategies for management of the education system and its efficiency, (4) education for social cohesion, economic growth and development. The concrete, already adopted, education system development goals are an increase in the ratio of enrolment and graduation rates, participation in tertiary education, increase in public and private expenditures, increase in the education level of early school-leavers, as well as better and more efficient organization of education and measuring of academic achievement.

In accordance with the above, the choice of education is regarded as the allocation of occupation. The decision to choose an occupation is affected, inter alia, by economic determinants (economic structure, the state of the economy, labour market, occupational structure, etc.) as well as socio-cultural and socio-psychological determinants (belonging to a particular class, family, school, peers, career counselling).
A decision on one's occupation and study choice after primary and secondary education is certainly one of the most important decisions in life. Regardless of its importance, if one happens to make a wrong choice, one should bear in mind and develop the idea of lifelong learning, training, retraining, etc.
It has become a necessity to invest a great deal in the training and development of individuals, in other words to invest in knowledge, because knowledge is the main source of competitive advantage and value added. This is why education and lifelong learning are the key drivers of social mobility.
European higher education institutions, including the University of Josip Juraj Strossmayer in Osijek, have accepted the challenge of developing the European system of education. In line with the initiative of the European Employment Strategy "New Skills for New Jobs" (European Commission, 2008), the vision of the Republic of Croatia is becoming increasingly dynamic. The Republic of Croatia needs to create a social and political framework and a challenging business

environment in which the majority of employed persons would have an opportunity to participate in the global economy through:

- Promoting and anticipating future skill needs;
- Developing consistency between individual skills and labour market needs;
- Bridging the gap between the world of education and the world of labour.

The above categories need to be analysed in their own environment and their implementation should be proactive to ensure further development of the higher education and the development of skills needed in the modern market conditions.

Educational structure of younger age groups is much better than that of the overall population, but it still needs improvement in terms of the following: increasing pre-school education coverage; increasing the rate of children and youth who successfully complete programmes they enrolled into; analysing the causes and preventing dropouts; greater consistency between curriculums and the needs of the labour market; increasing the percentage of university graduates and introducing new teaching methods into the teaching process.

Continuous accumulation of information requires one to filter it and choose that information which is essential for achieving the set objective function. To know how to use the information is more important than just to store it as a set of facts. In this context, a set of cognitive, affective, social and work place competencies is created.

Interactive teaching also develops essential skills such as communication skills and computer literacy, desire to improve performance in the pursuit of the highest quality, and taking the responsibility in the search for the objective function.

Interactive learning is a necessary and desirable aspect of the modernization of the traditional teaching. Interactive Internet-based teaching and learning provides a new dimension of flexibility in teaching. Its main advantages are choice, suitability and flexibility of the set tasks through interactive learning. This method of teaching changes the traditional position of the teacher and his team (teaching assistant, teacher's aide). Now they need to ensure continuous interaction and provide for interesting course contents. Teachers and their teams need to be stimulated and educated for this type of work with students.

Teachers have lost their traditionally dominant role. Nevertheless, they must keep this role so that they could properly conduct the teaching process. Their active role as lecturers has now changed into a passive role as guides of the teaching process, but now they need to develop interactive course contents. By creating interactive contents, teachers guide students towards perceiving the vital relationships and contents. The quality of interactive contents must not jeopardize the benefits provided by Web-based learning, because it must guide the students to a given objective function.

The teachers' role in the communication process has changed as compared to the traditional teaching, because teachers no longer have an active role as mediators between the course contents and students.

Teachers must now focus more on student performance assessment and adopt a more individualised approach to students. Their attention is now primarily directed at:

- Accuracy and consistency of lectures;
- Independent work, in terms of controlling the production and limited time for delivery;
- More frequent assessments (on a weekly basis)
- Individual communication with students, which, to the extent possible, should be organised on a daily basis.

Despite the popularity of the idea that the Internet is an interactive and educational medium, interaction via online learning must always seek to increase student motivation.

One of the modern teaching methods is an interactive Web-based system which was developed and has been used at Josip Juraj Strossmayer University of Osijek, Faculty of Economics, in an undergraduate course in *Urban Economics*.

II. CASE STUDY: THE FACULTY OF ECONOMICS OF J.J. STROSSMAYER UNIVERSITY IN OSIJEK

Research studies have shown that the best results in learning are achieved by a combination of passive and interactive learning. Unlike so-called passive learning, during which students mostly "observe" the learning process (by listening), interactive learning directly requires feedback information, i.e. active participation in the teaching process. Interactive learning is especially important for the development of cognitive abilities of students, which are crucial in modern management as well as in living environment. Harmonious application of ICT in academic education enables communication and efficient cooperation of the entire virtual environment of the university in the learning process, thus increasing the always necessary contactability of cooperation between the university and the economy.

Traditional education, which is based on education paradigm of knowledge reproduction, has been upgraded by means of a dynamised system in the teaching process based on application of information and communication technology.

The system was tested at undergraduate studies at the Faculty of Economics in Osijek, within the course *Urban Economics*: web site (http://www.urbanaekonomija.com/) and it was a part of the realisation of the teaching programme in the mentioned course.

The advantages of the above system are simplicity and low operation and maintenance costs. The system has been created as a version of WordPress system. WordPress is an Open Source Content Management System (CMS) and it is often used as an application for blog publication, supported by PHP and MySQL. WordPress has a range of characteristics, including plug-

in architecture and system template. With these characteristics it has also proved to be a good solution for use in the teaching process as it prefers the development of critical thinking and strengthens the role of team work. The possibility of using a web site (http:/www.urbanaekonomija.com/) within the education process has resulted in the following advantages:

- Possibility of strengthening the relation between theory and practice
- Application of interactive learning
- Platform for development of new creative ideas and seeking for possible solutions
- Joint (team) work on tasks and projects
- Opening space for discussion and comments on students' papers
- Expanding the source and theoretical knowledge by using the published online edition of course curricula (as e-textbooks)

The above advantages are prerequisites for expansion and possibilities of use of the described educational model.

The process of creating the system for web-based interactive realisation of curriculum of the course Urban Economics has been divided in certain phases, the most significant ones among them being:

1. Analysis of the current condition and current needs and interests - *realised*
2. Project realisation: application of the model of web-based interactive learning - *realised*
3. Monitoring project performance and justification – *realised / continued.*

In this structure the third phase arises, which is characterised by continuous improvement and monitoring of the system in all its segments. This segment is carried out by the team executing the course Urban Economics according to the curriculum: the teaching staff and associates: teacher's aides and students through continuous monitoring of the established system.

Indicators of successful implementation of the above system were monitored by means of a survey conducted among students who were logged in the system (N=125) in multiple iterations. The survey was conducted in the duration of the above mentioned course (on term basis). According to the analysis of the obtained results, it can be observed that most of the students were satisfied with system implementation. Here is the analysis of indicators related to the selection of a theme, searching the available literature, topic disposition, process of research and elaboration of the selected theme, and entering the themes (of students' papers) into the system.

Based on the information from *Google Analytics,* a system for tracking visits to a web site, it can be observed that the top seven themes in the system were viewed 5039 times in the academic year 2011/2012 and that the theme was downloaded 24 times. Further, these themes received 168 *likes* on Facebook, one of the most popular social networks.

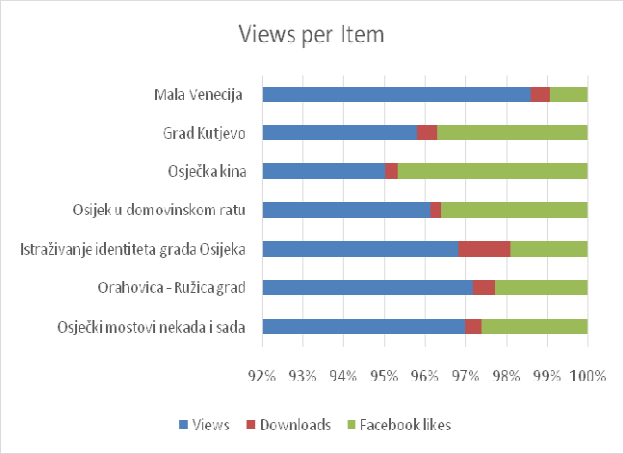
Table 1. Total statistics of views of the web content Urban Economics – category: students' papers

Total Views (7 presentations)	Total Downloads	Total **Facebook** likes
5039	**24**	**168**

Source: Developed by authors

Based on the presented indicators, each of the seven themes can be analysed separately. The following themes and papers attracted most attention and received the greatest number of suggestions and comments, thus occupying the first place according to the level of interest: *Cinemas in Osijek* and *Osijek in the Homeland War,* with more than 1000 views. These themes have also received the highest number of Facebook *likes,* whereas the most downloaded theme was *Researching the identity of the City of Osijek.*

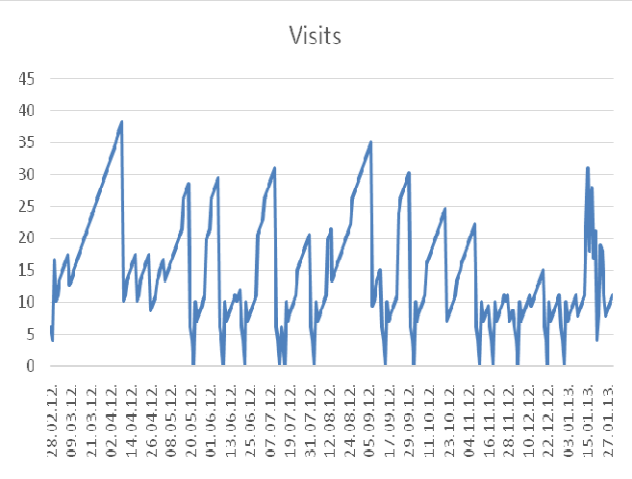
Graph 1. Individual statistics of the observed top 7 themes



Source: Developed by authors

According to the analysis of the number of visits to the web-based interactive system (Graph 2), below are the main intervals in which the web site was visited in the period from February 2012 to January 2013.

Graph 2. Visits to the web-based interactive system *Urban Economics* (period February 2012 – January 2013)



Source: Developed by authors

It can be observed from the graph that most visits to the web system for the purpose of viewing the themes took place during the course (March – April 2012), during the regular exam period (July, September 2012), and in the period when the beginning of the teaching process in the course *Urban Economics* in the new academic year 2012/2013 was expected (January, February 2013).

Based on the conducted analysis and performance of the established system (web site), a conclusion can be made that all listed processes ultimately create synergy, which results in better exam pass rates. Increased possibility of contactability, together with the adoption of new knowledge and skills, shows that the presented educational system with web-based interactive learning was successfully applied.

III. CONLUSION

Interactive learning is a necessary and desirable aspect of the modernization of the traditional teaching. Interactive Internet-based teaching and learning provides a new dimension of flexibility in teaching. Its main advantages are choice, suitability and flexibility of the set tasks through interactive learning. Teachers have lost their traditionally dominant role. Nevertheless, they must keep this role so that they could properly conduct the teaching process. Their active role as lecturers has now changed into a passive role as guides of the teaching process, but now they need to develop interactive course contents. Teachers must now focus more on student performance assessment and adopt a more individualised approach to students.

REFERENCES

[1] Čičin-Šain, M.; Marinović, M.; Pearce, C.E.M., Ljudski čimbenik u informacijskoj tehnologiji // Cjeloživotno učenje za održivi razvoj / Uzelac Vinka ; Vujičić, Lidija (ur.). Rijeka : Sveučilište u Rijeci, Učiteljski fakultet u Rijeci, 2008.

[2] Jovanovski, V., Razvoj karijere pojedinca u okružju menadžmenta znanja, doktorska disertacija, Ekonomski fakultet Sveučilišta J.J. Strossmayera u Osijeku, Osijek, 2012.

[3] Lamza-Maronić, M.; Glavaš, J.; Horvatin, T., Promoting Knowledge Society through Quality Assurance Model // 35th international convention MIPRO 2012 / Biljanović, P. (ur.). Rijeka : Croatian Society for Information and Communication Technology, Electronics and Microelectronics - MIPRO, 2012.

[4] Peko, A., Mlinarević, V., Jindra, R., Interkulturalno obrazovanje učitelja – što i kako poučavati, U: Izazovi obrazovanja u multikulturalnim sredinama, (Ur.: Peko, A., Mlinarević, V.), Sveučilište J. J. Strossmayera u Osijeku, Učiteljski fakultet u Osijeku i Nansen dijalog Centar Osijek, 2009.

[5] Suzić, N., Pojam i značaj interaktivnog učenja, Nastava, br. 1-2, str. 33-51, Banja Luka, 2003.

The role of computerisation and automation of school libraries management in teaching information literacy on the example of curriculum unit Online Library Catalogues

Denis Vincek
Osnovna škola Ante Kovačića (Primary school)
Vladimira Nazora 1, HR-49250 Zlatar, Croatia
Phone: +385-49-466 832 +385-98-223 746 Fax: +385-49-466 832
E-mail: denis.vincek7@gmail.com

Abstract – Croatian seventh grade elementary students should be completely qualified to independently browse their school library collection using its online library catalogue. Online library catalogues are an excellent example for a pupil to gradually and consciously use a computer in class to acquire information literacy skills, as described by Mike Eisenberg and Bob Berkowitz in their Big6 model. The results of a research conducted among elementary school librarians in the Krapina-Zagorje County in December 2012 indicate that only 14 out of 27 libraries (23 of which use Metelwin library automation software and are connected to the Internet) have their own online library catalogues, and only five librarians refresh their online databases every day. Forty-four percent of elementary school libraries do not own a computer intended for students, or have them but they are not connected to the Internet. Seven out of 27 elementary school librarians in Krapina-Zagorje County do not cover the curriculum unit Online Library Catalogues. The research shows a statistically significant positive correlation between having an online school library catalogue and a computer for students in the school library (and its Internet connection) and teaching information literacy on the example of curriculum unit Online Library Catalogues.

I. INTRODUCTION

One of the eight competences crucial for lifelong learning stated in the National Curriculum for Preschool Education and General Compulsory and Secondary School Education is the digital competence [1].

The National Curriculum states that digital competence entails students' qualification for confident and critical usage of Information Society Technology (IST) for work, leisure and communication. The ability to use the computer to retrieve the information via the Internet is listed as the basic information technology skill.

This paper will show the extent of computerisation and management automation for elementary school libraries in the Krapina-Zagorje County and confirm or reject the hypothesis that school librarians cover the curriculum unit Online Library Catalogues in which a student can be taught, through gradual computer usage, everything that is understood by the complex term of information literacy, a part of the elementary school curriculum. This will be done only in the schools that meet the following criteria:

- the library management is automated, which means that the librarian has his or her own computer with any kind of library automation software installed and Internet connection;
- the library has its own online library catalogue;
- the library has computers intended for student work which are connected to the Internet;

II. ONLINE LIBRARY CATALOGUES AND LIBRARY AND INFORMATION EDUCATION

According to the National Curriculum the students are introduced to the world of information by the school librarian as a regular part of the educational process as early as from the third grade on. The usage of computers in the acquisition of information literacy skills becomes especially apparent in the upper elementary school grades.

For instance, the seventh grade students use computers on a daily basis at home, in the school library, or in the IT classroom, and are generally highly motivated to learn any new computer-related teaching contents. After they have learned the teaching contents of online library catalogues, the students should be completely qualified to independently browse their school library collection using its online library catalogue. Not only will the students fully understand the bibliographic record, but they will be able to find out whether their school library has in its possession the desired book, magazine, audio and video materials, or any other library materials [2].

This generally includes information whether the school library has a certain assigned reading material, whether it is loaned out or not, and when it is due to be returned. In addition, a student can search the catalogue using different criteria, determine whether the library has enough copies of a certain assigned reading material, or

find out about other books written by the authors on the assigned reading lists.

III. METHODOLOGY

In the school year 2012/2013, the Krapina-Zagorje County had 33 elementary schools. Since this is not a large population, the plan was to gather data from all schools. Surveys were chosen as a data gathering method, because of their suitability for descriptive and causative research, but also because their sending costs are minimal and the examiner does not have to be present when the survey is being conducted. It involved 46 closed-type questions which were not too complicated. Since the school librarians are both educational and informational experts with a high level of computer literacy, the survey questionnaires were sent directly to them via e-mail and/or the schools themselves with a note to be forwarded to the school librarians.

The questionnaires were sent on 3rd December 2012 with a request to be returned via e-mail by 10th December. The librarians who had not replied by the given date were reminded of the survey and were asked once more to send the filled out questionnaire. By 23rd December, overall of 27 filled out questionnaires from elementary schools were received (23 via e-mail and 4 by fax machine), which constitutes 82 percent of elementary school population in the Krapina-Zagorje County and provides a highly representative sample, especially given the fact that the acceptable response rate for surveys is 30 percent for organisations, and less than 10 percent for Internet surveys [3].

IV. RESULTS AND DISCUSSION

A. The state of computerisation of school libraries in the Krapina-Zagorje County

All elementary school libraries employ librarians with a university degree in Library science, except for one school which employs a person with a university degree, enrolled in a two-year part time Library and Information studies programme. It should be noted that a small portion of school librarians work part-time, but this factor is not relevant to the data gathered and analyzed in this work.

The teaching contents of Library and Information Education which require computer usage can alternatively be conducted in the IT classrooms. Bearing this in mind, when asked whether their school has such a specialized classroom, 27 elementary school librarians gave a positive answer, which constitutes 100 percent of all answers.

The question whether school libraries have computers intended for student work received more diverse answers. As much as 30 percent of elementary school libraries do not own such computers, which is a shocking piece of information. However, the survey has shown that the remaining 70 percent do not have much to be proud of. Seven libraries have only one computer for student work, eight libraries have two, one library has four, two libraries have six computers, and one library stands out, so to say, with 13 computers intended for student usage.

Of these 19 elementary school libraries with computers for students, computers in four libraries are not connected to the Internet which leads us to conclude that only 15 out of 27 elementary school libraries (56 %) have computers that students can use to access the Internet for learning or entertainment.

When it comes to computers intended for school librarians, every librarian has his or her own computer. Seventeen computers were newly purchased for school librarians, seven were second-hand computers (mostly donations from one of Croatia's leading store brands) and three librarians did not know whether their computers were new or used, as they found them when they started working in their respective schools. So, although 63 percent of computers were newly acquired for the needs of school library administrators, the data clearly shows that they are six to ten years old on average.

B. Automation of management

It should be noted that the computers of two school librarians are not connected to the Internet. Not only does this cause difficulties in the part of the library management that this topic deals with, but the librarians are also unable to receive e-mails from their county teacher association chairperson, their designated advisors from the Education and Teacher Training Agency, and their county Central Department. They are also unable to follow the news from their profession through web pages of librarian institutions from their country and abroad, or receive information about the conferences being held.

When asked whether their school library uses library automation software, four librarians answered negative. The remaining 23 libraries use Metelwin library automation software, which implies that that this software package is present in all elementary school libraries with automated management in the Krapina-Zagorje County. Only 9 librarians (39 %) that use the Metelwin software stated that they had received some form of training for it, while others learned to operate it through some other informal means of education.

Of all the libraries that own the Metelwin software, only 8 (35 %) have input their entire library collection into the software, which is a prerequisite for completely automated library management. This percentage for other libraries varies from 62 to 95 percent.

School librarians were also asked particular questions about the assigned reading collections. Libraries use different classification systems to organize the books on the shelves, and Croatian school libraries use classification by age [4]:

1. books for the youngest students, such as picture books, short stories, and simple poems (1st and 2nd grade of elementary school, class mark M)
2. children's literature, i.e. fairy tales, fantasy tales, children's poetry, children's tales, and novels (from 3rd to 6th grade, class mark D)
3. youth literature, i.e. works with historical, social and war topics (7th and 8th grade, class mark O)

4. folk literature, i.e. folk tales, fairy tales, riddles, stories, poems (class mark N)

5. dramas used for making school plays (class mark I)

It is important to note that 21 out of 23 librarians (91 %) have input all the mentioned assigned reading materials into the Metelwin software, keeping in mind that students are the most common users of the library collection, particularly because of assigned and optional reading materials. The remaining two libraries have so far entered 68 and 90 percent of the assigned reading collection.

The fact that only 13 out of 23 (57 %) libraries with Metelwin software have input the entire teacher collections highlights the importance of assigned reading collection input for automation of library management. The percentage of teacher collection input into the computer for the remaining ten school libraries is relatively low – as low as five percent for some libraries. Seven school librarians have input 50 percent or less of teacher collection into Metelwin, and only three have input more than 60 percent.

The situation is somewhat better for reference collections (encyclopaedias, lexicons, dictionaries, atlases etc). This collection type is completely input in 14 out of 23 libraries (61 percent). Two libraries have input 50 percent of the reference collection, and four have input 80 percent and more. This is because reference collection is used by students almost every day for writing essays or doing homework, and is used in several grades for performing the programme of Library and Information Education.

It is clear that the majority of school librarians have opted for a systematic order of inputting the collection into computer – assigned reading, followed by reference collection and then teacher collection.

The AV collection (records, audio and video cassettes), which is not very common in school libraries, has been input into Metelwin in only 7 out of 23 school libraries with the Metelwin software. The same applies for electronic resources, such as CD's and DVD's, which are becoming more and more present in school libraries.

Interestingly, the survey shows that eleven libraries no longer use card catalogues system, one library still uses it, while the remaining libraries provided no answer.

Everything that was stated so far is, in fact, a prerequisite for automation of library management. For every librarian this term implies usage of computers for borrowing library materials.

School librarians with entire collection input into the computer were asked whether their libraries have finalized their classic inventory books. It is interesting to note that only three out of eight (38 percent) libraries have finalized and archived them. In the remaining libraries, the librarians still enter new titles both into the computer and the inventory book, although the borrowing of books and other resources is done exclusively through computers in both cases.

What about the 15 libraries that have not input their entire collection into Metelwin? In eight of them, new titles are input only into the computer, whereas the rest use both the computer and the inventory books. Nine libraries manage borrowing books and other resources through computers, five use both computers and the card system, while one library uses only the card system.

In the four libraries that are managed without a computer programme the items are, naturally, borrowed through the card system.

C. School libraries networking

Although Metelwin enabled online browsing of databases way back in April 1988 (the first online database being that of City library and reading room "Metel Ožegović" in Varaždin) [5], still all elementary school libraries in the Krapina Zagorje County do not use this privilege.

Only 14 out of 27 libraries (23 of which use Metelwin and are connected to the Internet) have their online library catalogues, which means that 39 % libraries still do not use the data transfer module (despite having all the prerequisites).

Five librarians transfer data to the Internet every day after their working hours, one after acquiring new library items, four librarians do so once a year, and the rest a couple of times a year.

Eight school libraries have a hyperlink to their online library catalogue in the "Library" section of their school's website, while three libraries have the hyperlink to their online library catalogues on the home page of their schools' respective websites. The remaining schools do not have a hyperlink to their online catalogue on their school's website because, as it was mentioned, some libraries do not even have their own library catalogues. However, it is worth noting that all online library catalogues can be accessed through Metelwin search engine [1] (Metelwin\Pretraživanje na internetu\Metelwin OPAC) filtering and choosing the wanted library. After all, each school's online library catalogue is a standard online public access catalog (OPAC) and is open to everyone, as there are no password restrictions [6].

D. Covering the curriculum unit Online Library Catalogues in practice

When it comes to teaching segments of information literacy that are a part of Library and Information Education and which require using a computer, thirteen school librarians teach them mostly in the IT classrooms, seven librarians use school library computers, four use both IT classrooms and library computers, and three librarians do not cover these units.

The curriculum unit Online Library Catalogues is covered by two librarians in the seventh grade, by eight librarians in the eighth grade, by ten librarians in both seventh and eighth grade (in seventh grade by using online library catalogues of their own school library, and in eighth grade by using online catalogues of other

[1] http://library.foi.hr/digi/metelwin/

libraries, such as those of National and University Library in Zagreb or the Zagreb City Libraries). Seven elementary school librarians do not cover that unit at all (Table I).

The librarians who cover the curriculum unit Online Library Catalogues emphasize searching by author or title (10), by all criteria equally (7), and by author (3).

Seven school librarians keep regular track of their online library catalogues' statistics for number of visitors, and six of them noted a significant increase in number of visitors shortly after covering the curriculum unit Online Library Catalogues.

The librarians who work in school libraries with online databases introduced their colleagues to them and explained the basic rules of searching. None of the school librarians have had a user query for reserving a library item through Metelwin so far.

The questionnaire also asked whether the school librarian has ever used the online library catalogue to answer a user query. Only six librarians gave a positive answer to this question. The following reasons were elicited by the librarians:

- I can quickly check the number of copies;
- I can do an additional online check if a certain book is not on the shelf;
- Even when I am not in my library I can answer a user query whether a certain book is in my library's collection, which books from certain authors I have, and when the book is to be returned if it is out loaned; and
- I can check whether the public city library has the book the user is looking for.

E. Solving information problems through computers – step-by-step through online library catalogues

The educational goals in the school library are, to name a few, developing information and research skills and teaching students to use the sources of information in the school library.

Online catalogues for school libraries and other library types are an excellent example on which a student can gradually and consciously use a computer in class to acquire information literacy skills, in stages as described by Mike Eisenberg and Bob Berkowitz. This is the so-called Big6 model [7], one of the most commonly mentioned models and one of the most renown and widespread standards for teaching information literacy.

Six skills for teaching information literacy are [8]:

1. Task Definition (identifying the information needed for solving the problem)
2. Information Seeking Strategies (identifying all possible sources and selecting the best ones)
3. Location and Access (locating sources intellectually and phisically and finding information within selected sources)
4. Use of Information (reading, listening, viewing in order to get the most relevant information from the source)
5. Synthesis (organizing and presenting the information)
6. Evaluation (judging the whole process of solving the information problem and judging the product)

What does this actually mean? Suppose a student needs to get the book “Kći Lotršćaka” by Marija Jurić Zagorka, which means he or she is aware of his or her information needs (1st phase).

This is followed by a calculated planning of information searching strategies. The students should be taught not to avoid the standard sources, although they will almost certainly not be their first choice. This is mainly because card catalogues are slowly but steadily becoming extinct in the school libraries. After all, only one librarian reported that card catalogues are used in his library. The students will identify that source, know that it exists, regardless of whether it is physically present in their library or not, but their first choice for solving their information problem will be the computer and online library catalogue.

The third phase is finding the online library catalogue. The school website administrator should place a hyperlink to the online library catalogue in the "Library" section, or even on the school's home page. However, it would be even better for a school website administrator to authorize the school librarian to independently maintain the "Library" section so that it does not just provide the hyperlink to the online library catalogue but also publishes other information regarding the school library, from newly acquired titles, announcements for various workshops and activities to reports and similar announcements.

TABLE I COMPARING THE SCHOOL LIBRARIES WHICH DO NOT COVER ONLINE LIBRARY CATALOGUES WITH HAVING THE METELWIN SOFTWARE, AN ONLINE LIBRARY CATALOGUE, AND COMPUTERS INTENDED FOR STUDENTS' WORK

Libraries which do not cover the curriculum unit Online Library Catalogues	Metelwin is installed and computer is connected to the Internet	Online school library catalogue	The library possesses computers intended for students’ work connected to the Internet
1	+	-	-
2	-	-	-
3	-	-	+
4	+	-	-
5	+	-	-
6	+	-	-
7	+	-	-

If a library has automated management, i.e. it uses library automation software and is connected to the Internet, which is the case for 23 elementary school libraries in the Krapina-Zagorje County, it is only a small step from making the whole locally registered database visible online. If the technology permits it, why not use the option?

The first time the author's school library database was transferred online, it was performed by a Metelwin employee by connecting to the library's computer, and the whole procedure took less than an hour.

Locating the online library catalogue is followed by finding information in it. Since the students' information needs for online library catalogues usually include finding assigned reading titles, Metelwin is so well designed that the students can just input a part of the title and choose the desired book among the given matches. For this to work it is important that a librarian sets aside at least 30 seconds of his or her time after work, which is more than enough to update the search database.

Not only will this put the data regarding new books that were input into the local base online, but it will also show how many copies of a certain item are out of library that day, or how many copies are on the library's shelves, and the user can use the online library catalogue to see when a certain copy is due to be returned to the library.

Transferring the data is fairly simple, and searching it is possible immediately after the transfer. The advanced search initially [5] allowed search by using any of the eight bibliographic record elements (search by author, title, place of publication, publisher, date of publication, UDC, ISBN, and key words). Today, it is possible to search the online library catalogue through simple or advanced search query (by choosing even more criteria and their combinations).

When the student has identified the online library catalogue as a source and has accessed the information in it, the fourth stage of acquiring information literacy is finding relevant information. What does this exactly mean? For every match of bibliographic record areas, the students have to choose the one that fits their needs the most. For instance, if a student is searching for an encyclopaedia that deals with rapidly developing professions or sciences, he or she will choose a newer edition of a certain title, i.e. a book with a more recent date of publication.

This is followed by allocation of the book on the shelf (with the aid of the call number found in the online library catalogue).

V. CONCLUSIONS

When covering unit Online Library Catalogues, a computer in the school library will allow the students to single-handedly solve their information problems, reach all the stages in achieving learning objectives, or interiorization of new ideas from the cognitive domain. This involves six levels of skills: knowledge, comprehension, application, analysis, synthesis, and evaluation as the highest form [9].

A school librarian should show the students the hyperlink to the online library catalogue, encourage them to make it the starting point of their research and instruct them to use it, as they will benefit from this knowledge not only during their elementary school education, but also throughout their high school education, as high school libraries have their own online catalogues (also mostly in Metelwin), so they will be familiar with the interface and the way it functions. It should also be emphasized that the knowledge and skills they acquire will be useful for searching online catalogues of other library types – public, academic and other libraries.

What librarians can use as an additional motivation is the fact that online library catalogues are not like classic Internet search engines in which one needs to input an exact and precise term in an electronic resource so as to avoid the Information roulette. When searching online library catalogue, the user will always hit the information jackpot, so to say, as it is enough to enter just one element or its part, such as its title or the author's first or last name when searching for items in library's possession. In addition, the students can use the knowledge and skills they acquire while covering the curriculum unit Online Library Catalogues, a part of the National curriculum for seventh and eighth grade, to find their way in the world of information throughout the rest of their education, outside of their school library, even outside the school itself.

When covering this unit, most librarians emphasize searching by title and author, but mention other search criteria as well. This highlights another feature of the modern library, and that is instructing students how to work individually, that is, to systematically and independently explore all the options offered by the online library catalogue, at home, or on the library computer. In addition to mentioned digital competence, this underlines the important role of school libraries in teaching students another competence from the National Curriculum important for lifelong learning, and that is "learning to learn" [1]. This implies teaching students to pursue and persist in learning, to organise their own learning, and to effectively manage time and information when working on their own.

The survey conducted among the elementary school librarians in the Krapina-Zagorje County shows that computerisation and automation of library management are crucial factors for complete and quality teaching of information literacy skills, although one should not ignore the proactive role of librarians and the co-operation from their school principals as they are the ones who decide on the acquisition of computers for students and the library automation software (as both come with a price).

Although this work used the curriculum unit Online Library Catalogues to describe in detail how a school library computer plays an important part in acquiring information literacy, it is a disastrous piece of information that in the 21st century as much as seven elementary school libraries (26 %) in the Krapina-Zagorje County do not even cover this unit. Research based on a sample of 82 % of elementary school libraries in the

county shows that owning a library automation software (independent variable) is not in direct correlation with covering the curriculum unit Online Library Catalogues (dependent variable), which means that a causative or correlative relation has not been found. However, the hypothesis was confirmed in other two aspects. There is a statistically significant correlation between having an online library catalogue in school and having a computer for student work and its connection to the Internet as the independent variables, with teaching information literacy through computer usage on the example of online library catalogues (dependent variables).

Regarding these two independent variables, establishing an online library catalogue is a process which depends entirely on the school librarian and his or her own initiative. If a school has the Metelwin library software and the librarian's computer is connected to the Internet, this would cost the school nothing and could be done through Metelwin's help-desk in less than an hour.

Insufficient number of library computers for student usage is the crucial problem. Most school use high-speed Internet (ADSL) with flat rate packages. Connecting the computer with the ADSL device through networking cables is a job that most school caretakers can do well enough on their own, and many schools use ADSL devices which allow wireless access with sufficient signal strength to cover the entire building (although employees and persons in charge in some schools do not know that).

Therefore, the only solution in turn would be to increase the number of library computers intended for students, which in turn requires money or "fighting" for computers from donations. But, if the society wants to educate a student informationally literate for the 21st century, this is the way it should be heading.

REFERENCES

[1] "Nacionalni okvirni kurikulum za predškolski odgoj i obrazovanje te opće obvezno i srednjoškolsko obrazovanje." Internet: http://public.mzos.hr/fgs.axd?id=18247, Oct, 2011 [Jan. 28, 2013].

[2] "Nastavni plan i program za osnovnu školu." Internet: public.mzos.hr/fgs.axd?id=14181, Aug, 2006 [Jan. 28, 2013].

[3] A. Tkalac Verčič, D. Sinčić Ćorić, N. Pološki Vokić, Priručnik za metodologiju istraživačkog rada u društvenim istraživanjima. Zagreb: M.E.P., 2011.

[4] J. Panjičanin, "Organizacija i poslovanje školske knjižnice (signaturne oznake)." Internet: http://www.knjiznicari.hr/UDK02/index.php/Datoteka:Organizacija_i_poslovanje_%C5%A1kolske_knji%C5%BEnice_(signaturne_oznake)_-_Julija_Panji%C4%8Danin.ppt, May. 8, 2012 [Jan. 28, 2013].

[5] M. Šoštarić, D. Tkalec, "Razvitak i primjena računalnog knjižničarskog paketa Metel Win," in Godišnjak Gradske knjižnice i čitaonice "Metel Ožegović" Varaždin, vol. II. Varaždin: Gradska knjižnica i čitaonica "Metel Ožegović", 1998, pp. 165-191.

[6] K. W. Craver, Teaching Electronic Literacy. London: Greenwood Press, 1997.

[7] D. Kovačević, J. Lasić-Lazić, J. Lovrinčević, Školska knjižnica – korak dalje. Zagreb: Filozofski fakultet, Zavod za informacijske studije Odsjeka za informacijske znanosti : Altagama, 2004.

[8] I. Stričević, "Informacijska pismenost i uloga dječjih knjižničara," presented at Stručni skup Djeca na internetu u knjižnici, Zagreb, 2011.

[9] M. Matijević, D. Radovanović, Nastava usmjerena na učenika. Zagreb: Školske novine, 2011.

Design and Development of School Assets Management System

Andreas Vladimirou[1] and Angeliki Kokkinaki[2]
[1]Information and Communication Systems Graduate Program, Open University of Cyprus, Nicosia, Cyprus
[2]Department of Management and MIS, School of Business, University of Nicosia, Nicosia, Cyprus
[1]andreas.vladimirou@st.ouc.ac.cy
[2]kokkinaki.a@unic.ac.cy

Abstract - This paper presents the design and development of a web enabled system supporting materials flows and relevant management processes in educational settings. Unlike other organizational settings, where warehouses are organized with specific input and output procedures, in educational environments (at least, in Cyprus) resources are allocated in various locations (offices, warehouses, laboratories, gym rooms); they are not uniquely defined, and most of the time, the monitoring of their usage or consumption are ad hoc. These conditions create challenges and inefficiencies. The proposed system supports standardized handling of the resources in educational settings and structured management of assets and consumables. It consists of an application running on handheld/smartphone that supports record keeping of school materials. Moreover, receipt of goods from suppliers, loan/return of items and consumption/disposal of materials can be recorded. They system also has a server based back-end module and a central database repository. Pilot testing of the developed prototype took place in five primary schools in Cyprus and assessments have led to positive comments and recommendations for further improvements. The evaluation of the system was completed by the Web enabled tools checking for aspects of usability, efficiency and accessibility resulting to above the average rankings.

I. INTRODUCTION

In educational environments there is a number of materials which are allocated in various locations including but not limited to offices, warehouses, laboratories, gym rooms etc. These are not uniquely defined and there are inappropriate ways of monitoring the usage/consumption. These conditions create an environment, where optimal management procedures cannot be supported.

We noticed the assets management problem in schools and we considered that is an interesting variant of the general case problem of assets management. We identified extensions of application as the educational environments are numerous and multidimensional as covers different levels of education.

Our aim was to design and develop a complete information system which will save valuable time for people involved as it will help for the easier, faster and effective inventory and management of assets. The remaining of this paper is structured as follows. Section II includes implementation issues and Section II concludes this paper.

II. SYSTEM DESIGN AND DEVELOPMENT

In this section, we present aspects of the analysis, design and implementation of the system.

A. *Current Situation in schools*

During the users' requirements solicitation phase, we collected and examined information about the procedures followed for stock taking and stock management in primary schools in Cyprus.

Upon communication with involved stakeholders, suggestions and recommendations on the desired/optimal functional requirements have been collected and analyzed. Based on those, we identified the technical requirements that would lead the system to operate in order to make the procedures easier, faster and for the proper stock management.

The problems in conventional school asset management system are as follows:

- Assets and their quantities are recorded on paper which can be accidentally destroyed, or lost.
- In case of changes in quantities, the person in charge has to search for the specific paper which records the item is and change the quantity.
- Lack or limited capability of reports.
- The physical location of items is not recorded.
- Stock taking is time consuming and follows ad hoc procedures.

B. *Software Analysis and Design*

For analysis and design of the system we used Unified Modeling Language (UML) and we followed the ICONIX methodology.

UML is a standardized general-purpose modeling language in the field of object-oriented software

engineering. The Unified Modeling Language includes a set of graphic notation techniques to create visual models of object-oriented software-intensive systems.

ICONIX is a software development methodology which allows systematic transition from the initial requirements as set by the client to the code which finally implements the requirements. The ICONIX Process uses only four UML based diagrams in a four step process. These diagrams are: use case, robustness, sequence and class.

We used UML because is flexible (stereotypes and profiles can let you tailor UML to your needs) and it supports plenty of tools.

We used ICONIX because it's traceable from one step to the next, it actually drives the Object Oriented (OO) design from the use cases, it works well in an "agile" environment and it's a practical approach that's been proven to work in the real world, on hundreds of projects.

The proposed Web-based application features are as follows:

- Add/Edit/Delete: Users, items, suppliers, warehouses, categories;
- Receipt of items;
- Disposal/Consumption of items;
- Set minimum stock quantities;
- Loan/Returns of items.

Also, it could generate nine reports, which are:

- List of: items, stock takings, disposals /consumptions, receipts of items;
- Items by supplier;
- Items with less than a specific quantity;
- Stock Movement;
- Loaned items;
- Loans History.

The mobile application features are:

- Stock Taking;
- Receipt of items;
- Disposal/Consumption of items;
- Add new item;
- Add person to loan items;
- Loan/Return of items.

C. System Implementation Technologies

The android operating system (OS) selected for the development of the mobile application because its widespread adoption.

The integrated development environment (IDE) used is Basic4Android. The main advantages using Basic4Android are:

- It uses object oriented programming language (similar to Visual Basic);
- It supports web services, xml, json and csv;
- It consist of a powerful designer and step-by-step debugger;
- Supports all android versions from 1.6 to 4.x.

The programming language used for the web application is PHP. It is simple, open source, the code is inserted directly into the HyperText Markup Language (HTML) and offers many levels of security to prevent malicious attacks. Also, PHP doesn't use a lot of the system's resources so it runs fast.

MySQL is the open source relational database management system (RDBMS) used. It is fast, can provide access to multiple users and it counts more than 11 million installations.

D. Pilot Testing and Evaluation

Pilot testing of the developed prototype took place in five primary schools in Cyprus and assessments have led to positive comments and recommendations for further improvements. The evaluation of the system was completed by the Web enabled tools checking aspects of usability, efficiency and accessibility resulting to above the average rankings. Some of the tools that used for the web application evaluation are: AgileLoad, loadster, webFT, SecureAssist, total validator, phpUnit.

Some of the recommendations of the end users which will improve the system are:

- Location/Warehouse Transfers;
- Orders to suppliers;
- Barcode usage to scan items from mobile device instead of searching them;
- Pictures of items to be included along with description.

III. CONCLUSIONS

In this paper, we presented aspects of the design, development and evaluation of the School Assets Management System. The system developed was based on the needs of primary schools in Cyprus with the aim to manage their assets easily, effectively and efficiently. After testing the system, and implementing changes that have been suggested the end users recommendations, the

system is ready to undergo a pilot operation phase and beta testing.

It is expected that we'll receive valuable feedback regarding the ease of use of the system, the required time for the processes supported and the overall management of the stock. It is expected to help develop the system in a fully operational version in the near future.

APPENDIX

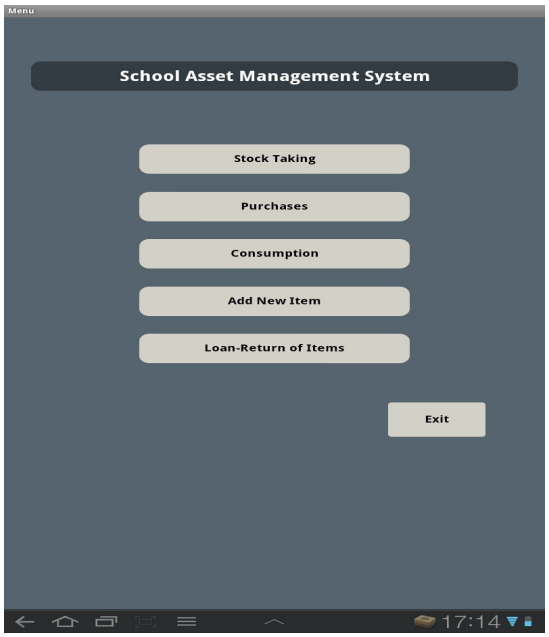


Figure 1. Mobile application screenshot 1

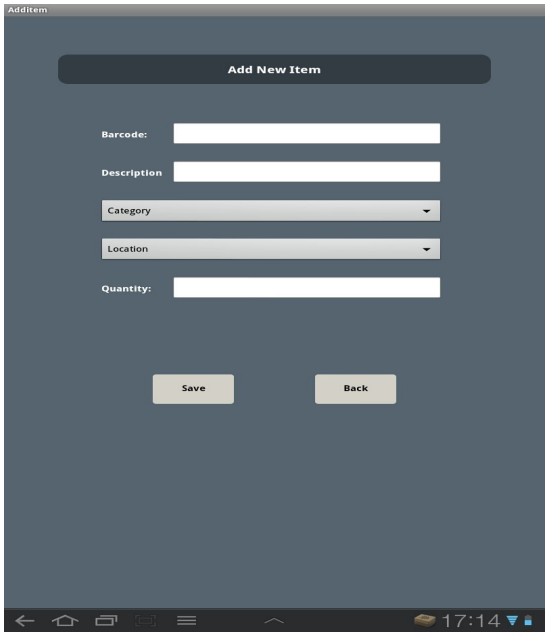


Figure 2. Mobile application screenshot 2

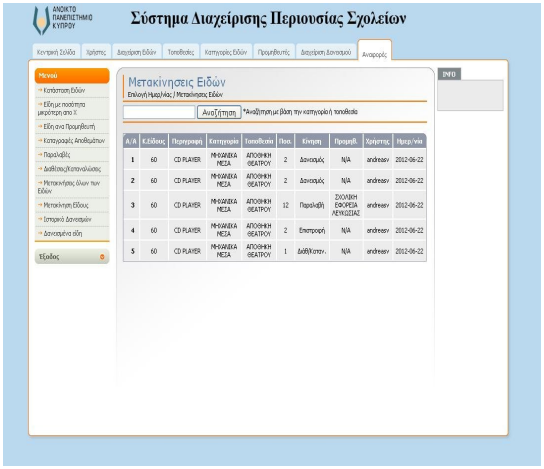


Figure 3. Web application screenshot 1

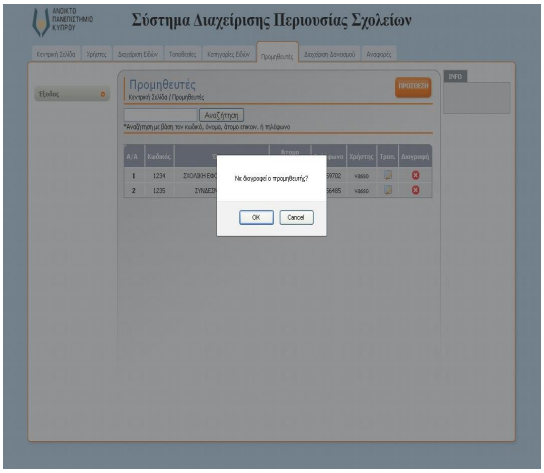


Figure 4. Web application screenshot 2

REFERENCES

[1] David Frederick Ross "Introduction To Supply Chain Management Technologies (second edition)". International Standard Book number -13 : 978-1-4398-3753-5. 2011 (Ebook - PDF).

[2] Barzan Tony Antal – "IT Inventory And Resource Management With OCS Inventory NG 1.02. Eliminate inventorying dilemmas by implementing a free & feasible IT Inventory solution.". May 2010.

[3] Manjunath Ramachandra – "Web- Based Supply Chain Management And Digital Signal Processing " New York 2010.

[4] J.A. Myburg – "Physical Asset Management Systems Concepts".

[5] Andrew S. Targowski & J.Michael Tarn - " Enterprise Systems Education in the 21st century " USA 2007.

[6] Alex Sefke - " IT Asset Management - "How to manage your information technology equipment " Great Britain 2003.

[7] G. Knolmayer, P. Mertens, A.Zeier, JT.Dickerbach - " Supply Chain Management Based On SAP Systems, Architecture And Planning processes"

[8] Dave Swartz, Ken Orgill - " Higher Education ERP : Lessons Learned" 2001.

[9] Max Muller - " Essenials Of Inventory Management" 2003.
[10] Microsoft Dynamic - " Inventory Management" February 2009.
[11] http://en.wikipedia.org
[12] www.sysoptima.com/erp/history_of_erp.php
[13] www.ebzberatungszentrum.de/pps_seiten/sonstiges/erp_engl.htm
[14] http://whatiserp.net/erp-basic-knowledge/erp-system/
[15] www935.ibm.com/services/uk/bcs/images/bcs_supply_chain443.jpg
[16] http://wiki.sdn.sap.com/wiki/display/ERPLO/ERP+EAM
[17] http://www.ema-inc.com/eam.html
[18] http://acqual.com/images/IT%20Asset.jpg
[19] http://cmckinnon.blogs.nuxeo.com/wpcontent/uploads/cmckinnon/images/6a010536291c30970b0120a6ead2e5970b-pi.png
[20] http://www.sama.org.za/media//DIR_36301/PhysicalAssetManagement_JMyburg.pdf
[21] http://www.samanage.com/solutions/it-asset-management.html
[22] http://timeislearning.com/wp-content/uploads/2011/11/inventory-management.jpg
[23] http://en.wikipedia.org/wiki/Enterprise_asset_management
[24] http://en.wikipedia.org/wiki/Supply_chain_management
[25] http://www.manageengine.com/products/service-desk/asset-inventory-management.html
[26] http://en.wikipedia.org/wiki/Warehouse_management_system
[27] http://it.toolbox.com/wiki/index.php/IT_Inventory_Management_with_SAManage
[28] http://www.inventoryops.com/articles.htm
[29] http://www.barcodesinc.com/articles/what-is-inventory-mangement.htm
[30] http://archive.sba.gov/idc/groups/public/documents/sba_homepage/pub_mp22.pdf
[31] http://www.youtube.com/watch?v=qkZQxXJuqKo
[32] http://en.wikipedia.org/wiki/Warehouse_management_system
[33] http://www.inventoryops.com/warehouse_management_systems.htm
[34] http://it.toolbox.com/blogs/wms-essentials/what-is-warehouse-management-system-wms-18819
[35] http://www.eweek.com/c/a/Mobile-and-Wireless/Android-Beats-iOS-in-Broad-Mobile-Market-10-Reasons-Why-104185/
[36] http://www.lifehack.org/articles/technology/five-reasons-to-choose-an-android-tablet-over-an-ipad.html
[37] http://www.phonearena.com/news/Which-platform-crashes-more-Android-or-iOS_id26542
[38] http://www.digitizd.com/2010/11/10/10-reasons-i-chose-android-over-an-iphone/
[39] http://www.comscore.com/Press_Events/Press_Releases/2011/12/comScore_Reports_November_2011_U.S._Mobile_Subscriber_Market_Share
[40] http://en.wikipedia.org/wiki/Android_(operating_system)
[41] http://www.bgr.com/2011/05/19/android-grabs-53-of-global-smartphone-market-share-ios-50-of-application-revenues/
[42] http://www.businessinsider.com/smartphone-survey-results-2011-4?op=1
[43] http://en.wikipedia.org/wiki/Mobile_operating_system
[44] http://halotech.org/showthread.php?405-10-Ways-android-Beats-other-OS-Android-vs-iPhone-and-Windows-Phone.
[45] http://www.networkworld.com/news/2011/080111-canalys.html
[46] http://el.wikipedia.org/wiki/PHP
[47] http://en.wikipedia.org/wiki/PHP
[48] http://el.wikipedia.org/wiki/MySQL
[49] http://en.wikipedia.org/wiki/MySQL
[50] http://www.davidtan.org/
[51] http://www.upriss.org.uk/perlcgi/l2.pdf
[52] http://w3techs.com/technologies/overview/programming_language/all
[53] http://langpop.com/
[54] http://www.infoq.com/news/2009/03/top-scripting-languages-php-ruby
[55] http://www.kepa.gov.cy/diktiothite/Portal/PortalDocuments.aspx?DocumentId=a735c138-a74e-483b-8720-31bc9aca7169
[56] https://elearning.cs.unipi.gr/?id=126
[57] http://dmst.aueb.gr/dds/ism/oo/indexw.htm
[58] http://www.dmst.aueb.gr/louridas/lectures/dais/uml/uml.html
[59] http://en.wikipedia.org/wiki/ICONIX
[60] http://en.wikipedia.org/wiki/Agile_software_development
[61] http://www.scribd.com/doc/85521860/7/ICONIX

Data mining in hybrid learning: possibility to predict the final exam result

Jasna Gamulin[a], Ozren Gamulin[b], Dragutin Kermek[c]
[a] University of Zagreb School of Medicine
Šalata 3, Zagreb, Croatia, Phone:+385 1 4590 268, E-mail: jasna.gamulin@gmail.com
[b] University of Zagreb School of Medicine, Department of Physics and Biophysics
Šalata 3, Zagreb, Croatia, Phone:+385 1 4566 797, E-mail: ozren@mef.hr
[c]University of Zagreb Faculty of Organization and Informatics, Department of Computing and Technology, Department of Theoretical and Applied Foundations of Information Sciences, Pavlinska 1, Varaždin, Croatia, Phone: +385 42 390864, E-mail: dkermek@foi.hr

Abstract - The hybrid learning environment that uses traditional lectures and examinations in conjunction with online learning resources and online assessment tools provides numerous data on students' activities and assessment scores which could be used for constructing a final exam result prediction model. In this paper the data on activities and assessments supported by information and communication technology (ICT) of 302 students enrolled in the first year Physics course of a biomedical university study program have been used to establish the correlations between scores on written midterm exams, scores on web-based formative assessment during seminar teaching, scores on web-based formative assessment during laboratory teaching, scores and time used for online self-assessment test, number of Moodle logins, number of approaches to specific Moodle resources and final exam result. As prediction methods the Principal Component Regression (PCR) and Partial Least Square regression (PLS) have been used, especially due to assumed multi-colinearity of predictive variables and dimension reduction requirement. The model could be useful for students and for teachers who would have the possibility to react and remedy the predicted final exam result if necessary.

I. Introduction

Data mining in education is a new and still developing interdisciplinary research area. Educational data mining (EDM) pursues the development of methods for exploring data acquired from educational context [1]. The hybrid learning environment which is the form of learning environments where e-learning is mixed with traditional teaching and/or human tutoring and where face-to-face communication with students is mixed with computer-mediated interaction [2] provides numerous data on students' activities and assessment scores which could be used for constructing a final exam result prediction model.

In this paper the data on activities and assessments supported by Moodle [3] Learning Management System and a previously developed web-based formative assessment tool named Pitalica of 302 students enrolled in the first year Physics course of a biomedical university study program have been used to establish the correlations between the formative and summative assessment scores, the number of Moodle logins, the time used for self-assessment test and the final exam result. This paper aims to investigate whether students' activities during the course in a hybrid learning environment can be used to predict students' performance at final exam and what would be the appropriate methodology. The idea for the paper appeared when the data on frequency of Moodle logins were observed and it was perceived that the peek days on Moodle corresponded with the dates of the three midterm exams and the final exam date (Fig. 1). The same tendency has been observed during period 2006-2011 when learning material was placed at physics.mef.hr. (Inset in Fig. 1).

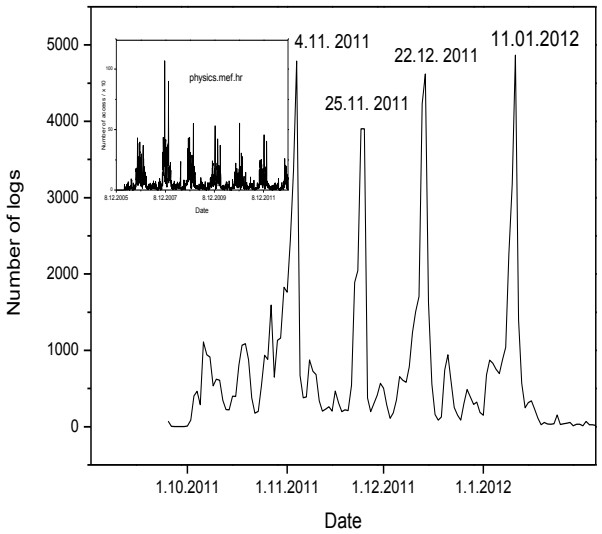


Figure 1. Number of Moodle and physics.mef.hr logins vs. midterm and final exam dates

This paper is organized as follows: first we have the review of the state-of-the-art scientific research on prediction of a student's performance using data mining. The underlying method of the study is presented followed by presentation, analysis and discussion of results. At the

end, main conclusions, limitations of the study and ideas for future research are presented.

II. Literature review

The idea, originating from data mining (DM) in e-commerce, to assess clients' interest in order to increase sales, is mapped also to educational context. The EDM is still lagging in progress compared to commercial DM, nevertheless there is a growing interest in applying DM to the educational environment [1]. There are some important issues that differentiate EDM from DM in other domains [4]. For example, objectives of EDM, being the improvement of the learning process, guiding students' learning or understanding educational phenomena, are difficult to quantify and measure [4]. In EDM there are also different stakeholders in respect to DM in other domains: students, teachers, course developers, universities, school/network/system administrators [1]. According to Romero et al. [5], these are the usual steps of EDM: data collection, data preprocessing, applying DM and interpreting, evaluating and deploying the results.

Data mining is used in distance learning environment [6], [7], [8], [9] and in hybrid learning environment [10], [11]. Prediction of students' performance is one of the most popular EDM applications. Jovanović et al. [6] applied classification models for prediction of students' performance. Model building process included data preprocessing, parameter optimization and attribute selection steps, which enhanced the overall performance. Additionally, they propose a Moodle module that allows automatic extraction of data. Dráždilová et al. [12] present a case study that aims to find latent social networks from a number of data collections stored in Moodle system logs and based on students' similar behavior. Nasiri et al. [13] are searching for ways of usage of EDM to estimate probability of dismissal, drop out and other challenges within the process of the graduation and also capable tools to predict GPA or even semester by semester grades. Cross-Industry Standard Process for Data Mining (CRISP) has been used as methodology of data mining. Weng [14] explored whether from infrequent data, information could be found that would be useful for teachers to find out which students need extra help in learning. Amoroso [10] investigated the use of online assessment tools for large classes. Correlation analysis and linear regression was used to ascertain relationships as well as impacts on overall grade. It was displayed that online assessments and traditional examinations both contribute to final grades for students in large classes. Attendance was not found to have either a strong correlation with final grade, nor be significant in predicting final grade. Students who performed better on online assessments also performed better on traditional examinations [10]. DeNeui et al. [11] examine the relationship between the frequency of usage of various utilities posted on Blackboard and student performance in a hybrid class. Results revealed a significant positive partial correlation between overall usage and their exam scores. Delgado et al. [15] applied artificial neural networks to obtain new knowledge from the Moodle logs. The information is combined with information on the students' final evaluation in order to obtain models that can predict students able to pass the exam. The proposed model shows that it is possible to predict those students with problems to pass a course with 80% of success. Kotsiantis et al. [8] point out that incremental learning ability has become more and more important for machine learning approaches in terms of explosive increase of data and information. They are proposing the combining classifiers as a new direction for the improvement of the classification accuracy (algorithms which try to forget irrelevant information instead of synthesizing all available information with the classic batch learning algorithms). Lykourentzou et al. [7] deployed a dropout prediction method for e-learning courses, based on three popular machine learning techniques and detailed student data. Myller et al. [9] are using linear regression finding method to predict exam results and to form heterogeneous student groups. Minaei-Bidgoli et al. [16] present an approach to classifying students in order to predict their final grade based on features extracted from logged data in an education web-based system. They designed, implemented and evaluated a series of pattern classifiers and compared their performance on an online course dataset. The genetic algorithm (GA) is demonstrated to successfully improve the accuracy of combined classifier performance about 10 to 12% when comparing to non-GA classifier.

III. Methodology

A. Methods

As prediction methods the Principal Component Regression (PCR) and Partial Least Square regression (PLS) have been used, especially due to assumed multi-colinearity of predictive variables and dimension reduction requirement**.** Software used here is Matlab with the addition of PLS Toolbox. In data mining and predictive modeling we usually have a large number of possible predictive variables. There from comes the demand for variable selection and dimension reduction in multivariate regressions. Another limitation is when several of the predictive variables are highly correlated. We have the multi-colinearity problem when a high degree of correlation among the predictive variables increases the variance in estimation of the regression parameters [17]. PLS technique is widely used in chemical engineering where the relationships between many predictor variables are not well understood [17]. According to Maitra and Yan [17] PLS technique is more efficient than PCA technique for dimension reduction due to the supervised nature of its algorithm. The similar finding was presented previously also by Yeniay et al. [18] who demonstrated that partial least squares regression yields somewhat better results in terms of the predictive ability of models obtained when compared to the other prediction methods. Haenlein and Kaplan [19] indicate that the PLS technique is especially appropriate

in cases in which constructs are measured primarily by formative indicators.

B. Dataset used

Data on assessment scores, time needed and Moodle logs during the course of 302 students enrolled in the first year Physics course of a biomedical university study program in the year 2011/2012 have been used in this paper. The gender structure of enrolled students is displayed in Table I.

TABLE I. GENDER STRUCTURE OF ENROLLED STUDENTS

	Enrolled (N)	Man	Woman
2011	302	121	181
%	100,00	40,07	59,93

Although authors usually take students' key demographic characteristics and their marks on a few assessments during the course as training set for predicting students' performance [20] in this paper we omitted the demographic characteristics and did not take them as predictor variables since in the specific situation the students' group is very homogeneous. Students enrolled in the observed study program had, with almost no exceptions, excellent marks during secondary school, are of the same age and have good ICT literacy.

TABLE II. DESCRIPTION OF THE VARIABLES

Indep. var.	Description
x_1	total scores of midterm exams (written).
x_2	scores of web-based formative assessment during seminars (Pitalica – seminars)
x_3	scores of web-based formative assessment during seminars and laboratory teaching (Pitalica – total)
x_4	scores of web-based formative assessment during laboratory teaching (Pitalica – laboratory)
x_5	number of logins to Moodle (learning material + general information about the course)
x_6	ability [23] calculated according scores of midterm exams
x_7	number of logins to Moodle learning material
x_8	number of login to Moodle general information about course
x_9	ability [23] calculated according to scores of self-assessment test on Moodle
x_{10}	time used for self-assessment test on Moodle
Dep. var.	
y_1	final oral exam grade at the first term calculated as the average of four questions' grades (Final oral exam consists of four questions and each question is graded. Final oral exam grade is calculated as the average of those four grades. Only first term is considered.)
y_2	final oral exam grade at first three terms calculated as the average of all three terms (is taken as variable in order to register also failures and to get a continuous variable)
y_3	final oral exam grade at all terms calculated as the average of all five terms (is taken as variable in order to register also failures and to get a continuous variable)
y_4	final oral exam grade after five exam terms (only the biggest grade achieved is considered. The information of possible exam falls is not registered)

Grading system in Croatia consists of 5 grades (range 1-5; 5=excellent, 4=very good, 3=good, 2=satisfactory, 1=fail).

The web application *Pitalica* has been introduced primarily in order to motivate students to prepare better and to participate actively during seminars. The application allows teacher to assess students during seminars and through their answers to detect the level of their former knowledge of the subject and understanding of the seminar topic and adjust the course of lecture accordingly. The results of the experiment during Physics course in academic years 2008/2009 and 2009/2010 have been described by Gamulin, Gugić, Gamulin [21]. After showing positive teaching results and good acceptance by students and teachers, the application has been extended also to the laboratory teaching [22] in order to allow the teacher to quickly assess students at the beginning of each laboratory exercise. It is assumed that a student will spend approximately 5 minutes on a computer answering to 4-6 questions related to the exercise that he/she is about to work off.

A student enrolled in the course has access to resources posted on Moodle: self-assessment test which could be taken only once, web-laboratory, simulations, video laboratory (short videos describing specific exercises), interactive exercises and PowerPoint presentations.

The Physics course is held at the beginning of the first study year and it is the first bigger exam during the study. It is observed that students are inexperienced in university type of studying and that their ability to learn and understand physics is not the only element which can influence their final exam results. Until the end of an academic year students have at disposal 5 exam terms. The question arises: after how many terms the best image of students' ability to pass the exam would be captured. To overcome this problem 4 different dependent variables are generated and used for PLS and PCR modeling.

C. Data preprocessing

LMS systems produce statistic reports but they are not designed in order to help instructors in bringing conclusions for the course or student potential. These systems are built and useful for administrative purposes. The usage of those data requires long and wearisome preprocessing [5]. In order to prepare data for DM the database is generated with tables exported from Moodle system with data of logs number, log date, type of accessed material, results from self-assessment test. Tables with results accumulated by means of Pitalica during web-based formative assessment on seminars and laboratory teaching are also added to that database. From there, ten independent variables are selected (Table II) in separate database table which is used to export data to format which can be processed by Matlab. Final extracted data are divided into two sets. One set is used for model calibration and cross validation and the other for prediction assessment of generated PCR and PLS models.

IV. RESULTS

First we calculated the correlation coefficients between chosen independent variables and proposed dependant variables to check if there is any significant connection between them (Table III). From the table we can observe that all independent variables with exception of x_8 and x_{10} have statistically significant ($p<0.05$) correlation with proposed dependant variables y.

TABLE III. CORRELATION COEFFICIENTS

Independent	y_1	y_2	y_3	y_4	Sign.
Dependent					
x_1	0.60	0.62	0.61	0.56	p<0.05
x_2	0.29	0.32	0.31	0.30	p<0.05
x_3	0.47	0.54	0.53	0.51	p<0.05
x_4	0.48	0.57	0.57	0.54	p<0.05
x_5	0.21	0.24	0.25	0.26	p<0.05
x_6	0.33	0.38	0.36	0.36	p<0.05
x_7	0.21	0.23	0.24	0.24	p<0.05
x_8	0.12	0.12	0.12	0.11	
x_9	0.33	0.40	0.40	0.38	p<0.05
x_{10}	-0.01	-0.01	-0.01	-0.03	

Two parameter types are observed to evaluate goodness of the model. The first one is root-mean square error (RMSE) which gives the information about the fit of the model to the used data set. Three types of RMSE are calculated for each (PLS and PCR) model. There is root-mean square error of calibration (RMSEC), root-mean square error of cross-validation (RMSECV) which measures a model's ability to predict new samples and root-mean square error of prediction (RMSEP) which is calculated when the model is applied to new data set provided that the reference values for the new data are known. The second parameter is correlation coefficient R^2 between measured values [18]. Three R^2 values were observed, for calibration (R^2 C), for cross validation (R^2 CV) and for prediction (R^2 P). Behavior of those parameters in dependence on number of used latent variables (PLS) and principal components (PCR) for different dependant variables y are presented in Fig. 2 and 3 respectively.

From Fig. 2 (PLS) we can see that values of both parameters are stabilized after third latent variable. The highest value of R^2 was observed for y_2 variable while value of RMSE is lowest for y_4 for calibration and cross validation, but in case of prediction with new data set lowest RMSEP is observed for y_2 variable. Considering all parameters calculated with PLS and also considering that differences between RMSEC and RMSECV for y_2, y_3 and y_4 are small, we can conclude that best fitting was achieved when y_2 is used as dependant variable.

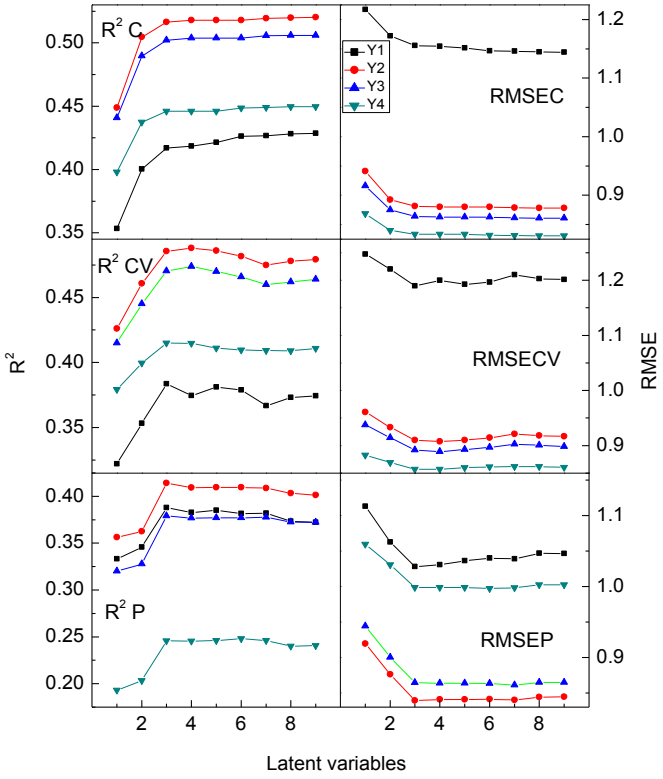


Figure 2. PLS technique - latent variables versus RMSE and R^2

In Fig. 3 changes of R^2 and RMSE for PCR model are presented. Variations of parameters are more emphasized in comparison to PLS model and stabilization of parameter value is achieved only after seventh principal component. The highest R^2 values and the lowest RMSE values for calibration and cross validation are reached for same dependent variables, y_2 and y_4 respectively. In contrast to PLS model, best fit for prediction with new set of data is achieved for y_1 dependant variable. In addition, it must be noted that value of RMSEP for all y variables fluctuates very strongly with increased number of included principal components.

Further, the percent of captured variance with number of components for both models and all y variables are compared. While total percentage of captured variance is similar for both applied models for x-block data the difference is observed for different y in y-block data. Captured variance for PLS model approaches its maximal value for 7 latent variables (Table IV) while captured variance for PCR model reaches its maximal value for 9 principal components (Table V). The latent variables could explain 52,02% of the variation, the PLS model reaching the maximal value with seven latent variables and the PCR model with nine principal components. Total captured variance is calculated for all four y variables and the result shows that total percentage is maximal for y_2. That result is in accordance with conclusions derived from Fig. 2 and Fig. 3.

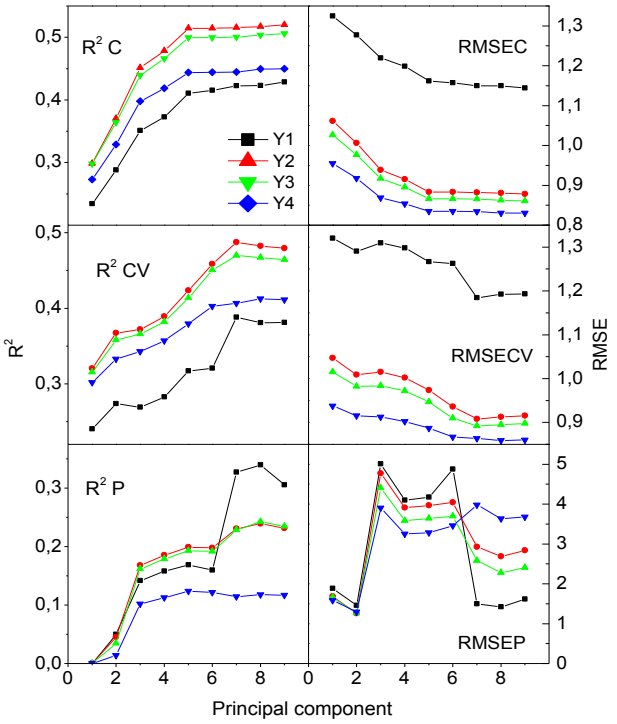


Figure 3. PCR technique – principal components versus RMSE and R^2

TABLE IV. PERCENT VARIANCE CAPTURED BY PLS MODEL

PLS	X-Block		y-Block	
Comp	This	Total	This	Total
1	31.58	31.58	45.14	45.14
2	18.52	50.1	5.36	50.5
3	11.74	61.83	1.21	51.71
4	15.3	77.13	0.15	51.85
5	6.29	83.42	0.09	51.94
6	4.33	87.74	0.06	52
7	3.09	90.83	0.02	52.02
8	4.07	94.9	0	52.02
9	5.1	100	0	52.02
10	0	100	0	52.02

According to the data listed in Table IV and Table V, the data and behavior of R^2 and RMSE presented in Fig. 2 and Fig. 3 we can conclude that for our data set the PLS model is more consistent and gives better prediction. The presented results also suggest that selection of correct set of depending variables is important step in building successful predictive model. In this paper and according to presented results, y_2 (final oral exam grade at first three terms calculated as the average of all three terms) is emphasized as most promising dependant variable.

TABLE V. PERCENT VARIANCE CAPTURED BY PCR MODEL

PCR	X-Block		y-Block	
Comp	This	Total	This	Total
1	33.79	33.79	29.85	29.85
2	20.68	54.47	7.17	37.02
3	16.4	70.86	8.14	45.16
4	8.6	79.46	2.69	47.85
5	6.58	86.04	3.6	51.45
6	5.3	91.34	0.01	51.46
7	4.15	95.49	0.1	51.56
8	3.4	98.89	0.18	51.74
9	1.11	100	0.29	52.02
10	0	100	0	52.02

V. CONCLUSION

Results presented in this paper indicate that in case of hybrid learning and mixed source of data (written midterm exams, Moodle and Pitalica) the prediction of final grade is possible but in case of models applied here the prediction is not very accurate. Observed parameters (R2, RMSE and number of components) indicate that PLS model is more consistent than PCR model, the conclusion also reached by [17].

Observed results also indicate that the choice of dependant variable can influence goodness of the model fit. In this paper the best choice proved to be y_2 variable (final oral exam grade at first three terms calculated as the average of all three terms). This choice of y variable is, according to the teacher's experience from many years period, the best reflection of students' habits in taking the exams.

Two things could be done to improve the prediction of students' final grade and results presented here. First, number of cases used in model calibration has to be increased. In order to do so and collect more data, the end of another academic year should be met. Second, other modelling and predictive methods should be used, for instance, neural networks and fuzzy logic in order to improve predictive results.

The predictor variable – time used for self-assessment test - was not found to have either a strong correlation with final grade, nor be significant in predicting final grade.

EDM could be used to predict students' performance at final exam and subsequently to improve the learning process since it gives the chance to students and to teachers to remedy the predicted result.

REFERENCES

[1] C. Romero and S. Ventura, "Educational data mining: a review of the state-of-the-art", IEEE Trans. Syst. Man Cybernet, C Appl. Rev., 40(6),pp. 601–618, 2010.

[2] G. Bubas, D. Kermek, "The Prospects for Blended Learning in Croatian Academic Institutions", *CARNet Users Conference CUC* 2004.

[3] Moodle, 2013. https://moodle.org/

[4] A. C. Romero and A. S. Ventura, "Educational data mining: A survey from 1995 to 2005", Journal of Expert Systems Applications, 33(1), pp. 135-146, 2007.

[5] C. Romero, S. Ventura and E. García, "Data mining in course management systems: moodle case study and tutorial", Comput. Educ. 51(1), pp. 368–384, 2008.

[6] M. Jovanović, M. Vukičević, M. Milovanović and M. Minović, "Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study", International Journal of Computational Intelligence Systems, 5:3, pp. 597-610, 2012.

[7] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques", Computers and Education 53 (3), pp. 950–965, 2009.

[8] S. Kotsiantis, K. Patriarcheas and M. Xenos, "A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education", Knowledge-Based Systems, 23(6), pp. 529-535, 2010.

[9] N. Myller, J. Suhonen and E. Sutinen, „Using Data Mining for Improving Web-Based Course Design", in Proc. International Conference on Computers in Education, (USA, Washington, 2002) pp. 959- 964, 2002.

[10] D. L. Amoroso, "Use of Online Assessment Tools to Enhance Student Performance in Large Classes", Information Systems Education Journal, 3 (4)., pp. 1-10, 2005.

[11] D. DeNeui, Daniel L.; T.L. Dodge, "Asynchronous Learning Networks and Student Outcomes: The Utility of Online Learning Components in Hybrid Courses", Journal of Instructional Psycholog, Vol. 33 Issue 4, pp. 256-259. December, 2006.

[12] P. Dráždilová, G. Obadi, K. Slaninová, S. Al-Dubaee, J. Martinovič and V. Snášel, "Computational Intelligence Methods for Data Analysis and Mining of eLearning Activities", F. Xhafa et al. (Eds.): Computational Intelligence for Tech. Enhanced Learning, SCI 273, pp. 195–224, 2010.

[13] M. Nasiri, B. Minaei and F.Vafaei, "Predicting GPA and academic dismissal in LMS using educational data mining: a case mining", 6th National and 3rd International conference of e-Learning and e-Teaching (ICELET 2012) pp. 53-58, 2012.

[14] C.H. Weng,"Mining fuzzy specific rare itemsets for education data", Knowledge-Based Systems, 24 (5), pp. 697-708, 2011.

[15] M. Delgado, E. Gibaja, M.C. Pegalajar and O. Pérez, "Predicting Students' Marks from Moodle Logs using Neural Network Models", in Proc. International Conference on Current Developments in Technology Assisted Education, (Sevilla, Spain, 2006), pp. 586-590. 2006.

[16] B. Minaei-Bidgoli, D. A. Kashy, G. Kortmeyer and W. F. Punch, "Predicting student performance: an application of data mining methods with an educational Web-based system", in Proc. 33rd International Conference on Frontiers in Education, (Colorado, Westminster, 2003), pp. 13-18, 2003.

[17] S. Maitra and J. Yan, „Principal Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression", Casualty Actuarial Society, 2008 Discussion Paper Program, Applying Multivariate Statistical Models, pp. 79-90, 2008.

[18] O. Yeniay and A. Göktaş, "A comparison of partial least squares regression with other prediction methods", Hacettepe Journal of Mathematics and Statistics, Volume 31 (2002), pp. 99-111, 2002.

[19] M. Haenlein, A.M. Kaplan, "A Beginner's Guide to Partial Least Squares Analysis", Understanding Statistics, *3*(4), pp. 283–297, 2004.

[20] S. Kotsiantis, C. Pierrakeas, P. Pintelas, "Predicting students' performance in distance learning using machine learning techniques", Applied Artificial Intelligence, 18: pp. 411-426, 2004.

[21] Gamulin, J., Gugić, J., Gamulin, O., "Improving classroom teaching in higher education environment using web-based formative assessment", MIPRO 2011-33rd International Convention on Information and Communication Technology, Electronics and Microelectronics - Proceedings, art. no. 5533585, p. 1001-1006, 2010.

[22] J. Gamulin, O. Gamulin, "Enhancing laboratory teaching in higher education environment using web-based formative colloquiums", MIPRO 2011-34th International Convention on Information and Communication Technology, Electronics and Microelectronics - Proceedings, art. no. 5967237, p. 1189-1194, 2011.

[23] F.B. Baker, The Basics of Item Response Theory. University of Maryland, College Park: ERIC Clearinghouse on Assessment and Evaluation, 2nd edition, 2001.

Students' attitudes to the use of Open Source technology in education

Krešimir Pavlina*, Borislav Petrović*
* Department of Information And Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
kpavlina@ffzg.hr

Abstract - The use of computers in modern life is almost unavoidable since computers are used in business and private life. Most users today use proprietary software, but with the arrival of the financial crisis forced savings raised the question of whether free Open Source technologies provide a good replacement for proprietary software. This paper presents the results of research about students' attitudes towards the use of Open Source technologies in education. Research was conducted by survey on primarily students of information and computer science since they are advanced users of information technologies they are competent to express their opinions whether Open Source technologies can provide good alternative to proprietary software and when Open Source technologies should be implemented in educational process.

I. INTRODUCTION

The use of specific and general software in education (from the operating system, office applications to web browsers) serves solving problems related to many areas, some of which are not included in the information technology education and other which are not part of the information technology field. In this way, one could say that information technology can affect most other scientific and practical problems, tasks and areas. Given this premise we can come to the conclusion that general computer literacy should not be reserved only for the chosen, but it should be encouraged, such as literacy in their mother tongue. It also means that the curriculum should expand content which will be included under term information and computer literacy. In this case, the opinion and educated decisions when choosing the software could come not only from IT professionals, but also by interested individuals from other professions. [1]

This paper is based on a survey of attitudes of students towards the use of Open Source technology in education, and includes primarily students of information and computer science studies and related fields.

Popularization of Open Source, which can be seen as an increased number of computers which are sold with preinstalled Open Source operating systems and the increasing occurrence of articles about Open Source software in professional computer magazines, indicates the trend of its increasing integration into everyday computing. This is the primary reason for the research about attitudes of students as future professionals on the market, on the application of Open Source technologies in education. [2]

II. METHODOLOGY

The survey included 197 respondents. According to the survey results youngest respondents were 18 years old, while the oldest were 32 years old. Most respondents were between 20 and 25 years of age (147 respondents). Majority of respondents were 24 years old (15.2%). Survey was conducted via SurveyGizmo on-line survey service.

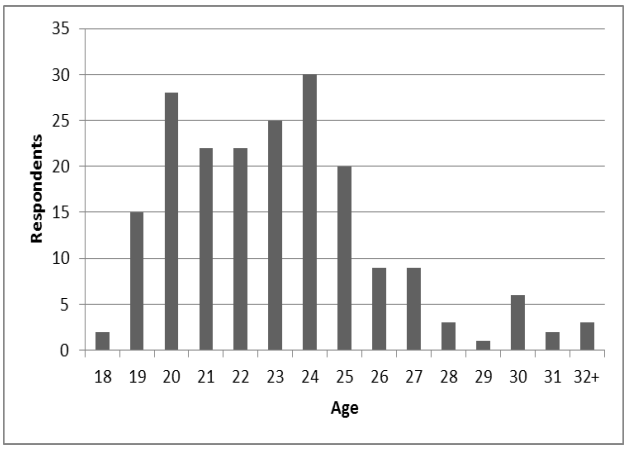


Figure 1. Respondents by age

By secondary education 52.3% of respondents finished gymnasium programs, 44.2% had attended vocational or technical secondary schools, while 3.6% received arts education.

Most of the respondents are currently enrolled at the undergraduate level (50.8%), then at the graduate level (43.7%), and finally the post-graduate level (5.6%).

Since Open Source technologies are still not universally recognized and accepted this study examined the attitudes of students of information and computer science since they represent advanced IT users who are more familiar with the advantages and disadvantages of Open Source technologies. The survey respondents were students of several higher education institutions that educate future information professionals.

TABLE I. RESPONDENTS AFFILIATION TO HIGHER EDUCATION INSTITUTION

Institution	Respondents	%
Faculty of Humanities and Social Sciences, University of Zagreb	61	31,1%
Polytechnic of Zagreb	57	28,9%
Department of Informatics, University of Rijeka	19	9,7%
Faculty of Electrical Engineering and Computing, University of Zagreb	14	7,1%
Faculty of Economics and Business, University of Zagreb	14	7,1%
Faculty of Geodesy, University of Zagreb	10	5,1%
Faculty of Organization and Informatics, University of Zagreb	3	1,5%
College of Information Technology Zagreb	2	1,0%
Other	17	8,6%

III. DATA ANALYSIS

A. *Sources of initial introduction to Open Source software*

When asked about how they were introduced to the concept of Open Source software, respondents could choose multiple options, allowing them to demonstrate the full range of information sources.

The results show that most of the respondents (80.3%), gathered information about Open Source software through personal research and interest in the software and operating systems. The next most common way of getting acquainted with Open Source software is the recommendation of friends, which makes 56.1% of total responses. In the domain of private interests, with the previous two methods, may be included magazines and articles related to this topic. Given the fact that recently more reviews of Open Source software can be found in information technology and other professional journals, it is not surprising that 23.7% of respondents name this as source of initial introduction to Open Source software. Except through a private domain, respondents were able to be aquatinted with this concept through "official" means, as a part of education, from elementary school to tertiary education. Even 35.3% of respondents found out initial information about Open Source software through the material and lectures at the university, 24.3% by recommendation of professor and 23.1% by recommendation of teaching assistants. The smallest percentage of the collected answers referred to, as expected, curriculum of elementary school (1.7%) and high school (5.8%) as sourced of initial introduction to Open Source software.

TABLE II. SOURCES OF INITIAL INTRODUCTION TO OPEN SOURCE SOFTWARE

Source	Respondents	%
Personal research	139	70,56%
Recommendation of a friend	97	49,24%
University lectures	61	30,96%
Recommendation of university professor	42	21,32%
Article in magazine	41	20,81%
Recommendation of teaching assistant	40	20,30%
Secondary school curriculum	10	5,08%
Elementary school curriculum	3	1,52%

Given that the highest percentage of responses is in the domain of the private interests of the respondents, it can be concluded that this was their reaction because of the lack of educational support, which can also be seen from the percentage of related materials in primary and secondary schools. Pretty encouraging is percentage of responses related to referrals from professors and teaching assistants at the higher education level, but it could be expected because education on this level provides specialized type of education within which we can expect better representation Open Source software.

B. *Attitudes towards Open Source software as an alternative to commercial software*

The issue was raised with the phrase "commercial software", but in order to eliminate possible confusion, in example we oppose OpenOffice (as an example of Open Source software) and Microsoft Office (as proprietary or commercial software). The reason for raising this issue lies in the rapid development and progress that has occurred in recent times, and which popularized Open Source software on the market. Working with Open Source software today, observing user experience, equates to work on proprietary software and tools, and this will be an analysis of this issue to confirm or refute. The analysis is conducted over the purified sample obtained by elimination of the respondents who said they were not familiar with the concept of Open Source software. Therefore analyzed are only the responses that are qualified to make judgments about whether Open Source software can serve as an adequate alternative to proprietary software.

The results show that 75.7% of respondents believe that Open Source software can provide a worthy alternative to commercial or proprietary software. The high degree of assurance that Open Source software can provide alternative to commercial solutions in everyday work demonstrates the need for the introduction of Open Source content to the regular curriculum.

C. *Should Open Source technologies be used in information and computer science education*

Majority of respondents (92.7%) think that the Open Source software should be used in information and computer science education, while only 7.3% of them responded negatively.

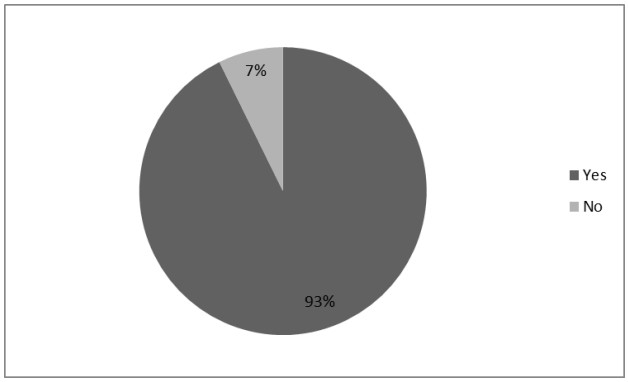


Figure 2. Should Open Source technologies be used in information and computer science education

D. *Should Open Source technologies be included in IT curriculum*

Majority of students (91,6%) think that Open Source technologies should be included in IT curriculum.

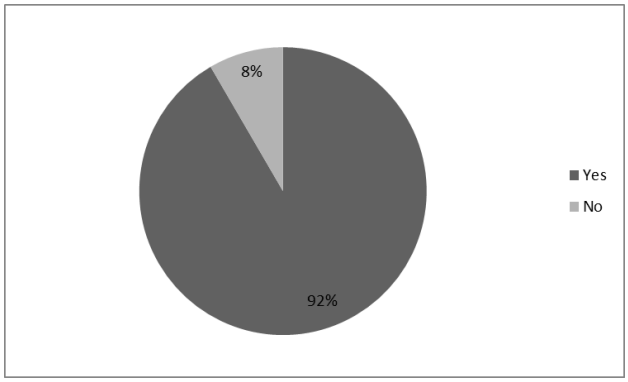


Figure 3. Should Open Source technologies be included in IT curriculum

E. *At which level should begin education about Open Source technologies*

Opinions about level at which should begin education about Open Source technologies are divided, so 49.1% of respondents think that Open Source technologies should be introduced into elementary school curriculum, while 35.2% think that children should start learning about Open Source technologies in high school. Introduction to Open Source technologies in higher education supports only 15.7%.

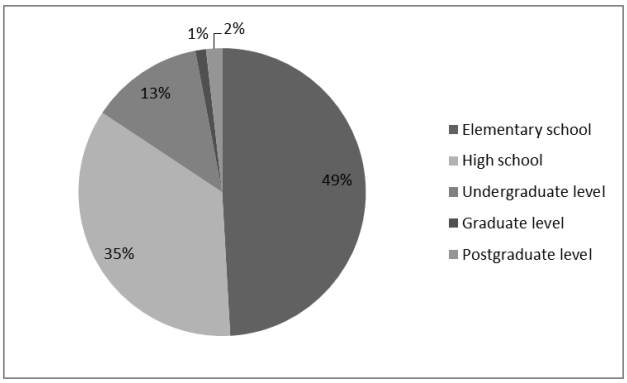


Figure 4. At which level should begin education about Open Source technologies

IV. CONCLUSION

The results of this research present students opinions that Open Source software can provide alternative to proprietary software which can be explained by development of Open Source software which have achieved user experience similar to proprietary.[3] The results also indicate that respondents think that education about Open Source software should start much earlier. Based on the responses of majority of respondents, it should start during primary or at least secondary education, which is in complete contrast to their educational experiences. According to that we can conclude that it is necessary to start a discussion related to the curriculum and technology that is used and taught in primary and secondary education because in rapidly changing world of information technologies it is necessary to implement changes into curriculum which will enable students to use available Open Source technologies.

REFERENCES

[1] K. J. O'Hara and J. S. Kay, Open source software and computer science education, Journal of Computing Science, Volume 18 Issue 3, February 2003, Pages 1-7

[2] P. Carmichael and L. Honour, Open Source as appropriate technology for global education, International Journal of Educational Development, Volume 22, Issue 1, January 2002, Pages 47-53

[3] S. W. van Rooij, Perceptions of Open Source versus Commercial Software: Is Higher Education Still on the Fence?, Journal of Research on Technology in Education, Volume 39, Issue 4, Sum 2007, Pages 433-453

More efficient training of teachers of informatics using modern information-communication technology in Croatian Education and Teacher Training Agency

E. Ban*, T. Bronzin**
Agencija za odgoj i obrazovanje, Rijeka, Hrvatska, eban@azoo.hr
Citus, Zagreb, Hrvatska, tbronzin@citus.hr

Summary - Education and Teacher Training Agency i s a public institution independent regarding its activities and management according to law. It is founded by Republic of Croatia, and seated in Zagreb. Its activities are normalized by Constitution and consider executing professional and advisory acts regarding education, in which an important role plays organizing and conducting training of educational workers and school principals. Because of complexity and specificity of content, educational methods and resources, special attention has to be given to training of informatics and computer science teachers. Intensive introduction of new IC technologies implies a demanding task of its proper use. Especially big expectations in the domain are represented by acquisition of digital competitions which is responsibility of educational system and its workers, who need the biggest support in their training. This paper is an overview of existing IS currently used in Agency and a more detail review of ICT usage in the system of teacher training. Question is raised will the introduction of modern information-communication technology advance and improve the system of teacher training, will it make it more efficient, and how can that be accomplished. Finally, there are given some propositions of possible solution of the problem.

I. INTRODUCTION

The development and rapid progress of modern information and communication technology (ICT) has significantly changed the school system, and the Ministry of Education and Sports has an impressive infrastructure that is able to provide a series of quality services [8]. In the educational process, the use of ICT and e-learning (LMS) is very well represented and welcomed as a teaching tool and an aid, not only in the subjects of which it is the object of study and teaching, but is increasingly used in the traditionally more conservative environments as well [9]. Nevertheless, the training of teachers of informatics, organized and conducted by the Agency for Education (the Agency) is mainly performed in the classical way through lectures, workshops and seminars. After describing the organization of the Agency and its activities, currently used training model is described, as well as some informal alternative options that are used, but are not a part of the system. To complement this model, the introduction of an additional system for training through e-learning according with the existing model is proposed. The assumption is that using this method would result in greater efficiency and better results, especially with the use of cutting-edge technology and positive examples of the good practice so far. Construction and application of such a system would be much easier if the legal framework through ordinances would regulate the rights and obligations of all stakeholders in the process.

II. EDUCATION AND TEACHER TRAINING AGENCY

With its tasks and activities of the Agency permeates the Croatian education system that in the school year 2012./2013. had over 57 000 employees [1]. Besides them, the system also includes more than 525 000 students in elementary and secondary schools [1]. The Agency is a public institution, independent in its operations and transactions in accordance with law founded by the Republic of Croatia, based in Zagreb, in 24th Badalićeva Street. Its activity is regulated by the Statute, and applies to professional and consulting activities in the field of education. Agencies involved in monitoring, promoting and developing education in preschool, elementary and secondary school education, adult education and education for children of Croatian citizens abroad and the children of foreign citizens. [2]

Agency's work is reflected in the performance of the following activities:

- participates in the design, development and implementation of the national curriculum,
- provides professional assistance and provides guidance to institutions, their directors and teachers, teachers and experts (hereinafter referred to as the educational workers) in the implementation of activities related to education,
- organizes and conducts training of educational staff and directors,
- conducts professional examinations of educational workers
- carry out the procedure for the professional advancement of educational workers
- gives opinions about the programs in preschool, teaching programs in primary schools and high schools, the general

education programs of secondary vocational education and on programs in adult education,
- gives opinions about the curricula of academic subjects in the process of verification of educational institutions,
- monitors the development and implementation of the Croatian National Educational Standard as part of the national curriculum,
- conducts the professional pedagogic supervision,
- participates in the development and implementation of student shows and competitions,
- participates in the development and monitoring of national programs,
- performs information-documentation and publishing activities,
- performs other tasks in the system of education in accordance with the law and regulations of the minister responsible for education [2].

Organizational units of the Agency are departments, divisions and subsidiaries (Picture 1):

- Department of legal, accounting and ancillary operations, which consists of: Department legal, personnel and supporting activities and accounting operations of the Department,
- Department of Administration Department consists of: Department of Common Affairs, Department of International Cooperation and the Department of Education and Training Director,
- Department for preschool and elementary education are: the Department of Preschool classroom instruction and the Department for subject teaching,
- Department of pedagogic supervision with the executors In Zagreb, Split, Osijek and Rijeka
- The Department of Secondary Education and
- Branches in: Rijeka, Split and Osijek [3].
Overall the Agency employs 130 people.

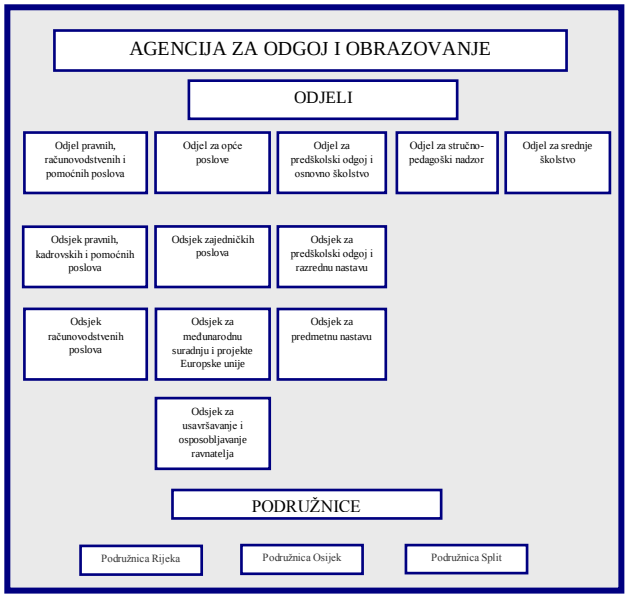


Fig 1. Organizational units of the Agency are departments, divisions and subsidiaries

III. IT INFRASTRUCTURE

The information system is mainly used for data processing, and not for its collection, storage, preservation and distribution. It can be considered small, well organized and with a low degree of automation [5].

Computer equipment, computers and network devices have different configurations and average age of a few years, but they are regularly serviced and restored. Most computers have operating system of MS Windows and versions depend on the quality and age of computers, and user software used for data processing and preparation of text documents is MS Word, also in different versions. Network access to the Internet is protected and is mainly used to print to shared printers and communicate via e-mail with a possibility of access via web interface.

Most of these systems are organized and operated in the same way so that the list of activities and various documents from the environment enter the system, and can then be processed, and ultimately be responded with written documents which represent output from the system [4]. Although there is a possibility of sharing files in a network, it is not used and it is one of the major drawbacks of this information system. Each of the parts of the system works independently when processing data coming from the environment, and after processing returns it back into the environment. Mutual communication is conducted through other media.

All parts of this system are effective in their own work. The problem is that they are separated from each other, and when working with data from the environment it leads to redundant operations with data and as a result redundant information. In this way formed a large amount of redundant data stored mostly in traditional media.

It is very important to connect all these pieces into a complete information system with a single database of users and all relevant information necessary for the Agency.

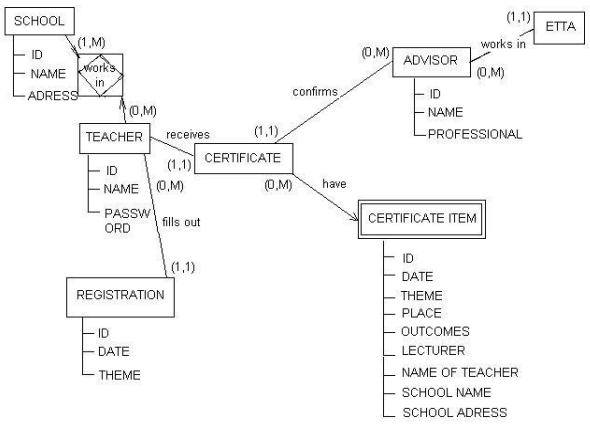


Fig 2. Entity-relationship diagram of system for vocational training

IV. TRAINING

Professional training of employees in education and training is one of the fundamental activities and is defined by the mission: The agency provides professional support and professional development for educational professionals in order to create a society based on competent and innovative educational workers [7]. This system provides an important component of the quality concerns for the timely and adequate information to teachers about contemporary developments in ICT, pedagogical and psychological expertise, methods and more.

System that includes professional development activity consists of several separate processes and procedures of which part is automated. Professional development is the right and obligation of every employee in the education system. Process is started by adviser who designs the form of professional development execution, which influences on how the content is prepared. Because of the fact that each school subject is specific, content is prepared separately for each subject. Usually the choice of speaker is connected tightly connected to the process of content preparation, because (often) speaker is involved in content creation as an expert. After schedule for training/workshop is set, all of these informations are published on the Teachers Training Agency (AZOO) web site in an online catalog. The catalog application that is located within the Agency's web site provides possibility of an on-line application form and registration for the selected training, workshop or symposium. The application provides the possibility for authorized persons to automatically generate certificates for the participants who attended the meeting after meeting director confirms participation of participants [14]. In the time of the launch of this application in the year of 2010 there were already visible benefits of this system of administering professional development. In the meantime, application grew and added various options that reduce visibility and ease of use by the user, and thus its effectiveness. It should be thoroughly processed by creating multiple smaller ones (Fig. 2.).

Such a system does not have the characteristics of distance learning, it uses only some of the features of this model. After the lectures, presentations material is published on the Agency's website, to be inserted as part of repository. Most of the meetings where teacher of informatics delivered lectures were broadcasted using Live Meeting Service. Today Microsoft Office 365 Edu / Lync components is used for live broadcast of meetings for recording them. This on-line services offer possibility to prepare recorded materials to be used as trainings on demand [6]. IT teachers in the last few years, had the opportunity to participate in numerous webinars and videoconferences through which, in their training they used modern IC technology, but this practice is sporadic and not a rule but an exception and depends on many external factors outside the system. However, even in these conditions it is possible to highlight the benefits of the underlying reasons for introducing the model of distance learning, and that is the opportunity to participation to event regardless of the location of the user, if the technical requirements are met and access to materials is available at any time. Although it is possible to make contact with the lecturer or director of the symposium, this option is rarely used, and becomes completely unavailable at the moment of completion of the symposium or training. The feedback, an important element of distance learning, gets lost and there are records of access to the materials or of time of using, as well as evaluation or use of materials. Once these materials are created and put into the repository they are very rarely changed after that.

Fig. 3. Number of professional development trainings

V. SYSTEM FOR TRAINING THROUGH E-LEARNING

A. Existing Resources

Existing ICT infrastructure has the quality and capacity to solve complex problems that could arise by introducing LMS which would assist in training at the Agency. Given the experience of using modern ICT and knowledge of algorithms working in this environment it should not be a problem for a computer science teacher to deal with such system [9]. This example might later serve as a model for the development of LMS for other subjects, and they themselves might play the role of a coach of the teaching staff.

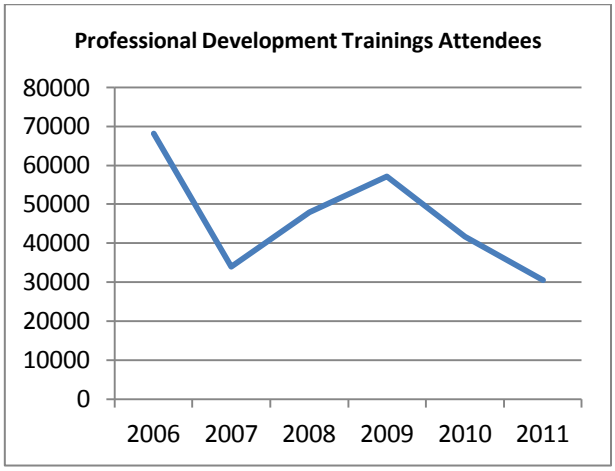


Fig. 4. Number of professional development trainings attendees

B. Benefits - the reasons for the introduction of the system

The particular geographical position and shape our country puts additional problem of physical distance that should be travelled to attend the symposium for the organizers of training. Winter weather conditions are an additional aggravating circumstance. Lack of funds in the school budget as a result of the global and local economic recession has impacted heavily on the number of participants in vocational training (Fig. 3 and 4) and almost halved as appose to the previous year. By introducing a model of distance learning and adequate LMS with support of modern tools this problem could be significantly reduced. In preparation for their work and daily responsibilities in the transfer of knowledge to younger generations teachers are continually under certain pressure monitoring and learning new technologies and thus adequate and freshly prepared working papers that can serve as an excellent addition to professional training facilities.

C. Description of the system

For this system could be chosen, Moodle, for instance, an open-source system, which has had a strong community of developers and that is constantly being improved. Web 2.0 tools would be suitable as well, as it is one of the characteristics that have a user interface based on user interaction with the system and its movement through hyperspace, showing hypermedia pages and offers a variety of tools to navigate in this space [15]. Because creative component of distance learning means and human support in the process could include professional managers of county councils, faculty mentors, counselors and teachers [12]. Some colleagues have attended the ELA-(E-learning academy) and acquired valuable knowledge about working in the distance learning system. They would also be able to participate in the system in the form of mentors and moderators of individual cycles as they have with their work and dedication shown extra qualities. For all these activities there should be provided adequate support and evaluation system for career advancement.

TABLE I presents the data of the Agency for the period since year 2006 by year 2011 the number of employees and the number of activities in the scope of activities of the agency involved, and the number of participants in certain activities.

The system can be improved with web recommender systems, which could facilitate the use of the system, and would also make working on system more interesting and dynamic. [11]. This system should not completely abolish present method of the organization and delivery of training that allows the realization of social contact which is important for a successful operation, but it can be reduced by using videoconferencing tools. [13]. After finalizing the components of training within the system, it would be possible to create a kind of e-portfolio in which includes all the relevant information necessary for the teachers CV.

Given the speed of change in ICT and its lifetime, one of the solutions could be setting the whole system in the cloud, which would permanently resolve the issue of maintaining the necessary hardware and software infrastructure required to operate the system. [10].

D. Initial indicators introduced components of the future system

If we assume that some of the described systems are actually an integral part of our future system for training, then the results and experiences obtained by their analysis could be used to assess the effectiveness and quality. Number of participants of webinars increases steadily as opposed to the number of participants who physically attend the symposium, whose number is significantly reduced (Fig. 1 and 2). The number and speed of introduced technologies could be an indirect indicator of the effectiveness of training. In this case, the great advantage has the system in which it is possible to re-examine the lecture and any other relevant information that the user needs at that very moment. The process of application and obtaining a certificate of participation in professional training is significantly improved and fully automated. This system thereby achieves greater speed, quality and credibility, and hence greater effectiveness. The biggest problem and possible restrictions on the introduction of this system could result in an additional reduction of funds provided for training which would prevent even the minimal integration of existing components which would form a single system for training with the help of ICT.

AZOO	2006	2007	2008	2009	2010	2011
Employees	138	136	134	134	135	133
Advisers	115	110	107	107	106	105
Administration	23	26	27	27	29	28
Professional support	1099	2423	3395	3837	3600	4282
Symposiums	1028	523	853	948	740	871
Participants	68180	33971	47873	57166	41623	30500
Professional exams	2107	2230	3091	3233	3095	3011
Promotions	554	783	811	1251	923	1788
Directors of ŽSV	1312	1420	1427	1432	1429	1429
Elementary school	866	926	928	934	927	927
Highschool	448	494	499	498	502	502
ŽSV conference	3059	3892	3052	4401	2519	1313
Participants	97711	94194	77519	111236	68421	14581

VI. CONCLUSION

This article describes a proposal for the training of teachers of informatics based on a hybrid model of distance learning with the help of modern ICT. Such a model with its advantages of distance learning could contribute to the quality and success of the overall need for continuous professional development in the field. The system should be based on adaptive hypermedia with the possibility of introducing a system for recommendations. The necessary material preconditions for the introduction of the system based on strong IT infrastructure and a number of highly qualified personnel already exists and is capable to carry out the initial phase of the application of such a model, but it is necessary to make additional efforts in terms of material preparation and mastering technology. All activities in this model of professional development should be introduced into the system of evaluation for advancement in the profession and be generally regulated by the purpose-legal acts. By analyzing outcomes of some elements of the future system that are occasionally organized, it can be assumed that the complete system of training using modern ICT would enable more effective training of Informatics teachers.

As the next step, it should be conducted a detailed analysis of the costs and feasibility of introducing such a system and compare the expected savings with the cost of the introduction. Furthermore, there could be conducted a survey using a questionnaire within the target group about the desirability of such a model of training and readiness to participate in its preparation, making and implementation.

REFERENCES

[1] Central Bureau of Statistics of Republic of Croatia, „Statističke informacije 2012," Zagreb, 2012.

[2] The Governing Council of AZOO."Statut Agencije za odgoj i obrazovanje," Zagreb, 2007.

[3] Croatian Education and Teacher Training Agency – structure. (n.d.) Retrieved 2013. from http://www.azoo.hr/index.php?option=com_content&view=category&id=34&Itemid=55

[4] M. Pavlić, „Informacijski sustavi", Školska knjiga, Zagreb, 2011.

[5] M. Pavlić, „Razvoj Informacijskih sustava — projektiranje, praktična iskustva, metodologija", Znak, Zagreb, 1996.

[6] AZOO-conferences – Retrieved 2013. from http://ucitelji.hr/Obrazovnisadržaji/Stručniskupovi/AZOOskupovi.aspx

[7] AZOO – work report 2011, Zagreb, 2012.

[8] Ministry of Science, Education and Sports, Informational infrastructure, Retrieved 2013. from http://public.mzos.hr/Default.aspx?sec=1961

[9] R. Zemsky, i W.F. Massy, „Ometena inovacija: Što se dogodilo e-learningu i zašto?" , 2006., Edupoint, 47(VI). Retrieved 2013. from http://www.carnet.hr/casopis/47/clanci/3

[10] Williams, Mark I. „A quick start guide to cloud computing: moving your business into the cloud", London, 2010.

[11] P. Brusilovsky, "Adaptive navigation support: From Adaptive Hypermedia to the Adaptive Web and Beyond, PsychNology Journal, Vol. 3, 2004, pp. 7-23

[12] N. Hoić-Božić, „Prilagodljiva hipermedijska programska potpora za učenje – doktorska disertacija", Zagreb, Fakultet elektrotehnike i računarstva, 2002.

[13] F. Alonso, G. Lopez, D. Manrique, J. M. Viñes, "Instructional model for e-learning with a blended learning process approach," British Journal of Educational Technology, vol. 36, no. 2, 2005.

[14] AZOO – conferences – Retrieved 2013. from http://www.ettaedu.eu/

[15] AHyCo – home page, http://ahyco.uniri.hr/portal/Pocetna.aspx, 2013.

[16] CARNet – e-learning akademy – Retrieved 2013. from http://www.carnet.hr/ela

Student's level of ICT literacy by study groups on Faculty of Philosophy, University of Mostar

A. Ivanković*, S. Špiranec ** and D. Miljko
* Faculty of Philosophy, University of Mostar/Department of Information Sciences, Mostar, Bosnia and Herzegovina
** Faculty of Humanities and Social Sciences, University of Zagreb/ Department of Information Sciences, Zagreb, Croatia
andreaivankovic@gmail.com; sspiran@ffzg.hr; drazan.miljko@gmail.com

Abstract - The past decade has brought an increasing rate of ICT use for learning, teaching, research and communication and has led to significant changes in student characteristics, needs and expectations. Development in these areas changes student profile everyday, which have prompted many universities to reconsider the quality of ICT literacy they provide to students. For this purpose, a study was conducted among students, which was used to determine level of ICT literacy of some study groups. The study involved 807 students. Since there are great combinations of two-subject study groups we considered only those that include more than 80% of students attended some type of ICT courses, groups that include 40% to 60%, and groups with less than 20% of students attended such courses. Respondents anonymously accessed the questionnaire, which consisted of 15 questions. Correct answers on a test were treated as total score, i.e. simple linear combination. Therefore, the differences among the 7 analysed groups were tested according to their overall result on the test. The results are interpreted in the context of the level of presence of ICT courses during the education. There was a statistically significant difference in the level of ICT literacy among the different study groups.

I. INTRODUCTION

Information and communication technology (ICT) has become a key tool in the 21st century and has had a revolutionary impact on how we see the world that we live in. [5]

Nowadays, the concept of literacy has changed. Being literate does not mean only writing, reading and numeracy skills, but requires a whole range of new skills. It implies different literacy skills such as ICT literacy, digital literacy, computer literacy, technological literacy, media literacy, information literacy, etc.

Literacy can be divided into two significant periods, the period before the occurrence of electronic computers and literacy after that. The first steps of ICT literacy occur with the appearance of electronic and later electronic calculating machines. [2]

Despite the development of the Internet and other technologies, today's college students might not have the ICT literacy skills - the ability to effectively research and communicate using technology - necessary to navigate and make good use of the overabundance of information available today. Just because someone can navigate to a Web site, or even know how to construct Web sites, it does not mean that the person can identify reliable, authoritative resources from the Web or know how to best interpret and communicate a Web site's content via a well-supported argument. In other words, technological competence alone does not equal ICT literacy. [3]

Therefore, discussion about ICT literacy should begin with the concept of information literacy.

In the year 1989, American Library Association (ALA), formulates one of the most commonly cited definition of information literacy. According to the ALA's Report (Presidential Committee, 1989), information literate persons are defined as "those who have learned how to learn because they know how knowledge is organized, how to find information and how to use them in ways others can understand. Those are the people prepared to learn through life". [6]

ICT literacy could be defined as the ability to use digital technology, communications tools, and/or networks to access, manage, integrate, evaluate, and create information in order to function in a information society. [3]. The variety of literacy definitions shows that contemporary societies demand a wide spectrum of different literacies. However, due to ubiquity of ICT's in modern society, there is no doubt that ICT literacy represents a prerequisite and core element of other, more complex literacies.

The use of ICT in education is a relatively new phenomenon. The past decade has brought an increasing rate of ICT use for learning, teaching, research and communication and has led to significant changes in student characteristics, needs and expectations. The development in these areas changes student profiles everyday, which has prompted many universities to reconsider the quality of ICT literacy they provide to their students. [1]

ICT has drastically changed the world and is still changing environment of learning and teaching as well as playing an important role to promote literacy among students. There is no doubt that ICTs are seen as centre of education in the 21st century. However, ICT skills and literacy are more often assumed than explored, in particular among younger users (e.g. pupils, students). Therefore, more research data is needed about the actual level of ICT literacy among different groups of users. This paper explores ICT literacy among students and examines

differences in relation to different study groups, which have integrated ICT subjects into their curricula to a different extent.

II. ICT LITERACY RESEARCH

A. *Methodology and respondents*

The main hypothesis of the research was that study groups differ in relation to the level of integration of ICT courses. As a consequence, a variation of ICT literacy among different study groups might be expected.

The research was undertaken among 807 students from different study groups, both undergraduate and graduate level (Figure 1). There were 211 (26,15%) male and 596 (73,85%) female students whose average age was 21,45 years.

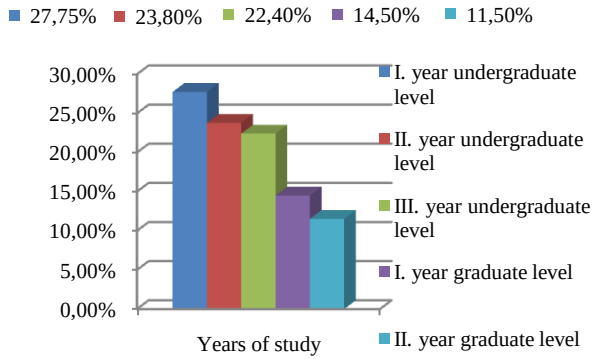


Figure 1. Respondents by study years

However, since there is a variability of combinations of double-subject study groups on the Faculty of Philosophy, it was difficult to categorize groups of those with and without some form of ICT courses in their curriculum.

Therefore, we considered only those study groups that include more than 80% of students who attended some types of ICT courses ((Journalism - 89,29%; Politology - 83,94%), groups that include 40% to 60% of students attended an ICT course (Psychology - 44,44%; Social work - 42,68), and groups with less than 20% of students attended such courses (Croatian language and literature and English language and literature - 3,77%, English language and literature and Philosophy - 10,14 %, English language and literature and German language and literature - 9,76%), i.e. 651 student included in this research. Also, for the purposes of determining the difference between the groups using the chi-square test, we considered only those study groups with more than 50 students.

The ICT literacy of students was tested with knowledge test consisted of 15 questions related to different aspects of the computer functioning and operating.

B. *Statistics*

Correct answers in the test of 15 questions were treated as total score, i.e. simple linear combination of correct answers, and the differences were tested with one-way analysis of the variance of independent samples, according to seven analysed study groups. Also, we used the chi-square test to determine the differences between the tested study groups, for answers on every specific question. A statistically significant difference in the level of ICT literacy among the different study groups ($p < 0,001$) was determined.

Statistical data analysis was conducted by computer program Statistica 7.0 (StatSoft, Inc.).

C. *Results*

ICT literacy test results are presented in tables. In the table we used the following phrases:

Study groups:

[1] Journalism

[2] Politology

[3] Psychology

[4] Social work

[5] Croatian language and literature and English language and literature

[6] English language and literature and Philosophy

[7] English language and literature and German language and literature

T- Correct answers

F- Incorrect answers

S -Total number of students by study groups

TABLE I. ICT LITERACY TEST

	Questions						
	Study groups						
	[1]	**[2]**	**[3]**	**[4]**	**[5]**	**[6]**	**[7]**
1	Witch are the digits the computer world is build on?						
T	75,00 %	73,91 %	89,09 %	59,49 %	69,81 %	85,51 %	68,29 %
F	25,00 %	26,09 %	10,31 %	40,51 %	30,19 %	14,49 %	31,71 %
S	21,41 %	21,10 %	8,41 %	24,16 %	8,10 %	10,55 %	6,27 %
2	What of the listed represents the operating system?						
T	41,43 %	40,58 %	69,09 %	34,39 %	45,28 %	62,63 %	53,66 %
F	58,57 %	59,42 %	30,91 %	65,61 %	54,72 %	37,68 %	46,34 %
S	21,44 %	21,13 %	8,42 %	24,04 %	8,12 %	10,57 %	6,28 %
3	What is Facebook?						
T	99,29 %	97,83 %	100,00%	99,37 %	100,00%	98,55 %	97,56 %
F	0,71 %	2,17 %	0,00 %	0,63 %	0,00 %	1,45 %	2,44 %
S	21,44 %	21,13 %	8,41 %	24,16 %	8,10 %	10,55 %	6,27 %
4	What is http?						
T	21,43 %	22,46 %	32,73 %	29,94 %	32,08 %	34,78 %	26,83 %
F	78,57 %	77,54 %	67,27 %	70,06 %	67,92 %	65,22 %	73,17 %

	Questions						
	Study groups						
S	21,44 %	21,13 %	8,42 %	24,04 %	8,12 %	10,57 %	6,28 %
5	Whose founder was Steve Jobs?						
T	83,57 %	84,67 %	92,73 %	64,56 %	86,79 %	84,06 %	85,37 %
F	16,43 %	15,33 %	7,27 %	35,44 %	13,21 %	15,94 %	14,63 %
S	21,44 %	20,98 %	8,42 %	24,20 %	8,12 %	10,57 %	6,28 %
6	Which number in the picture represents the space bar button?						
T	90,00 %	83,33 %	90,91 %	68,35 %	98,11 %	97,10 %	95,37 %
F	10,00 %	16,67 %	9,09 %	31,65 %	1,89 %	2,90 %	4,88 %
S	21,41 %	21,10 %	8,41 %	24,16 %	8,10 %	10,55 %	6,27 %
7	What shortcut means Cut?						
T	22,86 %	26,09 %	43,64 %	34,18 %	22,64 %	23,19 %	31,71 %
F	77,14 %	73,91 %	56,36 %	65,82 %	77,36 %	76,81 %	68,29 %
S	21,41 %	21,10 %	8,41 %	24,16 %	8,10 %	10,55 %	6,27 %
8	What is the latest operating system produced by Microsoft in the 2011. year?						
T	29,29 %	26,81 %	23,64 %	25,95 %	22,64 %	36,09 %	17,07 %
F	70,71 %	73,19 %	76,36 %	74,05 %	77,36 %	73,91 %	81,93 %
S	21,41 %	21,10 %	8,41 %	24,16 %	8,10 %	10,55 %	6,27 %
9	Which one of listed is an IP address?						
T	52,86 %	57,25 %	58,18 %	38,61 %	45,28 %	53,62 %	58,54 %
F	47,14 %	42,75 %	41,82 %	61,39 %	54,72 %	46,38 %	41,46 %
S	21,41 %	21,10 %	8,41 %	24,16 %	8,10 %	10,55 %	6,27 %
10	Which function on the computer reruns the computer?						
T	87,14 %	80,43 %	98,15 %	83,54 %	90,57 %	95,65 %	90,24 %
F	12,86 %	19,57 %	1,85 %	16,46 %	9,43 %	4,35 %	9,76 %
S	21,44 %	21,13 %	8,27 %	24,20 %	8,12 %	10,57 %	6,28 %
11	@ sign is an integral part of:						
T	89,29 %	87,68 %	96,36 %	86,08 %	90,57 %	94,20 %	90,24 %
F	10,71 %	12,32 %	3,64 %	13,92 %	9,43 %	5,80 %	9,76 %
S	21,41 %	21,10 %	8,41 %	24,16 %	8,10 %	10,55 %	6,27 %
12	Which one of the listed represents a wireless network?						
T	73,91 %	67,88 %	72,73 %	57,96 %	69,81 %	68,12 %	80,49 %
F	26,09 %	32,12 %	27,27 %	42,04 %	30,19 %	31,88 %	19,51 %
S	21,23 %	21,08 %	8,46 %	24,15 %	8,15 %	10,62 %	6,31 %
13	What is the meaning of the abbreviation PC?						
T	74,29 %	72,26 %	87,27 %	58,23 %	86,79 %	88,41 %	80,49 %
F	25,71 %	27,74 %	12,73 %	41,77 %	13,21 %	11,59 %	19,51 %
S	21,44 %	20,98 %	8,42 %	24,20 %	8,12 %	10,57 %	6,28 %

	Questions						
	Study groups						
14	What program is used for website browsing?						
T	49,62 %	44,53 %	47,06 %	28,38 %	62,75 %	43,33 %	48,65 %
F	50,38 %	55,47 %	52,94 %	71,62 %	37,25 %	56,67 %	51,35 %
S	21,88 %	21,05 %	8,39 %	24,34 %	8,39 %	9,87 %	6,09 %
15	What file extension makes only an image file?						
T	83,45 %	87,68 %	94,55 %	81,01 %	98,11 %	95,65 %	95,12 %
F	16,55 %	12,32 %	5,45 %	18,99 %	1,89 %	4,35 %	4,88 %
S	21,29 %	21,13 %	8,42 %	24,20 %	8,12 %	10,57 %	6,28 %

Differences in students ICT literacy among different study groups are presented in the following table.

TABLE II. DIFFERENCES IN STUDENTS ICT LITERACY

Questions		X^2	df	P
1	Witch are the digits the computer world is build on?	27,973	6	p<0,001
2	What of the listed represents the operating system?	31,418	6	p<0,001
3	What is Facebook?	4,028	6	p=0,673
4	What is http?	7,991	6	p=0,239
5	Whose founder was Steve Jobs?	35,549	6	p<0,001
6	Which number in the picture represents the space bar button?	57,776	6	p<0,005
7	What shortcut means Cut?	13,291	6	p=0,039
8	What is the latest operating system produced by Microsoft in the 2011. year?	3,005	6	p=0,808
9	Which one of listed is an IP address?	14,964	6	p=0,021
10	Which function on the computer reruns the computer?	18,563	6	p=0,005
11	@ sign is an integral part of:	6,902	6	p=0,330
12	Which one of the listed represents a wireless network?	13,098	6	p=0,042
13	What is the meaning of the abbreviation PC?	38,499	6	p<0,001
14	What program is used for website browsing?	24,189	6	p<0,001
15	What file extension makes only an image file?	23,031	6	p<0,001

The highest percentage of the accurate answers that the students offered to the question was "What is Facebook?" (98,93%), and therefore, there were determined no differences in knowledge among the study groups (p=0,673).

Furthermore, nearly 90% of students of the Faculty of Philosophy know what a sign @ is. Also, there were no differences between the study groups (p=0,330).

Almost 90% (precisely 87,90%) of the students know what file name extension indicates an image file. Regardless, there were some differences determined among the study groups (p<0,001). While most of the

groups have achieved scores above 90%, Social work students (81,01%) and studies with majority of ICT courses as part of their curriculum achieved less correct responses corresponding to other groups (Journalism (83,45%) and Politology (87,68%)).

A large number of students (87,14%) knows what function reruns computer. However, differences have been found in the frequency of correct responses among the study groups (p=0,005). Although all groups achieved accuracy above 80%, high scores of correct answers were achieved by Psychology students (98,15%), and a double-subject English language and literature and Philosophy students (95,65%) and least often Politology students (80,43%).

Also many of them (85,17%) can recognize the "space bar" key on keyboard illustration. However, Social work students have significantly lower accuracy (68,35%) in answering to this question (p<0,005), compared to students from other studies. Moreover, students from Croatian language and literature and English language and literature groups achieved almost 100% accuracy.

Slightly more than 80% of students (precisely 80,40%) know that Steve Jobs was the founder of Apple. However, we have determined differences in the knowledge of study groups (p<0,001) - while most of the tested groups in more than 80% of cases provide the correct answer to this question, Social work students achieved an accuracy of 64,56%.

Three quarters of the students (73,97%) know what the acronym PC means. Among them, with significantly less frequency (p<0,001), Social work students (58.23%) have given the correct answers.

About two-thirds of the students (precisely 72,48%) are familiar with the fact that computers are functioning across the binary system. However, differences have been found in the level of knowledge among the study groups (p<0,001). The most accurate answers were provided by Psychology students (89,09%), and least accurate responses by the study group Social work (59,49%). In both of these study groups, about 50% of the students attended some form of ICT courses.

The correct answer to the question “Which one of the listed represents a wireless network?” was given by 68,15% of students. However, differences have been found in the responses of study groups (p=0,042). The difference is reflected in a significantly higher frequency of correct answers of the English language and literature and German language and literature study group (80,49%) than the group of Social work (57,96%). In any case, all of the tested groups are achieving more than 50% of correct answers.

About 50% of students know which symbols represent the IP address. Among them, there are differences among students of different study orientation (p=0,021). The most correct answers obtained groups English language and literature and German language and literature (58,54%), Psychology (58,18%) and Politology (57,25%). On the other hand, less than half of Social work students (38,61%) and the Croatian language and literature and English language and literature students (45,28%) achieved the correct answers.

In the following we would show questions with the accuracy of answering below 50%.

Among all students, less than half of them can recognize, what an operating system is (45,18%). The comparison of study groups showed significant differences in representation of correct answers (p<0,001). Psychology students (60,09%) and English language and literature and Philosophy combination students (62,32%) gave the most correct answers. Opposed to them, Social work students provided correct answers in the lowest frequency (34,39%).

Slightly less than 50% of students have recognized the name of the program for Web browsing. With this question, we determined differences among different groups (p<0,001). The least correct answers were from Social work students (28,38%), and the most correct from Croatian language and literature and English language and literature students (62,75%).

It should be noted that a certain percentage of respondents (about 7%), marked two of the four answers at the same time, of which one was correct and the other not. Namely, these groups of students consider that both, Google and Mozilla Firefox, are intended for Web browsing.

Generally speaking, less than a third of the students (28,59%) know how to use a shortcut for the command "Cut". How to use this command usually know Psychology students (43,64%), Social work students (34,18%), and English language and literature and German language and literature students (31,71%). In contrast, students of study groups that in majority listened to some of ICT courses, achieved the correct answers in significantly lower frequency (p=0,039) (Journalism - 22,86%, Politology – 21,10%).

Only 27,26% of the students know what the abbreviation http stands for. However, there were determined no differences between the study groups, students in different study orientations equally know the answer to this question (p = 0,239).

The question that had the least correct answer (25,84%) is “What is the latest operating system produced by Microsoft in the year 2011?”. Among the study groups there were determined no differences in the frequency of (in) correct responses (p=0,808).

TABLE III. TABLE OF AVERAGE VALUES SCORED ON THE ICT LITERACY KNOWLEDGE TEST

Descriptive statistics (Include cases 1:654)			
Level of factor	***N***	***Average***	***Std.Dev***
Journalism	140	9,60550	2,654971
Politology	138	9,69286	2,795292
Psychology	55	10,90909	2,327554
Social work	158	8,47468	2,768566
Croatian language and literature and English language and literature	53	10,18868	2,038633
English language and literature and Philosophy	69	10,44928	2,054666
English language and literature and German language and literature	41	10,14634	2,545201
Total	**654**	**9,60550**	**2,678889**

F= 9,453; df= 6; p<0,001

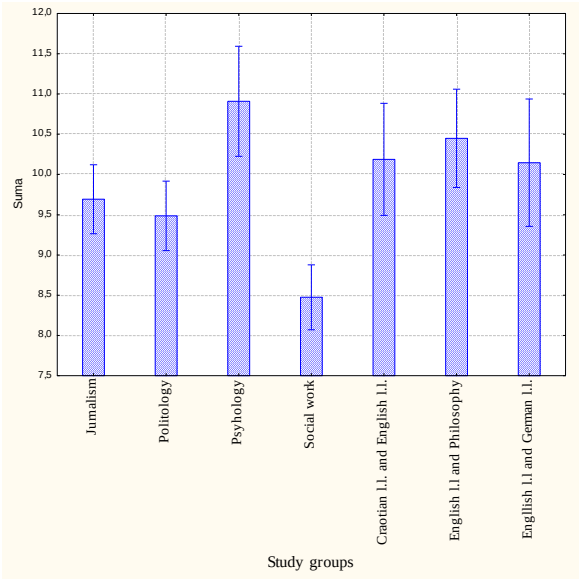


Figure 2. Graphic preview of correct answers on the test by study groups

III. DISCUSSION AND CONCLUSION

Ubiquity of ICTs has brought rapid technological, social, political and global economic transformation. Similar effects are visible in the educational field. Unquestionably, ICT has had an impact on the quality and quantity of teaching, learning and research in the education process. A variety of skills that are part of ICT literacy are supposed to be acquired through education. The question, however, remains, whether this process is effective and how different factors affect it (different study groups, level of integration of ICT courses etc.). A common expectation would be that a higher level of ICT literacy will occur in study groups with obligatory ICT courses.

However, our study has showed that the prevailing distribution of correct answers on specific questions is related non-specifically on study groups, such as Psychology, English language and literature and German language and literature, English language and literature and Philosophy, and to a certain extent on Croatian language and literature and English language and literature - a group whose students in the majority or at least 50% of cases have not attended a course that deals with ICT. Moreover, in most cases (questions), students of study groups that have in their curriculum ICT courses are not preceding, and even give significantly less accurate answers compared to the other groups.

It should be noted that the obtained results may depend on the frequency of computers used during study, for the purpose of meeting the obligations of students (seminars, presentations, literature search on the Internet), which trains certain groups significantly to work on the computer through non-formal education, and thereby it increases their ICT literacy.

Attending ICT courses in their curriculum does not contribute significantly to the ICT literacy, so it is advisable to analyze the content of ICT courses in study groups that are not directly involved in computer sciences.

Moreover, the correct answers to a test of 15 questions are treated as total score, i.e. a simple linear combination of correctly answered, and the differences were tested with one-way analysis of variance for independent samples.

There was a statistically significant difference in the level of ICT literacy among the different study groups ($p<0,001$). Scheffe's post-hoc procedure showed that this difference relates to ICT literacy of Social work and other study groups, except for Politology (between these two groups there were no significant difference in the test results). On average, Social work students achieved significantly lower scores on the ICT literacy test of most other groups. The differences among the other groups were not observed ($p>0,05$).

REFERENCES

[1] A. Martin and D. Madigan, Digital literacies for learning. Facet publishing, London, 2006.

[2] D. Radošević, "Informatika i uloga fakulteta organizacije i informatike u njenom razvoju", Journal of Information and Organizational Sciences, Vol. No. 16, 1992

[3] International ICT Literacy Panel, "Digital transformation: A framework for ICT literacy", Princeton, NJ , Educational Testing Service, 2002.

[4] Katz, I. R. 2005. Beyond technical competence: Literacy in information and communication technology. Educational Technology Magazine 45

[5] S. Moradi, A. Fallah and S. Ahmadi, "Presenting a practical framework for assessing and comparing the ICT literacy of principals and teachers in Iran's schools", Odgojne znanosti, Vol. 11 No. 1(17), 2009

[6] S. Špiranec and M. Banek-Zorica, Informacijska pismenost-Teorijski okviri i polazišta, Zavod za informacijske studije, Zagreb, 2008.

E_learning vs. Multimedia (Ivica & Marica – Hansel & Gretel)

Krunoslav Bedi
Graditeljska škola Čakovec
(School of Building and Craft Čakovec)
Športska 1, HR-40000 Čakovec, Croatia
Phone: +385 40 329 003 Fax: +385 40 329 027
E-mail: krunoslav.bedi@ck.t-com.hr

Nikolina Žajdela Hrustek
Fakultet organizacije i informatike
(Faculty of Organization and Informatics Varaždin
University of Zagreb)
Phone: +385 042 39 08 28 Fax: +385 042 21 34 13
Pavlinska 2, HR-42000 Varaždin, Croatia
E-mail: nikolina.zajdela@foi.hr

Tedo Vrbanec
Učiteljski fakultet Zagreb - Odsjek u Čakovcu
(The Faculty of Teacher Education)
Ul. dr. Ante Starčevića 55, HR-40000 Čakovec, Croatia
Phone: +385 040 370030 Fax: +385 040 370025
E-mail: tedo.vrbanec@gmail.com

Abstract. IVICA I MARICA (HANSEL AND GRETEL) is a project of designing an interactive-educational multimedia DVD for children. The very idea of designing the DVD stemmed from personal experience and suggestions from both teachers and students. Namely, the title of the DVD hides the education of children in the field of environment protection, litter and waste problems, the harmful agents that surround us and the polluters of today. From the very beginning, the efforts were aimed at motivating children to acquire knowledge of the pollutants of today. On the other hand, children get to know about the possibilities of sorting the garbage, i.e. of waste packages management as well as about the optimistic views of the future in this realm. Besides the educational part, the DVD as well contains the attractive funny contents via which the knowledge acquired through the use of the DVD and play can be checked.

The paper shows the parts and the process of creating of the multimedia DVD that were based on the previous own experience of students.

The purpose of the paper was to test the successfulness of the project-based out-of-classroom education environment of the secondary school students. The content was based on the project-based learning and new skills development through team work and honing of the vocational skills by means of multimedia.

Keywords. DVD, multimedia, e_learning, computer interaction

I. INTRODUCTION

The deployment of the information technologies is inevitable in everyday life so its deployment in the teaching process is obligatory. By means of the deployment of e-learning, the education and learning are enhanced. Students thus develop habits that are to accompany them in the future, in the course of the lifelong learning through the future education and-finally-at their workplaces. Besides that, the implementation of new information technologies inevitably changes the teaching process in terms of the reception of learning and training as compared to the traditional teaching process.

Many educators now believe that computers can be used to promote learning and development in early childhood education if they are used appropriately. Children need to be aware of the nature and uses of computers in order to meet the challenges presented by the present and future technological society [1].

This is why the objective of any teacher –on the individual level- is to combine various methods and procedures, define explicit rules and provide for a

good working environment, interesting conceptions and an appropriate level of expertise.

The educative multimedia DVD entitled "Ivica i Marica" can be used in educating of pre-school aged children and lower-grades primary school students in order to provide a facilitated process of acquiring and checking the knowledge related to the ecological matters by means of interesting animations and attractive pictures.

Technology has become user-friendly, and the cost of multimedia production has decreased, thus providing educators with additional degrees of freedom to explore alternatives to traditional education and assessment. [2]. The sensibilisation of conscious in children is required in order for the living environment to be conservated. The objective is to educate children to recognize the dangers stemming from the contemporary way of living (noise, radiation, pollution…) and the potential problems (waste, litter) arising in the environment protection realm. The project was being implemented with the objective of a multimedia influence on children. Under the mentorship of the author of the paper, two third-grade students started creating the multimedia interactive educational DVD entitled „Ivica i Marica“. A wide scope of software was used that is normally used throughout the four-year course. Namely, this is the web-designer/media technician course for whose implementation the entire necessary software infrastructure has been provided by the Ministry of Education and Sports of the Republic of Croatia. The very work was organized as a teamwork as it had been proved that it was far more efficient a work when one of the students did the programming and the other the visual designing. The students worked on the multimedia interactive educational DVD during their optional subject/activity time where the key role of the coordinator was assigned to the mentor, i.e. the optional subject teacher, who is the author of this paper. In their teamwork, the students chose a specific name, "DreamLand“", to work under and this can be seen from the Youtube attached trailer. In this way, the teamwork got a new dimension of being "professional" one and additionally motivated the students. Besides that, for the purpose of the State IT Competition – Infokup 2012. Primošten - a Facebook profile (a group) was created that additionally promoted the multimedia DVD.

II. E-LEARNING

The deployment of the information technologies is inevitable in everyday life so its deployment in the teaching process is obligatory. By means of the deployment of e-learning, the education and learning are enhanced. The multimedia content alone, however, does not necessarily result in significant positive learning performance and satisfaction. Moreover, it is expensive to design and develop multimedia instructional material. [3].

Students should develop habits that are to accompany them in the future, in the course of the lifelong learning through the future education and-finally-at their workplaces. Besides that, the implementation of new information technologies inevitably changes the teaching process in terms of the reception of learning and training as compared to the traditional teaching process. Similar research was already done: the researchers have developed a cartoon-style multimedia application whereas animated cartoons where designed from scratch using appropriate programs [4].

This is why the objective of any teacher –on the individual level- is to combine various methods and procedures, define explicit rules and provide for a good working environment, interesting conceptions and an appropriate level of expertise. Therefore, it is interesting to know what teachers think about possible advantages produced by multimedia tools and their limits. Since teachers are in charge of managing students learning processes, these benefits and weaknesses should concern learning [5].

III. MULTIMEDIA PROJECT

Liu & Hsiao [6] report the results of a research that made students take part in multimedia designing by means of using a project-based learning approach. They say that making students to take part “as multimedia designers (is) one type of project-based learning, which requires students’ active participation, and engages them in authentic problem investigations“[6].

Audiovisual media integrate the functions of vision and hearing, which means the sound and the picture and thus enhance the process of learning and memorizing [7]. Students are more effective if several senses and activities are deployed. Moreover, the interaction of the students and the digital learning material plays an important role in the process.

As the morphology of the term indicate, multimedia is eventually an amalgamation of numerous media and, when such materials are created, the skills that have to be taken into account are programming, graphic designing, producing multimedia elements, planning of projects, team working, critical thinking, problem-solving, interpretation and the visualization in an innovative way by means of using the contemporary information technologies. By taking part in a multimedia project, students “are engaged in a variety of activities from brainstorming, gathering and researching information, writing, creating art works, to programming and evaluating” [6]. McGrath et al. [8] say that the self-respect in students increased while they were taking part in implementation of a multimedia project.

The challenge to create a multimedia product for a target audience serves as the central curriculum activity to drive students to learn and solve problems along the way [9].

A. Multimedia-assited learning

The term *multimedia* is widely interpreted nowadays. It can be said though that the term is used within the context of the multiple meaning. However, it can be agreed on that multimedia involves perception, or, rather, receiving and processing information from one or more senses at a time. Various media through which the information is spread and in which the information exists are used for the purpose.

The multimedia elements are labelled as text, sound, graphics, video and animation. The core use of all the contemporary digital technologies in the teaching process is based on knowing the students' visual and listening skills. The interaction between a student and a digital teaching content plays a crucial role in the process. The multimedia-assisted learning has proved very effective, especially in teaching youngsters.

Applying these fundamental conditions to the context of multimedia learning, we can define interactivity as follows: Interactivity in the context of computer-based multimedia learning is a reciprocal activity between a learner and a multimedia learning system, in which the [re]action of the learner is dependent upon the [re]-action of the system and vice versa [10].

Multimedia content for the young are specially designed, enriched and interesting. No matter whether an adult or a child is implied, the process of the multimedia-assisted learning commences at the moment of reading a text, or at the beginning of watching a graphic image/animation and at the moment of listening to the speech, music or musical effects that motivate, direct, give advice, remind or provides a feedback. As its very name implies, multimedia is eventually an amalgamation of various media and, when such digital content is being created, what is to be taken into consideration is the skills of programming, graphic designing, planning, project implementation cooperation, organizing skills, critical thinking and problem solving, interpretation and visualisation in an innovative way and by means of using the contemporary information technologies.

The use of the definition made multimedia fit perfectly into the making of the DVD and brought the DVD onto a higher level of the quality of being interesting and practical. The primary objective was to provide for an abundance of the educative quality and the educative content as prescribed by the educational wireframe. The mission of attaining the complementarity of multimedia and educative qualities was thus completed.

B. A development environment

For the purpose of making the Ivica i Marica DVD and the follow up documentation, the following was used: Adobe Flash CS3 Professional, Adobe Premiere Pro CS3, Adobe Photoshop CS3, Adobe After Effects CS3, Adobe Illustrator CS3, Microsoft Word 2007, Camtasia Studio).

-Start up menu:

At the very beginning, i.e. after the introductory film has been shown, the start up menu appears (Picture 1.). The start up menu was made in the way that provides for a practical and simple use. The main elements of the menu are rooms. When the user places the mouse pointer on a certain object, the lateral panel-window opens offering short, important information on the form of pollution (Picture 2.). This principle of work facilitated the navigation and the acquisition of the content.

Picture 1. A section of the main scene - menu

Picture 2. Panels in the game

C. The process of making main menu and submenus

The process of making a multimedia DVD starts from drafts and the drafts are the basis for any future multimedia project development. The drafts as well include a simultaneous work on the purpose of the interactive DVD because the purpose of the DVD directly influences the design, which is the following step in the development of the DVD. The innovative quality is desirable in every sense because it provides for an interested user, which is the objective of making a DVD. It is as well very important to take care of the choice of colours as the colours are chosen in accordance with the content and the purpose of a DVD. This project was implemented with the objective of a multimedia influence on users.

The educative multimedia DVD was created in order for children to get to know more and in an entertaining way about pollutions and activities of the people aimed at protecting the environment by means of the waste recycling. (Picture 3.)

The interactive educative DVD puts multimedia and e-learning in the focus. The use of the information and communication technologies on children develops the active use of the cognitive processes such as the memory, the perception, problem solving, the cause-effect relationships and the independent thinking and the adaptation with the specific attitudes [11].

Picture 3. Entertainment (a quiz)

-Design book:

Once the theme and the content were defined, the Design Book was being written. The quality of the final DVD depends directly on the quality of the way the Book is written. Moreover, the book contained all information and instructions that are necessary for the graphic designer, the programmer and the animator in order to complete their tasks successfully. After the elements were designed in the Design Book, the respective tasks were assigned to the team members.

-Graphic Design:

The graphic designer played important role in designing the user interface that has to be functional, dynamic, understandable and interactive. Besides the role related to the user interface, the graphic designer had as well a role in developing the textures that were being used in the making of the DVD. The graphic designer used the following software: Adobe Photoshop CS3, Adobe Illustrator CS3 and Adobe Flash Professional.

Picture 4 shows the DVD cover.

Picture 4. DVD cover

-Programmer:

The task of the programmer was to coordinate all the graphic design concepts with the software solutions. The programmer used ActionScript as the basic software system that made the application work properly. The programmer as well showed the respective skills in the use of Adobe Flash CS3 Professional software.

-Animator:

The job of the animator was a particularly hard one. Namely, the animator had to comply with the requests of both the graphic designer and the programmer. The animator contributed to making the DVD by means of Adobe Flash CS3 Professional software.

-Video design:

The task of the video designer was to coordinate the multimedia and the content requests with functionality and the educative quality. He made his ideas true with a help from the following software: Adobe After Effects CS3, Adobe Photoshop CS3, Adobe Premiere Professional CS3.

-Audio design:

The whole audio system that in turn depended on the audio designer is of a crucial importance for the quality of the DVD. The audio designer creates, chooses and coordinates the audio background with the requests of the application.

IV. PLATFORMS, TECHNOLOGIES OF THE MAKING THE DVD AND WEB SITE

All the projects on which the team had worked before were intended primarily for the Windows platform. The main development environment was Adobe Flash. Within the DreamLand studio, there is an active website (http://www.dreamland.com.hr). The updates related to the work are posted on a regular basis and there the photo/pictorial content can be browsed.

V. RESEARCH METHODS

The purpose of this research was to define the extent of the quantity of the multimedia content and effects the secondary school-aged youngster can acquire within the curriculum framework. A targeted affirmation of the acquired knowledge is focused at the recognition of the activities that can contribute to enhancing the teaching process in the future working tasks/projects.

The questionnaire consisting of 8 questions was used on the sample of 116 students attending Web Designer and Media Technician subcourses within the Multimedia course. The students that took part in the survey were from all grades (1-4) of the course. The students expressed their opinions by means of Likert scale where they marked as 1 – strongly disagree with the statement, 2 – disagree with the statement, 3 – have not got explicit opinion on the matter and 5 – agree fully with the statement.

VI. RESEARCH RESULTS

The opinions of the students after they had taken part in the activities of the project of making a multimedia interactive DVD entitled Ivica i Marica:

I think that….	
…I have mastered a satisfactory level of knowledge and skills of creating multimedia applications.	
Mean	Std. Dev
3.896	1.211
…I have mastered a satisfactory level of practical …Skills required for getting a job in the production sector of economy.	
4.466	1.002
…I have mastered the use of (Adobe® CS4 Photoshop, Adobe® Flash CS4, Adobe® Premiere Pro® CS4 v4.0, Adobe® After Effects® CS4 v9.0. Cinema 4D®, 3D Studio Max etc.), the applications similar to those used in the production sector of economy.	
4.649	0.099
…I can successfully cooperate in a team work.	
3.991	0.995
…my knowledge and skills level in the area of planning and organizing resources (time, space, material, equipment) needed for fulfilling activities are satisfactory.	
4.557	0.775
…I can successfully take part in all the stages of planning and developing of a multimedia application.	
4.221	0.418
… I do have the skills of identifying and applying the relevant knowledge to solve challenging multimedia tasks	
3.339	1.128
…I do have the skills for a creative solving of multimedia problems.	
3.456	0.844

The results showed a high averages in students as to the knowledge and skills in the realm of multimedia (Mean = 3.896, Std.Dev = 1.211). The same goes for their perception of the possibilities of getting a job in the production sector of economy (Mean = 4.466, Std.Dev = 1.002). As to the use of the multimedia tools: Adobe® CS4 Photoshop, Adobe® Flash CS4, Adobe® Premiere Pro® CS4 v4.0, Adobe® After Effects® CS4 v9.0. Cinema 4D®, 3D Studio Max, the averages are even higher (Mean = 4.649, Std.Dev = 0.0994). Needless to say, all other parameters showed as well that the students appreciated the project activities more when particular economy or their future occupations tasks were implied.

VII. CONCLUSION

The primary objective of the educative, interactive and multimedia DVD is education through playing. By means of the multimedia DVD, the process of acquisition the knowledge about waste, waste fumes, pollution etc was facilitated. As a multimedia, entertaining and useful application, the Ivica i Marica DVD is intended for pre-school aged children and the children attending 1-4 grades of primary school and thus provides for effective learning through playing. Besides the educative part, there are as well lots of entertaining content.

By means of interactive learning, the students:

- acquire knowledge and cognitive skills
- can argument their attitudes and opinions
- ask questions
- actively listen to others
- come to the conclusion that the result belong to everyone, not to individuals only
- develop the sense of personal responsibility for attaining the goals of the group

It can thus be concluded that the level of self-reliance as to the acquired skills and knowledge was raised through the project activities. More than a decade ago, researchers have suggested that ‘‘as teachers have progressed from the use of blackboards and chalk, to overhead transparencies and computer-aided presentations, and now to multimedia, more research is needed to help guide the use of these tools to enhance learning” [12].

The Dreamland team recommends this application whose goal is to entertain and educate its users.

REFERENCES

[1] M.-I. Shahrimin, and D.-M. Butterworth, Young children's collaborative interactions in a multimedia computer environment, *The Internet and Higher Education*, *Volume 4, Issues 3–4, 2001*, pp. *203-215.*

[2] D. Vogel, and J. Klassen, Technology-supported learning status, issues and trend, Journal of Computer Assisted Learning, 17, 2001, pp. 104–114.

[3] H.-K. Sun, and P.-C. Cheng, The design of instructional multimedia in e-Learning: A Media Richness Theory-based approach *Computers & Education*, *Volume 49*, *Issue 3*, *2007*, pp. *662-676.*

[4] K. Dalacosta, M. Kamariotaki-Paparrigopoulou, J.-A. Palyvos, and N. Spyrellis, Multimedia application with animated cartoons for teaching science in elementary education, *Computers & Education*, *Volume 52*, *Issue 4*, *May 2009*, pp. *741-748.*

[5] A. Antonietti, and M. Giorgetti, Teachers' beliefs about learning from multimedia, *Computers in Human Behavior*, *Volume 22*, *Issue 2*, *March 2006*, pp. *267-282.*

[6] Y. Liu M, and Y. Hsiao, Middle School Students as Multimedia Designers: A Project-Based Learning Approach, Journal of Interactive Learning Research, 13(4), 2001, pp. 311-337.

[7] M. Cindrić, D. Miljković, And V. Strugar, Didaktika i kurikulum, IEP-D2, Zagreb, 2010.

[8] D. McGrath, C. Cumaranatunge, H. Ji Misook Chen, W. Broce, and K. Wright, Multimedia Science Projects: Seven Case Studies, Journal of Research on Computing in Education, 30(1), 1997, pp. 18-37.

[9] Y.-T. Sung, K.-E. Chang, and M.-D. Lee, Designing multimedia next term games for previous termyoungnext term children's taxonomic concept development, *Computers & Education*, 2008, pp. 1037-1051.

[10] S. Domagk, R.-N. Schwartz, and J.-L. Plass, Interactivity in multimedia learning: An integrated model, *Computers in Human Behavior*, *Volume 26*, *Issue 5*, *2010*, pp. *1024-1033.*

[11] N. Cevher-Kalburan, Ö. Yurt, and E. Ömeroğlu, The use of interactive CD-ROM in early childhood education: Teachers' thoughts and practices, *Procedia Computer Science*, *Volume 3*, *2011*, pp. *1555-1561.*

[12] T. Li-Ping Tang, and, M.-J. Austin, Students' perceptions of teaching technologies, application of technologies, and academic performance, *Computers & Education*, *Volume 53*, *Issue 4*, *2009*, pp. *1241-1255.*

ICT learning experience in the multilingual and multicultural environment of ERASMUS IP TAARLoG project

Janko Žufić, Hrvoje Cvilinder
Juraj Dobrila University of Pula, Department of Educational Science, Pula, Croatia
jzufic@unipu.hr
hrvojecvilinder@gmail.com

Abstract: This work describes the ICT learning experience during intensive TAARLoG (Traditional in the Alpe Adria Region – Local for the global) Erasmus project. Project was implemented intensively in Pula in October 2012 with the participation of 54 students from the universities of Pula, Trieste and Maribor. In the multilingual environment (Slovenian, Italian, Croatian and English languages) students were tasked to learn, among other matters: a) to record with a video camera; b) to process and edit recorded audio and video material with the program application Adobe Premier CS4. Feedback received from students, in the end of the project, reflected their satisfaction (on the scale from 1 to 5 where 1 stands for '*not satisfied*' and 5 stands for '*very satisfied*') as following: academic achievement: 4,47; personal achievement: 4,71; utilization of equipment: 4,34; abilities and experience of professors: 4,49; total quality of tutoring: 4,36; results of the adopted knowledge: 4,09; total mark for the project: 4,57. The applied teaching method was the *Project teaching*.

Key words: project, Erasmus (IP), project lecturing

I. INTRODUCTION

University „Juraj Dobrila" in Pula hosted the project „Traditional in the Alpe Adria Region – Local of the Global – TAARLoG" within the framework of the Erasmus Intensive Program, Lifelong Learning Program, Education and Culture DG, from October 7th to 19th, 2012. The project goal was to teach nonmaterial cultural heritage in the Alpe-Adria region and specifically in Istria. Traditional dances, traditional instruments and traditional songs from Istria were researched and taught. The entire project was documented by video camera recording.

In total, 53 students participated in the project IP: 10 from the University in Trieste, 19 from the University in Maribor and 24 from the University „Juraj Dobrila", Pula. Students were divided in three groups: Group for singing and dancing – 35 participants, group for playing instruments – 8 participants and ICT group (Information and communication technology) – 10 participants. The main focus of this article will be the teaching process and activities of the ICT group.

ICT group focused on two main outcomes: I) To gain experience in creating multimedia materials suitable for education in the child-care centers and first few grades of elementary schools, and II) To master the basic rules and skills in creating multimedia material, individually or as a part of a team. There were several subtasks within the second outcome: a) to introduce and master all functions of a video camera; b) to prepare and position a video camera for recording; c) to use the tripod; d) to use lighting in open and close spaces; e) to learn framing; f) to learn separate audio recording on a video camera; g) to transmit recorded material to a computer; h) to learn postproduction and framing; i) to record a reportage; j) to record a cultural performance; k) to record an interview; l) to record a short film; m) to write media coverage; n) to prepare final materials for web posting; o) to create a web site and post materials on the web sites; p) to learn basic diction… Teaching was based on the idea that the goal of integration of students and technology was not to expose students to the technology, as the technology in the teaching process was primarily a tool for the realization of the teaching process [1].

Other, equally important goals for Erasmus IP and TAARLoG project were the mobility of students, acquirement of new competencies for the labor market and nurture of "culture of diversity". In this project, the "*Culture of diversity*" implied respect for the cultural origin of people and battle against xenophobia (hatred towards foreigners) and inclusion of young people of different cultural and national identities that work together on achieving common goals – projects, tolerance, solidarity, self-respect, respect and intercultural understanding.

Before the project started, the teacher – tutor thought which approach to the teaching process to take. He didn't want to use traditional types of teaching because the final outcome would not be entirely accomplished. In the end, he opted for the method of demonstration and a segment from *The Theory of Constructivism* that focuses on students that are actively involved in the teaching process and come to their ideas through the individual experience. Path to learning is unique for each student, as they construct their own knowledge with an individual effort based on their own experience. In a manner of the Constructivism, efficient teaching is founded on the importance of the existing knowledge, immediate experience in reality, experience of the whole, reflectivity of learning and teaching, effective creativity, inner motivation, role of a teacher and the method of integrity of teaching [2].

Authors want to thank the National Agency for mobility and EU programs and the EU for the co-financing of this project.

Additionally, some of the characteristics of the temporary teaching are learning based on discovery and self-initiative, solving problems individually, divergent and creative thinking [3]. De Zan is emphasizing teaching models such as problem-oriented teaching, teaching based on problem solving, creative teaching, research-oriented teaching, project teaching etc. The common thread to all these models is the existence of the problem that needs to be resolved.

The Project Method that the teacher chose, some authors cite as the specific type of learning by discovery [4] and it is considered as the most complex method of learning by experience as it implies work of several groups of students on a complex problem that is situated in a real life environment [5]. Teacher needs to put an extra effort due to possible difficulties and uncertain outcome of the project, in spite of that, the method itself is very interesting and motivating for the students. Teacher plays the role of a project manager that supervises students during the entire process, while students work in a team environment, assign and complete their part of the tasks in order to successfully complete the project [6]. Dewey and Kilpatrick developed the project method between 1930 and 1940. Dewey wanted to form a school where students would research, create and experiment. School that is formally dividing educational contents into study groups, lessons, themes, facts and assignments is not consistent with the needs of students for a coherent learning. As an alternative, Dewey offered the idea of a school-laboratory that would stimulate students to work individually in workshops and laboratories and in natural environments. Students learn by reflecting on problems and trying to resolve them. Emphasis is put on the importance of solving problems that exist in the community and mutual assistance and integration in the community, while educational activity needs to be led by instinctive and impulsive activities of students. Organization of individual tendencies and activities of students are a prerequisite for their integration in the community, with an emphasis put on the collective consciousness and cooperative work. Kilpatrick continued to develop Dewey's concept and developed the project method that is based on the students working in line with their affinities and in cooperation with the teacher work in teams on specific projects. The project starts with the definition of a problem and presentation of possible solutions, it continues with planning and implementation of activities that lead to problem solving and ends with presentation of conclusions and their practical realization thus enhancing equal development of individual abilities and socialization as a result of a team work. Teacher organizes activities and advises thus enables students to be more independent and inventive in problem solving [7].

Using the project method in teaching involves a student-centered partnership between a student and a teacher, an openness to the problem' situations and tasks from the real life, correlation with other areas of science and human activities, methods of cooperation and team-work, development of organizational and communication abilities of students, application of new teaching methods.

Phases of the project teaching are: preliminary phase, planning, implementation, project presentation and valorization. Each phase consists of several steps [8].

The preliminary phase consists of: a) selection of a project theme – it has to be interesting, inadequately researched, from the area of student's affinities; b) definition of project tasks or goals, keeping final goals feasible.

Planning consists of: a) Preparation of students for the selected theme – directing them to the literature or showing them preliminary steps that would help them in the realization of the project and b) development of the project plan implementation that will include: choosing place and duration of the project, duration of each phase, selection of material and method of work, forming groups of students and assigning work within the groups, anticipating problems and their solutions.

Implementation consists of following the plan under teacher's close supervision.

Project presentation – in the end, results of all groups need to be integrated and implementation documented and presented via video recordings, posters, web sites or public lectures.

Valorization consists of: a) evaluation of the project's goal realization and b) critical analysis of the entire work on the project – its problems, successes etc.

A. Role of a student in the project tutoring

Project tutoring teaches students to learn individually and to be individually responsible. Students in most parts organize their learning processes. The problem is that schools in general do not provide these competencies to students and existing conditions in schools, such as large study groups, defined time frames, limited financial abilities and teachers, obstruct development of individual responsibility in students. These factors make the process of qualifying students for learning via project teaching that much harder [9].

Learning process in project teaching includes:

- definition of distinct goals
- formation of working groups
- decision on time frames and adhering to them
- work assignment and its coordination
- team-work
- constructive conflict management
- implementation of all activities to the end
- presentation of individual results to other coworkers in the team
- control and evaluation of one's own work and other team-mates work

It is quite common that students during teaching process do not know how to assume assigned roles. In spite of that, they have to do it. Project teaching is designed for the stimulation of new type of learning and development of new abilities.

Graphic presentation of the educational effects of the project teaching with the students at its center is presented in the figure 1, and it was derived from the modifications of the effects cited in [11 and 12]

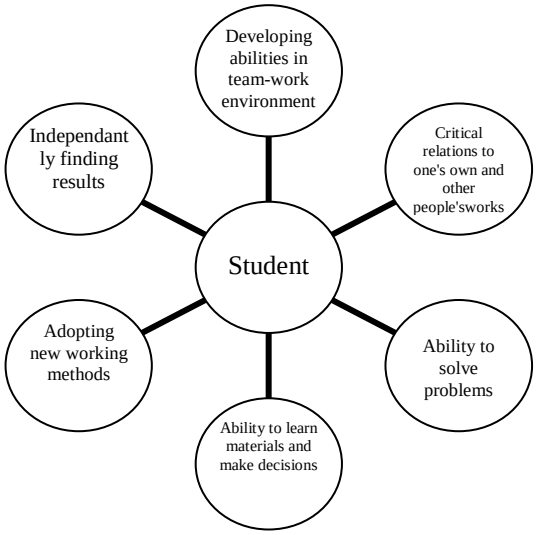


Figure 1. Educational outcomes of the project teaching

B. Role of the teacher in the project teaching

Teachers are assuming roles that outgrow the traditional roles. Depending on the project phases, roles will be changed in alternation. Teachers need to become familiar with the roles and help students understand what roles they assumed [9]. Roles of the teachers are:

- *Coordinators*: depending on the age group and project experience of students, teachers will at the beginning observe passively development of the project and only if necessary, coordinate the project. In this role, teachers are establishing external contacts and simplifying the path to these contacts for the students.
- *Moderators*: Teachers observe the course of learning and make sure that all participants in the projects have appropriate conditions for the participation in the learning process. Teachers as moderators will not have any impact on the formulation of the process, but they can open the free space and give wider prospective.
- *Coaches, advisors, assistances*: Teachers are always ready to provide directions and offer a helping hand in case those activities come to the dead-end or if problems arise.
- *Conflict managers*: Teachers need to mediate in situations when conflicts within the working groups arise, but also in the situations that involve groups and persons that are not part of the school environment. It is necessary to develop several scenarios and methods to bring parties to the conflict to a favorable solution. To prepare for this role, teachers might need to have additional qualifications.
- *Professionals*: In each project and in individual situations, teachers need to present their professional knowledge.
- *Teachers are learning too*: Project themes, in their complexities, sometimes outgrow the teacher's spectra of knowledge. Therefore, teachers on the project often acquire important learning experiences. Presentation of the role of a person who permanently learns is something that needs to be mentioned to the students in order to help them experience learning as the continues process.

Role of teachers in project teaching is presented in figure 2, and it was derived from the modifications of the role of teachers cited in [9 and 13]

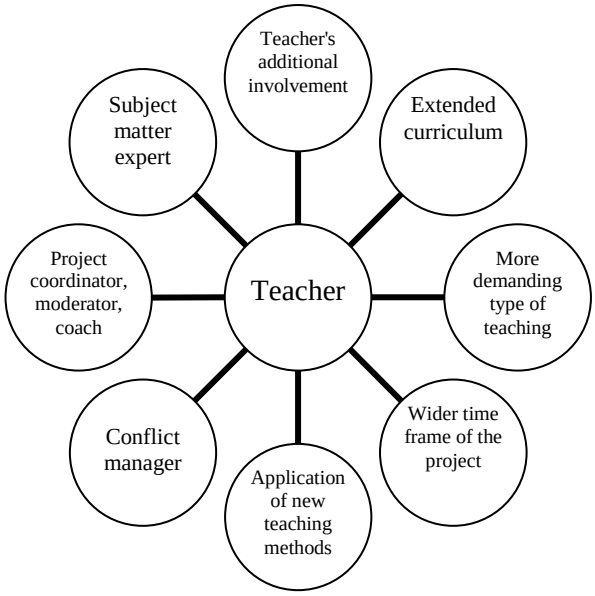


Figure 2. Role of teachers in the project teaching

II. PROJECT METHOD IN THE ACTIVITIES OF THE ICT GROUP WITHIN THE TAARLOG PROJECT

Phases in the project teaching in the activities of the ICT group were: preliminary phase, planning, implementation, project presentation and valorization.

A. Preliminary phase

Preliminary phase consisted of selecting goals, subtasks and selection of subtasks for the realization: students were introduced to the two main outcomes and subtasks. Some subtasks all participants needed to master, such as: introduction to the functions of a camera, learning how to prepare and position a video camera, learning to work with a tripod and lighting and basic formatting. For other tasks, students formed groups. An instruction was given that larger groups should consist of students from different universities. Working languages were English, Croatian, Italian and Slovenian. Table 1 presents subtasks/working tasks, number of students and their institutions, duration (in days) and task schedules

B. Planning

Planning consisted of: a) familiarization of students with the selected theme – each student/group of students were shown how to use the equipment, but students needed to use them creatively in line with their own initiative. For example, recording team consisted of two members. Students had to decide how to position the camera, choose exposition and frames they would shoot etc; b) writing a project implementation plan that would define place and duration of each phase of the project. Some of the project tasks required selection of equipment and work methods, for example for the recording cultural performance a plan on the number and positioning of the camera, selection of frames for each camera etc needed to be planed in advance. Forming groups was completed in the preliminary phase and work division in the groups was based on decisions who will start working and who will complete each element of the tasks. For example, group for recording interviews consisted of two students. One student started to interview while the second student recorded it with a camera and then they switched roles. This phase included also anticipation of problems that might arise and ways to resolve them. Examples: during panoramic shooting of the city, students had to go to the top of a building and the entrance door to the roof was locked, some students and teachers, participants in the TAARLoG project, refused to give interviews to the camera, while recording a short film, student-actor had to be persuaded to do it repeatedly etc.

C. Implementation

Implementation phase consisted of following the plan. After all work tasks were assigned and students introduced to the basics of their solutions, realization started with an active involvement of teachers that divided their attention equally to each group as the tie permitted. As this was an Intensive program, work day lasted for the entire day, with lunch breaks. That was the only way to learn and accomplish the outcomes due to the project methods and large number of tasks. Some elements needed to be repeated several times.

D. Presentation of the project results

In the end of the working day or at the beginning of the following day, each group presented results of their accomplishments. Students commented and elaborated on their work and offered arguments for their decisions and after that other students and the teacher would offer their comments. Some groups exchanged their work assignments. For example, group for recording and postproduction switched their groups in order to learn as much as possible and be involved in more than one phase of the material preparation.

Large part of the materials was not processed during the IP and it was left to be completed after the IP. In the end, it was concluded that too many tasks were assigned for the time dedicated for the project and number of students involved.

Most of the material was posted on the web sites and DVD media (over six hours of recorded and edited material) that were after the end of IP copied and disseminated to all participants to the project. As for the editing, initially was planned to publish all materials in Croatian, Italian, Slovenian and English languages. Later on, we had to abandon that idea and materials were published mostly in English language.

TABLE 1. SUBTASKS/WORK TASKS AND DIVISION BY PARTICIPANTS, DURATION AND SCHEDULE OF PERFORMANCE

Subtasks/project tasks	Participants[1]										Time [days]	Days of intensive program (IP)										Days after IP
	H1	H2	H3	I1	I2	I3	S1	S2	S3	S4		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	
Introduction and working with all functions of a video camera	x	x	x	x	x	x	x	x	x	x	3	x	x	x								
Preparation and positioning of video camera for recording	x	x	x	x	x	x	x	x	x	x	3	x	x	x								
Use of a tripod	x	x	x	x	x	x	x	x	x	x	3	x	x	x								
Use of lighting in open and close spaces	x	x	x	x	x	x	x	x	x	x	3	x	x	x								
Framing	x	x	x	x	x	x	x	x	x	x	3	x	x	x								
Separate audio recording on a video camera							x	x			2									x	x	
Transmission of recorded material to computers			x	x			x				9		x	x	x	x	x	x	x	x	x	
Postproduction of recorded material and framing	x	x						x			7				x	x	x	x	x	x	x	x
Recording and postproduction of a short film	x	x									3							x	x	x		
Recording a reportage			x	x	x	x				x	8			x	x	x	x	x	x	x	x	
Recording a cultural performance			x								2	X									x	
Recording interviews				x	x	x	x				4				x	x		x	x			
Writing media coverage	x				x				x		4							x	x	x	x	
Preparation (conversion) of recorded materials for web posting										x	5						x		x	x	x	x
Creation of web sites and posting materials on such								x			10									x	x	x
Basic diction	x					x		x			4						x	x	x	x		

[1] "H" marks students from Croatia, "I" from Italy and "S" from Slovenia. In this case, H1, H2 and H3 signify three students from Croatia.

E. Valorization

Valorization consisted of a) evaluation of the accomplishment of the project outcomes and b) critical analysis of the work on the entire project – difficulties, successes etc. Evaluation of the accomplishment of the project outcomes was carried out via a standard questionnaire provided by the National Agency. Table 2 presents questions and results of the students relevant for each ICT group. Critical analysis was conducted during the 4th phase and after the implemented IP when students could see the total results and they were very satisfied with their achievements. Students were quite successful and they were satisfied with their progress. Leader of the ICT group was also very satisfied with creative ideas and results that students achieved.

An important fact needs to be mentioned and that is that none of the students had any preexisting knowledge of any element used in their tasks (apart from basic computer knowledge: MS office, web and e-mail) and the accomplishment is so much greater when compared to their starting level.

III. RESEARCH

A. Goal of the research

Goal of this research was to obtain results relative to the satisfaction of students – participants in the Erasmus IP TAARLoG ICT group with the implementation of the project.

B. Participants in the research

Participants in this research were 53 students from the Universities in Trieste (10), Maribor (19) and Pula (24). They were divided in three groups: a) for singing and dancing; b) for playing instruments and c) for ICT. ICT group consisted of 10 students: 4 from Maribor, 3 from Trieste and 3 from Pula. As for the gender and age, there were nine females and one male of 21 to 23 years of age.

C. Methodology

For this research was utilized a standard questionnaire provided by the Agency for mobility and EU programs [10], and it was completed on the last day of the Erasmus IP project by students, participants to the project. Questionnaire was anonymous, but its completion was mandatory. Time was not limited and they had to grade answers from 1 to 5, where 1 stood for "Not at all" and 5 for "Very much". Data from the questionnaire are stored and processed in MS Office Excel 2003.

Questions relevant to the ICT group and received feedback are presented in Table 2. It also presents results for ICT groups and all participants in TAARLoG project.

IV. RESULTS AND DISCUSSION

Questions relevant to the ICT group and received feedback are presented in Table 2. It also presents results for ICT groups and all participants in TAARLoG project.

Students gave high marks. Standard deviations are rather insignificant and it indicates that there were not significant aberrations.

The fact that this was the first program implemented at the University of Pula and that project leads had no previous experiences in working with students with diverse language groups, mixed levels of academic achievements (different study groups and years of study), these marks are very high. During 10 days they worked intensively and we received few complains on the intensity of the program. This complains were anticipated and at times even justified, but after the final public performance, they were all satisfied. ICT group was satisfied with the achieved results and other students with the good performance.

TABLE 2. RESULTS AFTER THE CONDUCTED RESEARCH

Questions	All participants in the project			
	Average mark		Standard deviation	
1. How satisfied were you with the duration of the IP?	4,47		0,70	
2.Which were the factors which motivated you to participate?				
a) Academic	4,16		0,93	
b) Cultural	4,40		0,76	
c) Practice of foreign language	4,20		0,93	
d) Career plans	3,90		1,20	
e) European experience	4,37		0,91	
	All participants in the project		ICT group	
	Average mark	Standard deviation	Average mark	Standard deviation
3. Judgement of academic/learning outcomes of the IP:	4,47	0,73	4,44	0,68
4. Judgement of personal outcomes of the IP	4,71	0,53	4,55	0,68
5. How satisfied were you with the academic activities and the pedagogical aspects of the IP in terms of the following aspects?				
a) The number of hours taught	3,94	1,00	4,22	1,03
b) The equipment used	4,34	0,66	4,67	0,47
c) The capabilities and expertise of the professors	4,49	0,74	4,88	0,42
d) The overall quality of teaching	4,36	0,73	4,56	0,50
e) The expected learning outcomes	4,09	1,11	4,44	0,68
f) The activities besides the general course	4,47	0,74	4,44	0,83
6. Do you think participation in the IP will help you in your further studies/career?	4,60	0,67	4,22	0,91
7. Do you think participation in the IP will help you in finding a job?	3,80	1,28	3,56	1,64
8. Overall evaluation of the IP	4,57	0,67	4,56	0,50

Achieved results as well as the audience reflected a positive feedback to project leads on each outcome. ICT group received full feedback after materials were posted on the web site and after reception of over 6 hours of recorded materials stored on DVD media.

In addition to students who have self-evaluated their satisfaction with the project and audience that evaluated public performance, ICT team leader evaluated each task presented in Table 1. Table 3. presents tasks, number of students that participated in each task, an average mark received for each task and standard deviation of the received marks. Some of the tasks have not been included in the table as they were not completed during the course of the IP, but at a later date, for example: a) creation of the web site and posting of materials on the web sites ; b) conversion of completed materials into media (web and DVD) and c) finalizing DVDs containing materials. It has all been done after the project.

TABLE 3. ACHIEVED RESULTS

Subtasks/project tasks	Participants	Average mark	stdev
Introduction and working with all functions of a video camera	10	4,80	0,42
Preparation and positioning of video camera for recording	10	4,90	0,32
Use of a tripod	10	4,80	0,42
Use of lighting in open and close spaces	10	4,30	0,82
Framing	10	4,30	0,82
Separate audio recording on a video camera	2	5,00	0,00
Transmission of recorded material to computers	3	5,00	0,00
Postproduction of recorded material and framing	3	4,33	0,58
Recording and postproduction of a short film	2	4,00	0,00
Recording a reportage	5	4,00	0,71
Recording a cultural performance	1	4,00	n/a
Recording interviews	4	4,25	0,50
Writing media coverage	3	4,33	0,58
Basic diction	3	3,67	0,58

The reason for the untimely completion of all tasks is the overestimated possibilities of realization. In spite of diligent work of both students and leads, number of tasks was just too great. That was also one of the very few of students' objections. One of the factor was also the fact that none of the students, participants to the project, had any preexisting knowledge of the use of video technology and basic knowledge of the use of information technology.

Marks given by the teacher indicate very good results. Such results have not been achieved during regular teaching activities carried out at the University "Juraj Dobrila", Pula, by the same teacher. Some of the reasons are: a) teaching method – project lecturing; b) twice as many classes than during regular teaching; c) better motivation of the participants; d) joint and close work of students and leads (from 7-9 hours per day); e) cooperation between teams and team members.

Another positive outcome was that as the project was progressing, all 10 students were gaining more self-confidence. They have also exhibited great level of creativity in finding solutions to the given tasks.

Some of the observed difficulties were as following: a) more time consuming than during regular classes; b) permanent small communication issues due to the language barriers; c) students expected more free time and at times failed to give an adequate contribution to achieving results; d) student's inexperience in team-work, therefore more time was needed for adaptation.

V. CONCLUSION

This study presents experiences from the implemented the international Erasmus IP TAARLoG project. It described primarily experiences of the ICT group (information and communication technology) in its approach and method of teaching. Both students and leads are satisfied with the implemented project. The chosen approach of the project teaching might be time consuming, but accomplished results bring a long lasting quality. One of the most important positive result was the achieved international experience and higher self-confidence in working.

REFERENCES

[1] Docerman, D. (1998.) "Great teaching in one computer classroom" Tom Snyder Production Inc, Watertown

[2] Podravec, D. (2003.) "Radost učenja, integrirana i projektna nastava u osnovnoj školi" Osnovna škola Franje Viktora Šignjara, Virje

[3] De Zan, I. (2005.) "Metodika nastave prirode i društva" Školska knjiga, Zagreb

[4] Bognar, L., Matijević, M. (2002.) "Didaktika" Školska knjiga, Zagreb

[5] Varošanec, S. (2006.). Učenje otkrivanjem. http://web.math.hr/nastava/mnm1/ucenje.doc (viewed 13.2.2013.)

[6] Bjelanović Dijanić, Ž. (2012) Računalo u istraživačkom radu učenika u nastavi matematike, Napredak, 153 (2) 203-218)

[7] "Metodika i komunikacija e-obrazovanja - Pedagogija" http://www.carnet.hr/referalni/obrazovni/mkod/pedagogija.html viewed 13.2.2013.

[8] Bjelanović Dijanić, Ž., Tkalec, M. (2010.) "Fotostrip u integriranoj i projektnoj nastavi hrvatskog jezika, francuskog jezika i informatike u srednjoj školi" Bjelovarski učitelj 15 (1-2), pp. 38-45.

[9] Auchmann, M., Bauer, L., Doppelbauer, A., Hölzl, E., Winkler, S. (2002.) "Projektna nastava – prijedlozi za primjenu" K-education, Sarajevo

[10] Erasmus Intensive Programme (2012.) "Final Report Form for Students", Lifelong Learning Programme, Education and Culture DG - Obrazac završnog izvješća za studente - minimalne stavke http://www.mobilnost.hr/index.php?id=467 (viewed 02.01.2013.)

[11] Garmaz, J. (2006.) "Izazovi HNOS-a: projektna nastava u vjeronauku", Crkva u svijetu, 41,2, 166-192

[12] Jensen, E. (2003.) "Super nastava, *Nastavne strategije za kvalitetnu školu i uspješno učenje"*, Zagreb, Educa

[13] Matijević, M., Radovanović, D. (2011.) "Nastava usmjerena na učenika", Zagreb, Školska knjiga

Incorporating social network workshops in elementary school curriculum

Martina Draženović
Ban Josip Jelačić elementary school, Zagreb, Croatia
martina.drazenovic@gmail.com

Abstract - The main purpose of elementary school is to prepare students to be educated, critical thinking individuals who can provide and help their community. Community is now elevated on the virtual level and specific rules of conduct apply. School can influence student's behavior while online. Every informatics teacher can help their community by making workshops for students, their parents and colleagues about social networks. The goal is to raise awareness, educate and prevent online deviant behavior in our community. Those workshops can make a difference not only for our students, but also for their parents and colleagues as well.

I. INTRODUCTION

Social networks have become an important part of life in a way that they enable dual living, both virtual and real life. As a teacher, it has come to my attention that some students can develop deviant behavior while using social networks. This behavior manifests in cursing, insulting other individuals, creating hate groups and befriending strangers, and as a result consequences are visible in real world. While exchanging information with colleagues and students, I have started developing an idea of organizing social network workshops for each of these groups of people separately, including their parents.

II. SOCIAL NETWORK WORKSHOPS

A. About workshops

The general idea is to create workshops which would educate students about social networks and which would be implemented as a part of school curriculum. There would be three kinds of workshops: for students, their parents and teachers. Each of the workshops would consist of five parts: Introduction to social networking, Benefits of using social networking, Dangers of social networking, Laws (which apply while an individual uses Internet) and Netiquette. The reason why there is not only one workshop (for students) is that we all, as a community, must be familiar with the situation which today's students are experiencing.

B. Coverage of social networks in elementary school informatics textbooks

In the informatics textbook for fifth grade you can not find any information about social networking, but only about rules of conduct while online. [1] In the sixth grade informatics textbook isn't any information related with social networking or rules of conduct. [2] In the seventh grade informatics textbook there also isn't any information addressing social networking. [3] In the eighth grade informatics textbook you can find the critical approach toward Internet browsing, but nothing about social networking. [4]

C. Workshops for students

On the basis of research regarding coverage of social networks in elementary school informatics textbooks and elementary school students using social networking [5], I concluded that informatics class in elementary school should incorporate social networking as a part of the curriculum. The students are aware of social networking and are using them to connect with their friends and people they know, but unfortunately, some of them are not familiar with risks and rules of conduct which apply while online. To raise their awareness about this situation, I have created a series of workshops which would cover most of the information relevant for a proper use of social networking. Workshops are divided into five parts:

1. "Introduction to social networking" would cover the general information about the Internet and the history of social networking.

2. "Benefits of using social networking" would emphasize the positive sides of social networking such as connecting with people we know, easing the organization of numerous events, information distribution and possibilities of business opportunities.

3. "Dangers of social networking" would address the possible problems that person can experience. This is the most important part of the workshop because it identifies numerous problems and deviant behavior that can occur while online, such as : open profile negligence, identity theft, befriending strangers, impersonation, intimidation, spying, abuse of person's photos, cursing and insulting individuals or groups of people, obsessive collecting of "likes" (while on Facebook), sexting, "liking" groups and communities that encourage and approve hate speeches towards individuals or groups of people (while on Facebook) and potential problems with employment. The main point of this part of the workshop is that virtual behavior has its consequences both in real and virtual world.

4. As a consequence of deviant behavior or a simple lack of information, there are laws which apply and they would be covered in the fourth part of the workshop called "Laws (which apply while an individual uses

Internet)". Those laws are divided into two groups; criminal [6] and misdemeanor [7].

5. The last part of the workshop that would be covered is "Netiquette". Netiquette is a set of social conventions that facilitate interaction over networks, ranging from Usenet and mailing lists to blogs and forums [8]. Specific rules of conduct apply both while online and real world and students have to be aware of those.

D. *Workshops for parents*

Parents should be, if not yet, familiar with social networking and should also be aware of benefits and consequences. The reason why to include student's parents in these workshops is that school can cover only so much and family must keep student's internet behavior in control. I have held 19 parent–teacher conferences in my school regarding this subject during which I have learned that many of my student's parents do not possess elemental knowledge about this subject. They were surprised that criminal laws can apply on their children as well as them (if the child is under 14 years of age) if children's conduct while online should consist of deviant behavior previously mentioned. Some of them are under the impression that whatever happens online stays online, without any consequences in the real world and that must be changed.

E. *Workshops for colleagues*

Colleagues can exploit benefits of social networking in a way that they could improve their approach in teaching a certain subject. They could eliminate the boundary standing between them and their students and make their classes modern and more interesting. They should also be familiar with all pros and cons regarding social networking in order to improve their class.

F. *Pilot*

I am currently holding pilot workshops for students in my school. They include students in seventh and eighth grade (total of 67 students are involved). Students have shown positive reactions and genuine interest regarding this subject. We incorporated these workshops in our curriculum as a part of the chapter about Internet, chapter about creating web pages, chapter regarding presentations and class project. The result will come in a form student knowledge of social networking, web page regarding specific social networks (Facebook for the eighth grade) and web page regarding dangers of the Internet (for seventh grade), as well as presentations which would appear in the official school website and posters displayed in school.

G. *Collaboration within community*

If a certain problem should appear, we should be able to sort it out as quickly as we can and we can't do it without the support of our community. This no longer solely includes the school they attend, but also the community they live in. In order to provide quality response, education is crucial. Information is easy to reach if a person knows where to reach it which makes it hard and unacceptable for the people with no elementary knowledge about informatics. That should change for all students, their parents and teachers so that we all can collaborate and extend our support for the individual. The point is a fast and efficient transfer of information in order to improve our community.

H. *Expected results*

I am hoping to successfully incorporate social network workshops in curriculum as well as to educate and raise awareness regarding social networking and also to ease access of the information regardless of one's knowledge about informatics. The first results of the pilot workshops held with students should be visible during this school year.

III. CONLUSION

Social network workshops should make a difference not only for students in understanding and improving their conduct and knowledge while using social networks, but also their parents and my colleagues, teachers. They should be a different, thorough and interesting way of approaching and addressing this topic in school as a part of curriculum.

REFERENCES

[1] http://www.scribd.com/full/28177389?access_key=key-jeuryk4ch5dq7gpiwy0

[2] http://www.scribd.com/full/28177591?access_key=key-27vyfd01wbw3xgbr7gt2

[3] http://www.scribd.com/full/28179488?access_key=key-2m7oos3nqan2murcg0vv

[4] http://www.scribd.com/full/28178029?access_key=key-yrxapdz0zdmr1chhogj

[5] The role of Facebook in elementary school education, Draženović, Martina, Osnovna škola bana Josipa Jelačića, Zagreb, Hrvatska, 2012.

[6] http://www.zakon.hr/z/98/Kazneni-zakon

[7] http://www.zakon.hr/z/52/Prekr%C5%A1ajni-zakon

[8] http://en.wikipedia.org/wiki/Netiquette

[9] http://www.mup.hr/83255.aspx

[10] http://www.cert.hr/

[11] Informatika 8 : udžbenik iz informatike za 8. razred s multimedijskim CD-om / Fredi Glavan, Alfa, 2011.

[12] Informatika 7 : udžbenik iz informatike za 7. razred s multimedijskim CD-om / Ida Srdić, Goran Kadić, Branko Hrpka, Alfa, 2008.

[13] Informatika 6 : udžbenik iz informatike za 6. razred s multimedijskim CD-om / Ida Srdić, Branko Hrpka, Goran Kadić, Alfa, 2011.

[14] Informatika 5 : udžbenik iz informatike za 5. razred s multimedijskim CD-om / Ida Srdić, Alfa, 2011.

[15] http://www.slideshare.net/vhrzica/privatnost-na-drustvenim-mrezama

[16] http://www.slideshare.net/vhrzica/national-cert-azoo

[17] http://www.slideshare.net/vhrzica/digitalni-tragovi

[18] http://www.slideshare.net/vhrzica/uloga-policije-u-zatiti-djece-na-internetu

[19] http://www.slideshare.net/vhrzica/internet-novo-igralite-azoo2013

[20] http://www.slideshare.net/PogledKrozProzor/sigurni-internet-prevencija-nasilja-nad-djecom-na-internetu

[21] http://www.slideshare.net/unicefceecis/booklet-unicef-final21

[22] http://www.slideshare.net/PogledKrozProzor/predavanje-o-sigurnijem-internetu-za-djecu-i-mlade2012

[23] http://www.slideshare.net/vhrzica/mup-8868918

Mathematical Modelling, Computers and GeoGebra in University and College Mathematics Education

T. Sekulić*, Đ. Takači **

* Technical College of Applied Studies, Zrenjanin, Serbia

** Faculty of Natural Sciences and Mathematics/Department of Mathematics and Informatics, Novi Sad, Serbia

tsekulicvts@gmail.com

Abstract - In this paper we present a way of introducing innovation in the teaching process by applying new teaching methods that are based on the principles of active learning and mathematical modelling, empowered by computers and mathematical software GeoGebra.

The paper deals with experiences from teaching process, concerning both aspects: teachers and students point of view. Step by step procedure of making mathematical model is explained – from starting preparation, to implementing GeoGebra in mathematics education.

Direct application of new teaching methods based on principles of active learning, and mathematical modelling are illustrated in the paper by examples processed by using mathematical modeling as teaching method and GeoGebra as software tool.

I. INTRODUCTION

The process of mathematical education in our country and worldwide, is currently suffering many changes. Changes reflects in inplementation and application of new teaching methods in mathematics education at all levels, from elementary school up to university. In addition, the new teaching methods are expected to stimulate students to see and experience mathematics differently, through practical application, and recognizing it in everyday real life.

Mathematical modelling has a history as well as mathematics itself, but it can not be said for the history of the application of mathematical modelling in mathematics education. Currently are active many researches related to the application of mathematical modelling in the classroom, concerning students performances and training teachers to use the new teaching methods.

The need for mathematical modelling in mathematics education is mostly reflected in the fact that students coming out of schools and colleges with a relatively high level of theoretical knowledge, but at the same time they are not adequately prepared to apply this knowledge in practice. Nowadays, business and economic interests of employers require that students are prepared and trained to solve specific problems, which means that they must be ready to adapt on the new situations and required tasks.

This paper was designed to give the complete overview of the mathematical modelling, and in particular its application in the mathematics education through concrete examples realized with the help of computers and their application in higher mathematics.

All examples are implemented using the software package GeoGebra 4 [11].

II. MATHEMATICAL MODELLING

A. *The Concept of Mathematical Modelling*

The mathematical model is a mapping between reality and mathematics. The purpose of the mathematical model is reflected in drawing conclusions about reality based on the model. The crucial thing for any mathematical model is that it has to lead to something, meaning that mathematical model is the basis for drawing conclusions about reality which can then be tested experimentally.

There are different types of mathematical models:

• Normative models, most notably the models that are related to income, the methods of choice, ...

• Descriptive models, which are further divided into models that predict (such as weather forecast), models that explain (for example, model of the rainbow, explaines why we can see it), and models that describe (for example, the development cycle of a virus).

Mathematical modelling is the concept about which researches are still leading very lively debates. This term has a very different interpretation, and because of its

importance and application in the mathematics education, it is suggested that these interpretations join together, in order to make a starting point for any researcher.

The central point in mathematics education should be creation of students ability to apply mathematics in a various real life situations.

Mathematical modelling is connected to one of the main goals of education, which is related to the understanding of real-world phenomena and their connection with mathematics. In any case, the use of mathematical modelling in mathematics education is inevitable if we are to fulfill the ultimate goal - to train students to recognize the real-world situations where different mathematical theories can be applied.

In [2] *Blomhøj* refers at three major reasons for the integration of mathematical modelling in the mathematics education, based on his research results:

(1) Mathematical modelling bridges the gap between students real-world experiences and mathematics. Besides, it motivates students to learn mathematics, provides direct cognitive support for the creation of new concepts and creates mathematical culture in terms of describing and understanding the real life situation.

(2) The development of high-tech society, demands competencies that relate to the setting, analysis and criticism of mathematical models, and also competencies concerning social aspects of development in terms of adequately educated workforce.

(3) Mathematical modelling at different levels of complexity plays an important role in the forming and functioning of society based on high-tech development. Imperative to maintain such a society is the development of individuals and experts who will be able to interpret and use the results of modelling.

Currently, mathematical modelling is not yet sufficiently present in mathematics education, but intensive efforts are being made in adapting the curriculum and application of these teaching methods. The training and education of mathematics teachers for the application of mathematical modelling in practice is also one of the future goals for promoting mathematical modelling in education.

B. *Mathematical modelling process*

Establishing a relationship between mathematics and the real world is actually a process of a mathematical modelling. Modelling process is not complicated and there are different approaches and points of view concerning the modelling phases. Especially, particular attention should be given to the aim of mathematical modelling process, because this determines the approach to modelling, modelling cycle selection and modelling competencies that such action later should acquire.

Depending on the author, there are several different cycles of mathematical modelling. Mathematical modelling cycle, used as a base of research that will be presented in this paper was first presented by *Gloria Stillman and Peter Galbraith (2007)* [4], [10].

Since this mathematical modelling cycle is present in teaching practice, special attention has been devoted to the transitions between phases and potential students blockages that could arise in the modelling process. Fig. 1. shows the mathematical modelling cycle, used as a base of this research.

This mathematical modelling cycle consists of several phases, which are: selection of **real-world situation** that will be modelled, based on the selected real situation the **real world problem** is observed, **mathematical model** based on the real-world problem is made and **mathematical solution** is derived from mathematical model. If the **solution is accepted**, we make a **report** about it. If not, we can do **revision** of the modeling process, and mathematical model itself.

Particularly important in the process of mathematical modelling are also transition between phases. Transitions in this modelling cycle are as follows:

1. Understanding, structuring, simplifying, interpreting context, understanding the problem, simplifying the interpretation, ...
2. The mathematical description of the model, ie. mathematical model is obtained by adjusting the real situation to mathematical content using the assumptions, formulas.
3. Mathematical solution is obtained by mathematical operations.
4. The interpretation of the mathematical results includes analysis of the results and their interpretation in the context of real problems.
5. Communication, verification (if the model is satisfactory).
6. Repeating of modelling process (if the model is not satisfactory).

Fig. 1 illustrates the mathematical modelling cycle that was used in the study, as well as all its phases and transitions.

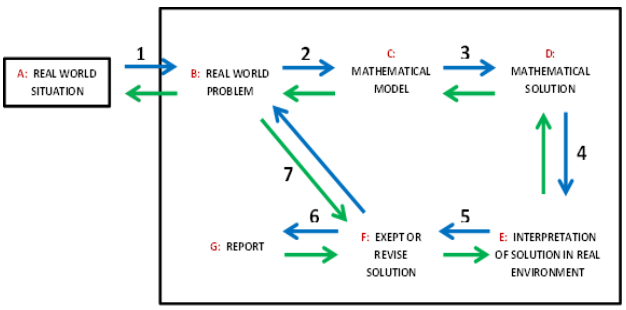


Figure 1. Stillman-Galbraith mathematical modeling cycle

This modelling cycle has been widely accepted for use in mathematics education, because of flexible transitions between modelling phases and extensive opportunities for discussion, publication of assumptions and conclusions that they provide, which is really essential in the mathematics education, if the goal is to get students actively involved in the learning process.

III. THE MATHEMATICAL MODELLING IN PRACTICE

A. *The organisation and realisation of research*

Respecting the basic principles of mathematical modelling that are listed in the previous section, in the next section of this paper will be given a description of the research related to the application of mathematical modelling in higher mathematics education.

The main goal of the research was to investigate the influence of mathematical modelling application on the formation of the students basic concepts of the calculus at the College of Applied Studies. Therefore, it was necessary to conduct an experiment, which consisted of two parts. In the first part we designed an effective example which students used to model the basic terms regarding the monotonicity of function of a real variable, and in the second part of the experiment we examined the effects of mathematical modelling application in mathematics education at the college level.

The research was conducted within the college first year Mathematics course, with weekly teaching hours of (3 +3). With the students of the control group were used conventional teaching methods for teaching lectures on functions monotonicity, while students of the experimental group were taught the same teaching unit by mathematical modelling. For that purpose, we designed an interesting situation, which students used as real-life situation in mathematical modelling process. By modelling this situation students came to the basic concepts of function monotonicity.

By its very nature, the differential calculus is an abstract, fairly complicated and demanding theory. Because of the presence of highly abstract concepts, students find it difficult to understand the true meaning of the calculus, and especially to see a little further, i.e. to recognize the situations where this theory can be applied. That is why in the learning process are often present questions as: "*Why are we doing this now?*", "*Where will we use this?*" and "*Do you ever use this in real life?*". Such questions are direct indication that teaching process is not effective and not well designed. The need for expending horizons for new ways to present the calculus, especially, function monotonicity were imposed. This new ways included teaching methods which will be colorfull, interesting and fun for students. The lecture should be based on a completely different principles than the classic, so that students have the impression that they are doing something fun and interesting, and not to deal with some abstract concepts. In doing so, the course of the lecture is designed to meet all the basic principles of mathematical modelling process and to follow all the steps of this process.

For the purposes of this study we used the Internet resources [12] which contains photos, illustrations, charts, and its interesting story illustrates the process of forming the function monotonicity concept in a very interesting manner. Presenting the basic informations related to the function monotonicity in this way allows students easier to understand and faster to adopt these concepts.

Lecture with the experimental group of students was held following the stages of the mathematical modelling process , using cycle-*Stillman Galbraith*.

B. *The experimental lecture*

First stage – real world situation*:* Students are first introduced to an interesting story, which may not be an example of a real life, but it is very descriptive and can be identified and implemented with a range of situations.

The story is based on the famous cartoon where the main character, Willie E. Coyote is trying to catch the Road Runner. He is crossing the wire holding the anvil in his hands. Due to the weight of the anvil, the wire stretches, lowering the Coyote on the ground. He at that point drops the anvil, and since he is on the wire, he was launched into the air. For a period of time he is in the air flying up and then reaches a maximum height and begins to fall down. Finaly, he falls to the ground. [13]

After the students have heard the story, they tried to interpret it and to set the problem based on the understanding of the story. Solving the problem will further lead to the mathematical model representing a given situation.

Observing the movement of the cartoon hero, students noted some changes in his movement, meaning that at some point his height is increasing, and decreasing in the other. They also identified moments when the cartoon character reached the maximum and minimum height.

Second stage – real world problem statement.: Thru active discussions, students recognized model within cartoon events. Students very fast noticed that the movement of the cartoon character can be represented as a function and tried to express it mathematically. This was the most appropriate moment for implementing computers in modelling process. GeoGebra was used as additional help for connecting real life situations with mathematics.

Third stage – mathematical model: Students made conclusions based on the presented situation, and described situation formally, using first derivation. Mathematical model was realised using GeoGebra. With the help of GeoGebra we created one simulation, which consisted of three elements.

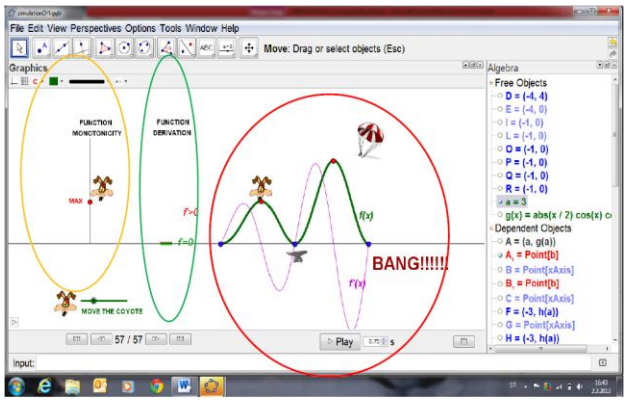


Figure 2. GeoGebra animation of real situation

The first element marked in red color represents the function graph and its derivation. The animation shows the path of the cartoon hero, represented mathematicaly by function graph. This is a very effective and obvious way to associate the real situation with the mathematical content.

Yellow colour marked second element of GeoGebra animation used for representing changing of function values. Specifically, this element of animation shows how function changes it's values, by using image of cartoon hero moving up and down in line with the changing value of the function. This section was inserted in an animation to illustrate the increase or decrease in function values. We used this part of the animation to explain students the connection between real situation (coyote is moving), with mathematics (function graph).

The third section, which is circled in green shows the sign of the functions first derivative. All three sections are connected by the same variable, so that at any time students can connect position of the cartoon hero with function monotonicity and with the sign of the functions first derivative. It was deliberately done in this way, in order to point out to the students the connection between functions first derivative and it's monotonicity.

Fourth stage – mathematical solution: Students configured rules and steps necessary for examming and finding function extrema. After calculation, the students have returned to the cartoon story and tried to make conclusions and interpret the results, in order to perceive the mapping between the formal results and the actual real situation.

In this stage of the mathematical modelling process, GeoGebra also has played a vital role. Specifically, the animation not only reflects the real situation on the mathematical model, but it can be manipulated, which is the primary purpose of mathematical modeling – to make model whose parameters can be changed, in order to simulate changes in real situations.

Without the use of computers in mathematical modelling process it would not be possible to manipulate with the model and especially to perceive the real situation. In particular, by using GeoGebra is possible to "launch" the model. In this way the maximum effects are achieved in terms of understanding the concepts that students have encountered in the process of modeling.

Fifth stage – real world meaning of solution: Students again compared obtained results and compared it with cartoon situations**.** In addition, they worked thru some examples, to confirm their results.

Students used GeoGebra for processing new examples because they also had a better view of the situation and they manipulated the mathematical model with less effort, instead they did it manually.

Sixth stage – revise model or accept solution: Students talked together thru mathematical model**.** To verify the model, students are given a few simple tasks, which consisted of testing monotonicity of functions, using the conclusions from the previous stages of mathematical modelling process.

IV. GEOGEBRA AS A TOOL FOR THE MATHEMATICAL MODELLING

Mathematical modelling process without the use of computers can not achieve it's maximum effects. This is because the model does describe the real situation, but the manipulation of the model without the use of computers is not enough descriptive and effective. With use of computers mathematical modelling process get much richer, and the results obtained by this method in mathematics education are extremely positive. Most notably is the positive reactions of students on using mathematical modelling with computers, because it is known that the new generation of students are favorable towards the to new technologies, particularly computers. Also, this kind of mathematics education organization, with the help of mathematical modelling and computers, helps mathematics education to run according to the principles of active learning. In this way, students are constantly activated during the learning process and come to their own conclusions, and that implies significantly enhanced effects of learning.

GeoGebra is chosen in this study as a software package that will accompany the process of mathematical modelling for several reasons. First, GeoGebra is a free software package, powered by Jawa, easy to install on the Windows platform. GeoGebra is also very easy to use and both, teachers and students do not need any special training for its application. Mathematical modelling is already quite a demanding process, speaking in terms of teacher training and in terms of its application, so that simplicity of use GeoGebra comes very nicely. However, although simple to implement, GeoGebra is still very powerful mathematical software package which can be used for implementing and ilustrating any mathematical theory, no matter how demanding this theory is.

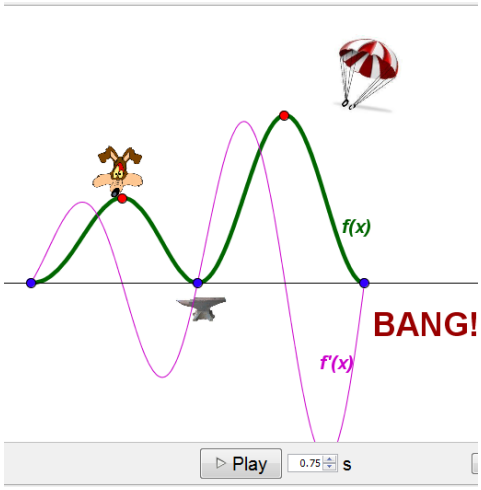


Figure 3. Function and it's derivation

Thus, the GeoGebra can be used for representation of function graph (Fig. 3), and there are also options allowing to display functions first derivative as was used in the first segment of the animation of the mathematical model described in the previous section.

GeoGebra's other options that were used in this study are related to inserting images and text. Images in GeoGebra can be further animated, and it was used in the study for animation of the cartoon hero picture, resulting picturing the whole situation with mathematical terminology.

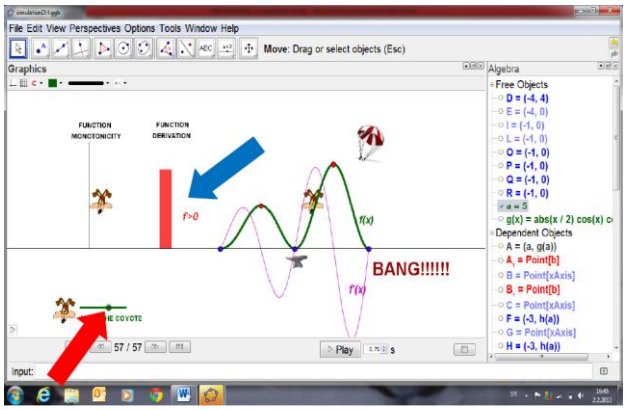


Figure 4. Manipulation with the slider realised in GeoGebra

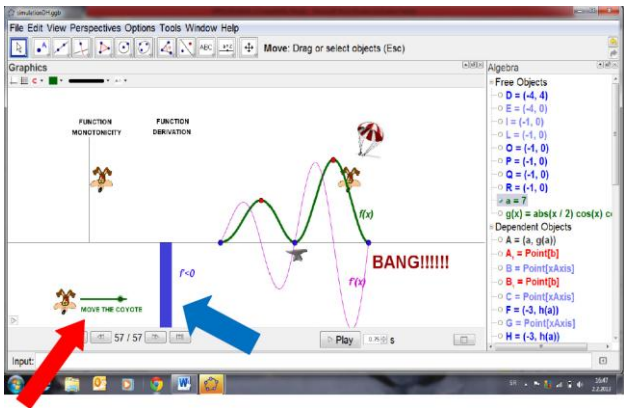


Figure 5. Manipulation with the model realised in GeoGebra

The key point in this study has found an GeoGebra option called slider. This option gives the user the ability to set the interval and the steps in this interval by which the value of a variable is taken. It is possible to launch slider and animate various structures in GeoGebra, and that is the one of GeoGebra's main advantages as seen from the mathematical modelling point of view.

On the figures above we can see the changes caused by moving the slider in the bottom left corner of the screen. So, when the first derivative is positive, the function is increasing (Fig. 4) and when the function first derivative is negative function is decreasing (Fig. 5).

In this case, the slider is present in each section of the GeoGebra animation. In the first section slider is used to run the cartoon character walking on the function graph. In the second section slider starts cartoon character image that goes up and down, and illustrates the rises and falls of the function values. Finally, in the third section, slider creates polygons depending on the sign of the function first derivative, and by that draws a connection between the sign of the function first derivative and its monotonicity.

The key point in this model is precisely the possibility of GeoGebra to create structures such are slider, function, functions first derivative, animation of pictures, etc., because in this way created mathematical model is evoked.

GeoGebra really provides great opportunities for the application in the mathematics teaching practice, at all levels of education, and in the present study, only some of its real capabilities are shown.

V. CONCLUSIONS

„…an important aspect of applied mathematics which is emphasised by all groups is mathematical modelling, that is, the modelling of real situatons in mathematical terms. This aspect of applied mathematics is vital in many areas, but unfortunately often overlooked in university applied mathematics."

Mc Lone Report (1973)[1]

Mathematical modelling can undoubtedly be widely used in mathematics education. The greatest contribution of this method is reflected precisely in linking the real world with the formal world of mathematics, because it opens up great opportunities for the application of mathematics in real situations.

On the other hand, mathematical modelling allows simulation of real situations outside of laboratory conditions, which greatly simplifies the process of research.

Implementation of computers in the mathematical modelling process makes this method much more efficient and enables easier manipulation with mathematical models.

Studies have shown that the effects of this method in mathematics education are extremely positive.

On the other hand, GeoGebra has proved to be a very convenient tool for the application and implementation of computers in mathematics education. It should be emphasized the ease of working with this software package and positive impressions of students and teachers who have used it.

The future of mathematical modelling method is definitely in education, and together with GeoGebra it will undoubtedly occupy an important place in the education of the young experts whose priority will be applicable knowledge and finding appropriate solutions for problems from the real environment.

REFERENCES

[1] Baumslag, B., Fundamentals of Teaching Mathematics at University Level, World Scientific Publishing Company, 2000.

[2] Blomhøj, M., "Different perspectives on mathematical modelling in educational research – categorising the TSG21 papers", Paper presented at the 11th International Congress of Mathematics Education, Monterrey, Mexico, 2008.

[3] Blum, W., Galbraith, P., Henn, H.-W., & Niss, M. (eds.), "Modelling and Applications in mathematics education", The 14th ICMI study, (The new ICMI Study Series),Vol. 10, New York, Springer, 2007.

[4] Galbraith, P., Stillman, G., "A Framework for Identifying Student Blockages during Transitions in the Modelling Process", ZDM, Vol. 38(2), 143-162, 2006.

[5] Goos, M., Stillman, G., Vale, C., Teaching Secondary School Mathematics - Research and Practice for the 21 st Century, Allen&Unwin, 2007.

[6] Henn, H., W., "Modelling in school - chances and obstacles", The Montana Mathematics Enthusiast, Monograph 3, pp. 125-138, 2007.

[7] Kaiser, G., Modelling and Modelling Compentencies in School", Mathematical Modelling (ICTMA 12) Education, Engineering and Economics, 510 pp, Chichester: Horwood, 2007.

[8] Kaiser, G., Maaß, K., "Modelling in Lower Mathematics Secondary Classroom - Problems and Opportunities", In Blum, W., Galbraith, P., Henn, H.-W. & Niss M.(eds.), Modelling and Applications in Mathematics Education, New York, Springer Academics (former Kluwer Academics), 2006.

[9] Sekulić, T., "The role of active learning and mathematical modelling in modern mathematics education", Computer algebra and dynamics systems in mathematics education-CADGME 2012, International conference, University of Novi Sad, June 2012., Novi Sad, Serbia

[10] Stillman, G., Galbraith, P., Brown, J., Edwards, I., "A Framework for Success in Implementing Mathematical Modelling in the Secondary Classroom", Mathematics: Essential Research, Essential Practice. Volumes 1 and 2, Proceedings of the 30th Annual Conference of the Mathematics Education Research Group of Australasia, 2007.

[11] www.geogebra.org

[12] www.karlscalculus.org

[13] Wile E. Coyote and The Road Runner – Beep Beep, Merrie Melodies Series, Warner Bros., 1952.

Application of Gantt Charts in the Educational Process

Andrea Bednjanec,
School of Electrical Engineering (Elektrotehnička škola), Zagreb, Croatia
andrea.bednjanec@gmail.com

Martina Filipović Tretinjak,
School of Electrical Engineering (Elektrotehnička škola), Zagreb, Croatia
martinafilipovic.tretinjak@gmail.com

Abstract - Gantt charts, a project management tool for planning and scheduling projects, can be used in educational processes. With the increasing duties, individuals are responsible to effectively manage time and complete tasks. An advantage of Gantt charts is their graphical overview. They visualize beginning dates, ending dates and the important duration parameters helping to track various projects over specific time periods. In the educational process there are many tasks which need to be planned such as school activities, school curriculum, competitions, school projects, intern´s activities, internship worksheets, teaching process, introducing curriculum fields, allocating time of the certain curriculum fields and subjects, defining efficient class strategies showing how many time will be spent doing various activities, guidance for lesson development, self-evaluation, etc.. Gantt chart's graphical representation is also very useful in the field of the teacher preparation. There are three components important in teacher preparation: teacher knowledge of the subject to be taught, knowledge and skill in how to teach and determining the time educational content will be introduced. Experience from the School of Electrical Engineering in Zagreb shows that use Gantt chart's in self-evaluation, intern´s activities planning, lesson planning and teacher preparing make teaching more effective.

I. INTRODUCTION

A table is the most common planning tool in the educational process. But they do not give a visual overview of all planned activities or events. Using them, it is impossible view the time of events that partially or completely disrupt planned tasks. In addition, tables and text files do not provide a visual connection ability nor the sequence of events which the pieces of the plan.

In the School of Electrical Engineering in Zagreb some teachers make researches on usage Gantt charts in the teacher preparation and planning the internship activities and teaching process. To facilitate the planning of the teaching process and ongoing monitoring and self-evaluation of our work and achieve higher levels of educational outcomes, we decided to improve planning some segments of our educational work with Gantt charts The result of this work is better preparation of teachers and their ability to easier organize and maintain the educational process. That results in greater satisfaction in everyday teacher´s work and achievements greater competences of students and their satisfaction with the educational process.

Gantt charts visualize each distinct activity as a block of time. Their main advantage is graphical overview. Each block of time is labeled with a title to show the task and the amount of time that the block represents. Gantt Charts the time each activity is expected to take and the order in which activities should occur. Each plan needs to be detailed with large tasks broken down into smaller subtask what makes easier to keep track of how it is getting further.

They are used for planning the time overall project would take and where the pressure points in the project will occur. Gantt charts are helpful tool to see immediately what should have been achieved at any point in time. They are also help in troubleshooting of potential delays seeing how remedial action may bring the project back on course.

It was found that this way of preparation and planning segments of teaching process is more effective, and that usage of Gantt charts can be involved in more school activities, school curriculum, competitions, etc.

We explored on the internet if other teachers in the world have experience with the use of Gantt charts, but we haven´t found a lot of documents. We have found only a few examples. Staff of the Francis Marion University, South Carolina, USA, has made Gantt chart to provide a general timeline of activities designed to help their University to earn re-accreditation. Institute of Technical Education Singapore use Gantt charts for Time Management Schedule in Completion of Supervised Field Teaching of participants who acquire the Pedagogical Certificate in Technical Education. University of Canberra has organized work in groups defining and setting time frames using Gantt charts. Board of Regents of the University System of Georgia (developed by its GLISI, Georgia's Leadership Institute for School Improvement) in A Performance-based Learning Module

for Georgia's Educational Leaders has recommended Gant charts as a project management tool. The Trafalgar School at Downton, UK, uses Gantt charts to teach students how to learn effectively and to take responsibility for their own time management.

Gantt charts can be created using the many free tools from the Internet, and can be used templates in MS Office (MS Word, MS Excel). Gantt chart can even be made in pencil on paper, without the use of computers. Their advantage is a deeper analysis of the problems we are discussing and raising awareness of key points where problems may arise. Shift plan is easily resolved in electronic form, what improves prospects for its final execution. We usually use KS Project Planner 2011, a freeware tool for Windows XP to Windows 8, 32-bit and 64-bit. However, we are aware that at the moment this kind of planning is not suitable for teachers who are not well in computer science.

Application of Gantt charts in the educational process is discussed is the SWOT analysis given in Table 1.

Based on the SWOT analysis can be concluded that for a wider use of Gantt charts in the collective should be a critical mass of people with ICT skills and the will to engage themselves.

TABLE I. SWOT ANALYSIS OF THE GANTT CHARTS APPLICATION IN THE EDUCATIONAL PROCESS

	Strengths	**Weaknesses**
Internal	1. Graphical overview used in planning educational processes. 2. Simple way to monitor and evaluate progress through schedule status. 3. Easy way to manage deviations from the curriculum. 4. Easy to detect disadvantages inside the curriculum. 5. No software implementation costs.	1. Takes time to create Gantt chart. 2. Changes to the schedule require a redrawing of the chart. 3. Curriculum structure must be completed before the chart can be drawn.
	Opportunities	**Threats**
External	1. Good reputation as an innovative educational institution. 2. Creating an innovative working climate. 3. Provides efficient time management. 4. Enables effective distribution of activities.	1. There is no positive attitude in the school towards change. 2. Some staff members are reluctant to learn necessary IT skills. 3. Gantt chart should not deviate from the specified plans and programs.

II. EXPERIENCES WITH USE OF GANNT CHARTS IN PLANNING

Our research is based on experience in the planning internship activities in planning during the educational process, planning and scheduling of lessons, educational outcomes and student´s projects, and thus preparing teachers and classes constantly just evaluating our teaching.

A. Internship activities planning

One of the most important components in preparing new teachers is a successful internship experience. During internship program intern is involved in many tasks which need to be organized through different activities such as: participating in faculty and school committee meetings, attending professional development seminars**,** collaboration activities between mentor and intern, keeping an internship diary and taking part in various school and extracurricular activities.
Most activities are carried out simultaneously. In order to increase the efficiency of the internship and to stimulate the motivation of the intern, internship program needs to be carefully planned. Duration of each activity varies across the year and depends on the correlation between mentor and intern's curriculum.

The main purpose of Gantt chart is to notice and if it's possible to avoid overlapping of intern's tasks. It provides a simple way to monitor and evaluate intern's progress through activities that have been completed. Through Gantt chart intern can also learn self-evaluation which enables him to achieve high quality educational outcomes and become an active participant in his professional improvement.

Chart shown in Figure 1 is based on a common school calendar, in which there are two semesters. At the beginning of the internship, intern observes mentors' teaching to achieve basic competencies. Afterwards, mentor observes interns' teaching and in meetings gives post-observation feedback for the interns' lessons.

During entire internship, intern is taking notes what allow him to record and apply mentors' teaching methods.

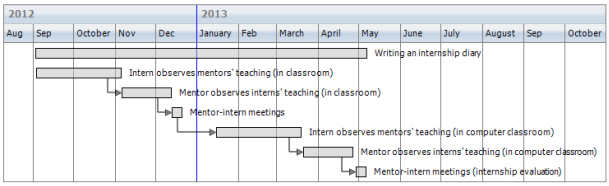


Figure 1. Internship activities planning

B. *Planning and scheduling of lessons*

Over the last 5 years Gantt diagrams are applied for schedule planning of lessons and lab exercises in the Electrical Measurements and in the Computer Networks. The goal of planning is to enable students to achieve theoretical learning outcomes required for the acquisition of competencies by performing laboratory exercises on time.

Figure 2 shows a Gantt chart schedule planning lessons for one class. Connection of instruments to measure current, voltage, resistance, frequency in the circuit is labeled as TEACHING UNIT 1, Connection of the wattmeter in a circuit - direct measurement of active power is labeled as TEACHING UNIT 2, Repeat of wattmeter is labeled as TEACHING UNIT 3, Semi-direct measurement of active power is labeled as TEACHING UNIT 3, and Measurement of active power by wattmeter in direct connection is labeled as LABORATORY EXERCISE. Scheduling depends on students´ timetable and is different for each class.

Figure 3 shows a Gantt chart schedule planning lessons and activities for two of third classes attending the same program (Computer Technicians). Each class is divided in two groups. Most of students from those two classes attends Cisco Networking Academy classes to and prepare from Certificates. They have Cisco lessons Saturdays, every two weeks. For that classes they have to be good prepared, reaching educational achievements on the Cisco Academy CCNA level. Gantt charts are helpful in scheduling teaching units, computer exercises, exercises on equipment, online tests, etc. Description of the router is labeled as TEACHING UNIT 1, Basic router configuration lab exercise is labeled as LAB EXERCISE 1, Construction of routing tables is labeled as

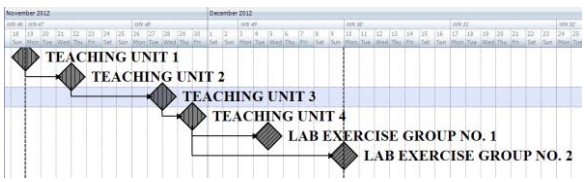


Figure 2. Electrical Measurements schedule planning

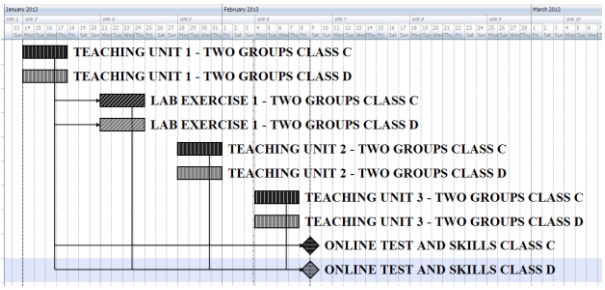


Figure 3. Gantt chart schedule planning lessons and activities in Computer Networks

TEACHING UNIT 2, Routing and switching function tables is labeled as TEACHING UNIT 3. Static routing is labeled as TEACHING UNIT 4. Cisco Networking Academy activities are labeled as ONLINE TESTS AND SKILLS.

In last 5 years research is made on 120 randomly selected students from third class attending the both, Computer Networks and Cisco Academy, CCNA program. Their lessons and educational achievements were planned using Gantt diagrams. Figure 4 shows results of researches.

C. *Students´project activities planning*

In the last two years students of the fourth class are involved in projects based on collaboration among Internet Technologies competences (Computer Technicians) and Politics and the Economy educational achievements Technicians for Electrical Machines with Applied Computer Science). During the school year they make three projects. Students involved with Politics and the Economy educational achievements order from students who are involved with Internet Technology skills a poster of imaginary political party, a promo-video for that party and a web site of an imaginary firm. They order the same thing in three classes and in each project choose one work as the best what meets their demands. Students are very engaged in those projects and work on them very likely.

Figure 5 shows Gantt chart of student´s project

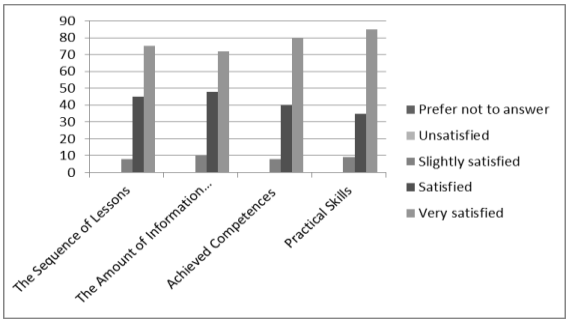


Figure 4. Research on satisfaction of students about teaching process in Computer Networks

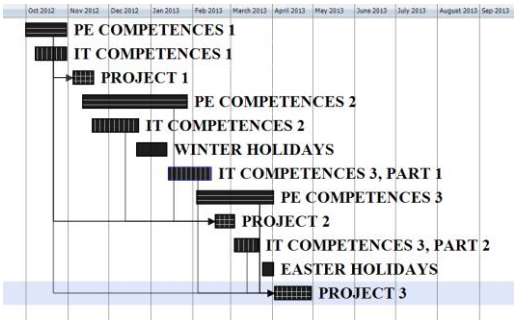


Figure 5. Gantt chart shedule planning student´s projects

activities schedule. PROJECT 1 is made after Computer Technicians achieve competences in editing of pictures (IT COMPETENCES 1) and Technicians for Electrical Machines with Applied Computer Science get competences to order appropriate pictures, text and color combination for an imaginary political party (PE COMPETENCES 1). After that Computer Technicians achieve competences in creating a video from pictures, and Technicians for Electrical Machines with Applied Computer Science get competences (PE COMPETENCES 2) to order appropriate promo-video for their imaginary parties. Computer Technicians in PROJECT 2 make videos (based on IT COMPETENCES 2) and must obey their demands doesn´t matter if their views are different from political party they work for. PROJECT 3 is made when Computer Technicians achieve competences in creating web site from a template (IT COMPETENCES 3) and Technicians for Electrical Machines with Applied Computer Science learn what is important for promotion of the business (PE COMPETENCES 3).

III. POSSIBILITIES OF GANTT CHARTS APPLICATION IN THE SCHOOL MANAGEMENT

In the school management there are many possibilities of Gantt charts application. The aim of that tool is to improve school curriculum planning, managing different activities and time. Two common examples are possibility of schedule teaching units based on correlations among the subjects, and planning of school activities as a part of the school curriculum. Involving of Gantt charts in the school management planning can be complicated in some schools because it demands a big engagement of different school stuff.

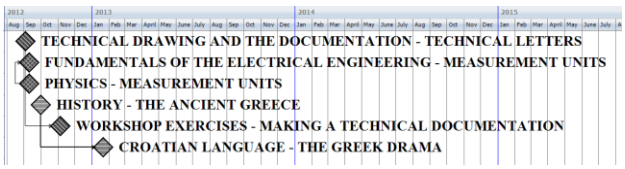


Figure 6. Gantt chart correlation among subjects schedule

A. *Gantt charts based on correlations among the subjects*

In the School of Electrical Engineering in Zagreb, for example, in the first class there are a lot of teaching units that correlate. Similar situation is in the upper classes, too. When a teaching stuff work as a team planning together the levels of educational achievements, results would be saving the time and reaching a higher levels student´s competences.

When is needed to set the beginning of an action or event with the completion of the previous one, is practically to display it using Milestone chart. Diagram shows only the events that must be brought into line. Figure 6 shows a Milestone Gantt chart based on beginning of several correlations in the first months of the school year. Those correlations are among: History (The Ancient Greece) and the Croatian Language (The Greek Drama), Fundamentals of the Electrical Engineering (Measurement Units) and Physics (Measurement Units), Technical Drawing and Documentation (Technical Letters) and Workshop Exercises (Making of Technical Documentation).

B. *Gantt charts based on planning the school curriculum activities*

Gantt chart tools enable permanent notes about the state of the each phase of the project. If an activity lasts a long time, it is possible to calculate what part of the activity is finished.

It is shown in the Figure 7. School curriculum activity IPA results implementation in the classroom activities – Career Guidance to Students lasts from the September to the June. It includes five Workshops and the one sub-activity for students, called Career Based Corner. The picture shows the situation in January 2013. Two of workshops are the part of that activity: Workshop Writing a CV, the application, simulation of talks with the employer is already finished, and workshop Self-employment will begin. Different labeling graphs can be distinguished actions that have been completed.

Figure 7. Gantt chart of the school curriculum activity: IPA results implementation in the classroom activities – Carrier Guidance to Students

IV. CONCLUSION

The Gantt charts application in the educational process improves planning and makes a teaching easier. This way of preparation and planning segments of teaching process is more effective. Besides a planning of teaching process and lessons scheduling usage of Gantt charts can be involved in more school activities, school curriculum, competitions, etc. Selection of the Gantt charts software depends on application of the chart and on user´s ICT skills.

REFERENCES

[1] http://www.gantt.com/

[2] Elektrotehnička škola, Zagreb, Konavoska 2: Školski kurikulum 2012/2013

[3] http://www.docstoc.com/docs/51927218/Gantt-Chart-for-Planning-Accreditation-Activities

[4] http://www.docstoc.com/docs/135594311/Academic-Year-Simple-Gantt

[5] http://edt.ite.edu.sg/eMentor/04_htm/sftmaterials_04/TimeMgt_04.doc

[6] http://edt.ite.edu.sg/eMentor/07_htm/sftmaterials_07/TimeMgt_07.doc

[7] http://www.canberra.edu.au/studyskills/learning/groups

[8] http://www.galeaders.org/sites/default/files/pbm/Leading%20a%20Team%20to%20Develop%20Project%20Management%20Tools.pdf

[9] http://www.trafalgarschool.com/

[10] http://www.kroll-software.de/products/projectplanner_e.asp

[11] http://www.businessdictionary.com/definition/Gantt-chart.html

[12] Elektrotehnička škola, Zagreb, Konavoska 2: Godišnji plan I program rada 2012/2013

[13] http://edt.ite.edu.sg/ActionResearch/index.htm

[14] http://www.easyfreeware.com/freeware/gantt_chart.html

[15] http://www. cisco.com

Interaktivnost u učionicama u kontekstu novih obrazovnih tehnologija i alata

K. Blažeka, Tehnička škola Čakovec, Čakovec, Hrvatska
kristinka.blazeka@ck.t-com.hr

Sažetak - Posljednje godine obilježene su sve većim učešćem obrazovnih tehnologija koje ohrabruju *online* i hibridno iskustvo učenja izvan granica određenih zidovima učionice (adaptivne platforme učenja, interaktivni i video-materijali za učenje, *online* stvaranje obrazovnih sadržaja, *Massive Online Open Courses (MOOCs)* itd.). Takva nova okolina učenja navodi nas na potrebu razmatranja ostvarene interaktivosti između učenika i nastavnika kao i učenika međusobno kad se oni fizički nalaze u učionici. Primjetno je da većina pukih prezentacija računalnih simulacija u stvari limitira vršnjačku konverzaciju. S druge strane, nastavnici koji potiču uvođenje novih digitalnih uređaja kao što su tableti i pametni telefoni uz pomoć odgovarajućih usluga i aplikacija, kroz tako realizirane virtualne aktivnosti i simulacije daju veći podstrek učeničkoj interakciji kroz dodatnu popratnu diskusiju, upite, debate i vršnjačku konverzaciju čime povećavaju njihovo razumijevanje koje je najučinkovitije upravo tada kad su zajedno u učionici. Određene nove tehnologije (3D projektori, *Xbox Kinect*...) također ohrabruju učenike sklone taktilnom i vizualnom učenju, a digitalni udžbenici omogućuju im dijeljenje bilježaka i oznaka koje vode ka diskusijama u skupinama. U radu se naznačuju, razmatraju i uspoređuju neke mogućnosti suvremenih digitalnih tehnologija, uređaja ili alata koje pridonose povećanju interaktivnosti u učionicama, a uz kvalitetnog nastavnika vidno poboljšavaju rezultate obrazovnog procesa.

I. UVOD

Suvremene tehnološke inovacije omogućile su transformaciju procesa učenja i poučavanja mijenjajući model distribucije i potrošnje informacija i znanja. Takve nove okolnosti zahtjevaju spremnost prihvaćanja novih načina medijske interakcije, preuzimanje pune kontrole nad načinima na koje konzumiramo sadržaje te spremnost pristupati sadržajima u bilo koje vrijeme i na bilo kojem mjestu. Razumljivo da sada te i slične vještine učenici moraju usvojiti tijekom svog obrazovanja.

Ovdje se ne radi o modelu koji škole mogu izabrati za prihvaćanje, ovo je model koji škole trebaju primjenjivati kako bi učenicima dale umijeće cjeloživotnog učenja i kritičkog promišljanja. U taj su proces najizravnije uključeni njihovi nastavnici. Kako bi ga uspješno proveli, edukacija nastavnika morala bi obuhvatiti usvajanje znanja o integraciji tehnologije u nastavu, ali i transformaciji metodologije poučavanja sa utjecajem na pedagogiju poučavanja. Na taj bi način nastavnici mogli usvojiti znanja o interaktivnoj pedagogiji koja promovira implementaciju učenja i poučavanja intenziviranjem upotrebe odgovarajućih digitalnih alata.

Mnoga istraživanja pokazuju kako općenito integriranje tehnologije u nastavu teži ostvariti promjenu sa na nastavnika usmjerenog učenja na okolinu učenja usmjerenu na učenika. U takvoj konstruktivističkoj učionici učenici teže više raditi timski, imaju više mogućnosti izbora i igraju aktivniju ulogu u svom učenju (Mize&Gibbons, 2000, Page, 2002, Waxman, Connell i Gray, 2002). Dodatna je pri tom dobrobit u tome što nastavnici mogu učinkovitije diferencirati nastavu ovisno o stilu učenja pojedinog učenika.

Zahvaljujući računalima, multimedijskoj interaktivnosti i sustavima za upravljanje učenjem (LMS) tako je naglasak stavljen na učenje, a ne poučavanje, pa je na učenicima puno veća odgovornost. S obzirom da se učenje većinom obavlja samostalno i uz pomoć interaktivnih medija, nastavnici se mogu usredotočiti na učenike koji trebaju pomoć.

Pri tom, interaktivna i multimedijalna obrada podupire trajnu manifestaciju zanimanja za sadržaje učenja, a učešće interaktivnih medija pobuđuje aktivnosti učenja. Učenici mogu aktivno i individualno objašnjavati sadržaje i sami ih obrađivati. Sve ovo vodi ka boljem razumijevanju i obradi informacija.

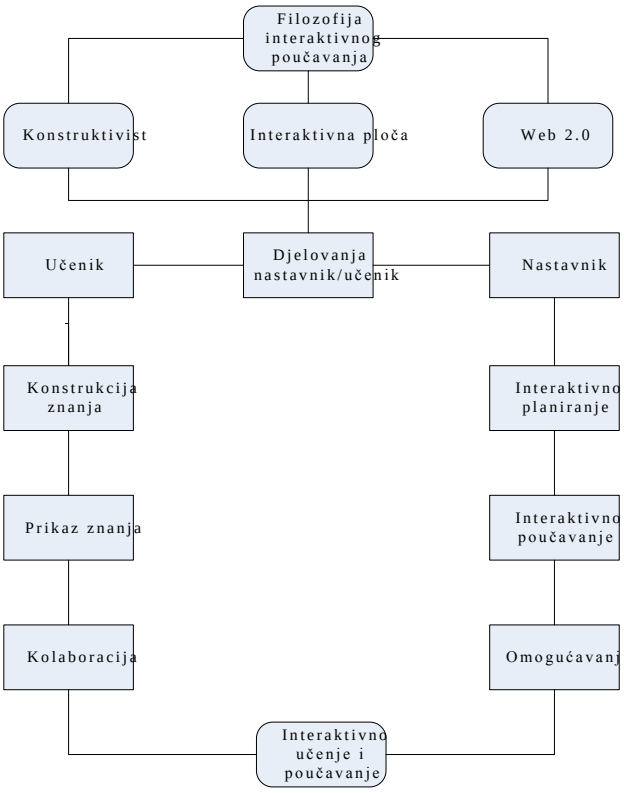
Takvo učenje podržavaju različiti alati, no evidentna je i opasnost da nastavnici neće uvijek te alate korisititi na odgovarajući način. U nedostatku odgovarajućeg obrazovanja i podrske događa se da nastavnici tehnologiju koriste kao potporu tradicionalnoj, na nastavnika orijentiranoj paradigmi koja ne uzima u obzir prednosti interaktivnih mogućnosti ovih tehnologija (npr. interaktivne ploče koja nije samo snažan obrazovni alat već i platforma za integraciju drugih tehnologija koje podupiru nastavu kao što su to web 2.0 alati pomoću kojih se pristupa interaktivnim na webu baziranim alatima za različite aktivnosti). To se događa kad nastavnici nisu uvježbani razmišljati o učenju i poučavanju kao interaktivnom procesu koji ohrabruje korištenje tehnologije kako bi se ostvarila interaktivnost u učionici [6]. Moguće je zato da tehnološke alate nastavnici koriste kao potporu tradicionalnoj pedagogiji poučavanja koja se ne temelji na interaktivnom modelu, umjesto interaktivnoj.

Zato bi svi odgovorni u obrazovnom sustavu morali biti svjesne načina na koji nove generacije učenika uče i vrše međusobnu interakciju, interakciju sa tehnologijom i sa životom općenito. Kako su tradicionalni modeli poučavanja vezani uz na nastavnika usmjereni pristup, a nova tehnologija potpora je potrebnoj i očekivanoj transformaciji na pristup usmjeren učeniku, evidentno je da je za zahtjevano interaktivno učenje i poučavanje nužno iskoristiti tehnologije i alate koji u nastavnom procesu aktivno angažiraju i učenike i nastavnike.

II. INTERAKTIVNA OKOLINA UČENJA I KONSTRUKTIVIZAM

Integracijom tehnologije mijenja se proces učenja i poučavanja u interaktivnu okolinu učenja. Pri tome se tehnička i pedagoška interaktivnost objašnjava kombinacijom interaktivnog poučavanja i učenja podržanog tehnološkim alatima.

Uključimo li, dakle, u poučavanje projekcijski sustav, Internet vezu, interaktivnu ploču i web 2.0 alate stvaramo interaktivnu okolinu učenja. Slijedom toga, okvir koji se primjenjuje na takav novi oblik poučavanja, interaktivno učenje i poučavanje, predstavlja zajedništvo teorije (konstruktivizam), interaktivnog hardvera (interaktivne ploče) i odgovarajućeg web 2.0 alata (slika 1, [6]).



Slika 1 Komponente teoretskog okvira (prema [6}, pojednostavljeno)

Ovako definiran okvir može uključivati bilo koji alat koji prirodno vodi prema interaktivnim okolinama učenja. Kombinacija je teorije učenja i alata pa uključuje konstruktivizam, interaktivno poučavanje, interaktivno učenje, interaktivne alate (hardver i softver) i interakciju između učenika i nastavnika.

Pojmovno konstruktivizam čini konstrukcija učenja ili znanja naglašavajući aktivnu ulogu učenika u njihovom osvješćivanju s obzirom na njihovu okolinu i njihova iskustva u toj okolini (Abdal-Haqq, 1998, Vygotsky, 1978). Temelj su, dakle, aktivni učesnici, a interaktivne ploče u ovom kontekstu naglašavaju upravo to. Pri tom interaktivna ploča pomaže ostvarenju interaktivne okoline učenja tako da učenicima nudi različite mogućnosti povezivanja sa sadržajem, a njihova vizualna priroda olakšava razvoj shema potrebnih za kreiranje novog učenja („*learning by doing*“). Kombinacijom vizualizacije, interaktivne ploče i interaktivnih alata učenicima je omogućeno kreiranje znanja kroz konstruktivističku paradigmu 21. stoljeća.

Na taj je način interaktivno učenje povezano sa učenjem uz pomoć tehnologije i rezultira novim procesom interaktivnog učenja i poučavanja.

Kombinacija konstruktivizma, interaktivne ploče i web 2.0 alata jedna je mogućnost u kojoj su i učenici i nastavnici u središtu procesa. Web 2.0 alati predstavljaju novi pristup interakciji sa sadržajima jače obilježenim sudioništvom i interaktivnošću. Pri tom su nastavnici odgovorni za planiranje, poučavanje i pojednostavljenje dijelova integriranih s tehnologijom, a učenici za konstruiranje i demonstriranje znanja kao i suradništvo sa kolegama prilikom kreiranja znanja kroz upotrebu alata koji im pomažu pristupiti informacijama i procesirati ih.

III. INTERAKTIVNO UČENJE I POUČAVANJE I OBRAZOVNE TEHNOLOGIJE

Formalno, obrazovna tehnologija je „primjena tehnoloških procesa i alata koji se mogu koristiti za rješavanje problema nastave i učenja“ [7]. Uključuje principe, procese i proizvode koje nastavnici i učenici upotrebljavaju kako bi unaprijedili učenje [8].

Obrazovna tehnologija omogućuje modifikaciju okoline učenja i poučavanja, s općim ciljem konstruiranja i predaje optimalnih iskustva učenja, a u nastavku teksta naglasak je na njenom programskom i hardverskom dijelu potrebnom za ostvarenje interaktivnih iskustava učenja.

U suvremenom su obrazovanju obrazovne tehnologije sveprisutne i neizbježne pri omogućavanju učenicima da pristupe stvarima koje bi inače na drugi način bile nemoguće, nesigurne, nepraktične ili nemotivirajuće. Uz njihovu pomoć učenici uče više, a nastavnici poučavaju bolje. Zato ne čudi što se i dalje obavlja niz istraživanja u traženju najboljeg načina upotrebe tehnologije s ciljem poboljšanja obrazovne produktivnosti, pri čemu su od velikog značaja upravo interaktivno poučavanje i interaktivno učenje.

Interaktivno poučavanje obilježavaju predavanja oplemenjena aktivnim demonstracijama, pri čemu nastavnik integrira različite tehnologije i oblike medija kako bi ohrabrio kognitivnu participaciju kod učenika.

Interaktivno učenje znači da su učenici aktivni sudionici u procesu učenja. Ako primjerice okolina učenja integrira interaktivnu ploču, učenici su verbalno ili fizički s njom u interakciji. Ovakvo na učenika usmjereno učenje slijedi principe konstruktivističkog učenja, kao sastavnice interaktivne okoline učenja. Učenike se ohrabruje da upravljaju svojim učenjem i konstruiraju značenje.

Interaktivnost pri tom može imati različite definicije. To može biti *click* na poveznicu *(link)* i pristupanje tekstu ili kompletiranje radnog lista, a uz interaktivnu ploču interaktivnost znači da nastavnik i učenik izvode fizičku aktivnost prijenosa npr. geometrijskog lika ili riječi. Obje aktivnosti dovode do interakcije sa kognitivnim procesima pojednostavnjujući konstrukciju znanja. Interaktivnost također znači i aktivni angažman nastavnika i učenika u razgovoru.

IV. PROGRAMSKI DIO INTERAKTIVNIH OBRAZOVNIH TEHNOLOGIJA

Stalno promjenjiva priroda inicirala je da te tehnologije danas možemo podijeliti u najmanje devet kategorija [1]: aplikacije za obradu teksta, programska potpora za organiziranje ili „oluju mozgova“, alati za prikupljanje podataka i analizu, programska potpora za komunikaciju i suradnju, instrukcijski mediji (učenik kao potrošač), stvaranje multimedije (učenik kao proizvođač), nastavne interaktive, izvori baza podataka i referenci, te kinestetičke tehnologije. Njihova upotreba u nastavi danas ima posebno značenje s obzirom na zahtjev za poticanje međupredmetnih kompetencija učenika i njihovu pripremu za cjeloživotno učenje.

TABLICA I. DEVET KATEGORIJA TEHNOLOGIJE (aplikacije, prema [1])

Kategorija	Definicija	Primjeri
Aplikacije za obradu teksta	Primjene koje stvaraju dokumente sa tekstom ispisanim na linearne ili vizualne načine	Google Docs, Microsoft Word
Programska potpora za organiziranje ili „oluju mozgova“	Programska potpora koja korisnicima pomaže organizirati razmišljanje, povezivanje i kategorizaciju ideja te prikazivanje procesa	Webspiration, Inspiration, SmartTools
Alati za prikupljanje podataka i analizu	Alati koji omogućuju učenicima prikupljanje i analizu podataka	SurveyMonkey, Microsoft Excell, eClicker, Poll Everywhere
Programska potpora za komunikaciju i suradnju	Programska potpora koja mijenja ili unapređuje tradicionalne oblike komunikacije video, audio, tekstualnom ili kombinacijama; korisnicima omogućuje dijeljenje i raspravljanje ideja, slika, poveznica i dr.; omogućuje zajednički rad čak i kod geografski razdvojenih sudionika	Skype, Facebook, Twitter
Instrukcijski mediji (učenik kao potrošač)	Tehnologije koje osiguravaju ili olakšavaju stvaranje videa ili zapisa namijenjenih upotrebi pri učenju	BrainPOP, Discovery Education Streaming
Stvaranje multimedije (učenik kao proizvođač)	Tehnologije koje omogućuju korisnicima kombiniranje audio, video, glazbe, slika, crteža ili kombinacije u konačni proizvod	PowerPoint, Keynote, Glogster, iMovie
Nastavne interaktive	Tehnologije kojima rukuju učenici kako bi povećali razumijevanje vještine ili koncepta uključujući igre, manipulative i programsku potporu koja procijenjuje učenika i diferencira aktivnost ili kurikulum bazirano na potrebama učenika	MathBoard, Star Chart
Izvori baza podataka i referenci	Izvori koji korisnicima osiguravaju informacije i podatke	RubiStar, Visual Thesaurus, Wikipedia, GapMinder
Kinestetičke tehnologije	Tehnologije u međudjelovanju sa korisnikovom geografskom ili fizičkom lokacijom ili kretanjem (očekuje se veliki rast slijedećih godina)	Nintendo Wii, Xbox, Kinect, GPS uređaji

Ne umanjujući važnost ostalih, vezano uz interaktivnost valja prokomentirati nastavne interaktivne aplikacije. Najvažnije karakteristike tih nastavnih interaktiva su da osiguravaju povratnu vezu, prepoznavanje, omogućuju nagađanja, pretpostavke, upite, nelingvističke prikaze, zadavanje domaćih zadataka, omogućuju vježbanje, stvaranje i ispitivanje hipoteza.

Pri tom je naglašen značaj osiguravanja upravo povratne veze. Osiguravanje učenicima korektivne, pravovremene i na kriterije, odnosno cilj učenja fokusirane povratne veze stvara okolinu učenja koja podupire učenje. Tako primjerice igre i simulacije omogućuju učenicima i nastavnicima neposredno preusmjeravanje ili korekciju nerazumijevanja.

V. IZAZOVI UPOTREBE TEHNOLOGIJA U NASTAVI (HARDVERSKI DIO)

A. *Pametne ploče*

Interaktivne bijele ploče smatraju se jednom od najrevolucionarnijih obrazovnih tehnologija, no još uvijek se javlja nerazumijevanje kako ih koristiti kao nastavni alat i time smisleno i kreativno iskoristiti njihove tehnološke blagodati i izazovno bogatstvo nastavnih sadržaja.

Kako uporaba pametnih ploča u nastavi dokazano pospješuje učenje, ne iznenađuje podatak da praktički sve škole u Sjedinjenim Državama posjeduju pametnu ploču ili više njih, dok je u Velikoj Britaniji ona prisutna u gotovo svakoj učionici, a u nekim područjima Njemačke već je oko četvrtine nastavnih prostora pokriveno interaktivnim pločama (npr. Hamburg).

Ispravno upotrebljene, interaktivne ploče osiguravaju prikaz i interaktivnost, pogodne su za učenje kroz igru, promoviraju interakciju učenika sa sadržajem, snažniju vizualizaciju koncepata i kroz njihovo aktivnije učešće značajno povećavaju njihovu motivaciju te oni pokazuju veće zanimanje za ovakvu slikovitiju i dinamičniju nastavu. U budućnosti će ipak biti potrebno više ukazivati na različite mogućnosti dohvata na internetu dostupnih digitalnih nastavnih materijala obzirom da mnogi nastavnici o tome ne znaju dovoljno. Uz to, ako se koristi kao jedina tehnologija, neka izvješća pokazuju i kako su ipak pri tom jedva dotaknuti „individualizirani oblici učenja, nastava orijentirana kompetencijama ili suradničko učenje u nastavnom procesu, a ploča je sveukupno svoje prednosti pokazala više u nastavi usmjerenoj na nastavnika nego na učenike“ [1].

B. *Tablet računala*

Tablet računala imaju mogućnost da u osnovi trajno promjene medijsku integraciju u kontekst učenja.

Njihovim uvođenjem u škole omogućava se poboljšana interakcija između učenika i nastavnika, ostvaruje se bolji pristup nastavnim sadržajima i olakšava učenje prilagođeno individualnim sposobnostima

Učenici uče više i trajnije, povećana je fleksibilnost i mobilnost nastave, olakšano je projektno učenje i problemska nastava. Zbog njihove pouzdanosti, raspoloživosti, intuitivnog korištenja i cijene upotrebe vrlo su primjenjivi za korištenje u nastavi. U odnosu na

prijenosna računala, uređaji su lakši i mobilniji, manje podložni softverskim problemima. Uz to, imaju potencijal poticanja nastave usmjerene na učenika, no podržavaju i timsko i individualizirano učenje.

U didaktičkom smislu uvođenje tablet računala pozitivno utječe na motivaciju učenja, interaktivnu nastavu i poticanje različitih tipova učenja. Kroz interaktivne nastavne sadržaje različitih aplikacija mogu se lakše predstaviti i individualno obrađivati razna područja, pri čemu interaktivni sadržaji omogućuju višekanalno učenje. Brojnost aplikacija, mogućnosti interakcije i vizualizacija održavaju nastavu zanimljivom i pokazuju raznolike mogućnosti učenja, a kroz interaktivnost potiče se vlastita odgovornost i timski rad. Uz pomoć aplikacija učenici se u nastavi mogu i povezivati, a motivaciji pridonosi i brisanje granica između privatne i školske komunikacije. Sve ovo može dovesti do dubljeg povezivanja i ukorijenjivanja opsega znanja. Cilj je da ovakav uređaj postane centralom osobnog upravljanja znanjem.

Prihvatljivost uređaja kao alata za učenje trebala bi biti jače naznačena u odnosu na uobičajene strategije koje ne posustaju (papir i olovka), a još bi se trebala etablirati i odgovarajuća sasvim promijenjena kultura učenja. Također, tableti ne bi trebali u potpunosti zamijeniti vrijedne udžbenike nego ih obogatiti faktorom interaktivnosti.

Tehnički, pažnju treba obratiti na instalaciju i aktualizaciju aplikacija i njihovo usklađivanje sa interaktivnim pametnim pločama i sustavima za upravljanje učenjem (LMS).

Nadalje, radi neograničenog korištenja ovih uređaja prilikom razmjene među korisnicima evidentna je potreba za povećanjem stupnja personalizacije. Iz istraživanja objavljenog u [5], uloga ovog uređaja u nastavi ovisi o stupnju upravljanja procesima učenja dionika. Što je jača personalizacija uređaja ili bi mogla biti toliko se češće postiže na učenika usmjereni način rada. Nadalje, mobilni i personalizirani uređaji pridonose nestajanju granica školske i privatne uporabe medija, odnosno integraciji digitaliziranih životnih prostora učenika.

C. Pametni telefoni

Dobijen je niz iznenađujuće pozitivnih saznanja koja pokazuju koliko su ti mali mobilni kompjutori sa *touchpad* tehnologijom i sve većim mogućnostima sposobni dopuniti najrazličitije aspekte i ostvariti učinkovitije i produktivnije okruženje učenja.

Optimističnim se pokazalo i saznanje da je ove uređaje moguće čak i bez specifičnih programa za učenje smisleno i na razne načine upotrebljavati u nastavi i poticati izjednačavanje učenika. Uočeno je i poticanje socijalne interakcije porastom izravne komunikacije među učenicima, ali i sa nastavnicima. Olakšana je individualizacija nastave s obzirom da su učenici zahvaljujući vlastitom uređaju mogli napredovati vlastitim tempom ostvarujući odgovarajuću razinu uspjeha.

S obzirom da je određene akcije CMS teže izvesti preko pametnog telefona, pokazuje se potreba za njihvoim *mobile-friendly* verzijama [2].

VI. ZAKLJUČAK

Posljednjih se nekoliko godina primjećuje snažan pomak sa upotrebe alata i aplikacija na lokalnim računalima prema upotrebi mnoštva novih tehnologija smještenih „u oblaku“ *online* nudeći učenicima raznolika iskustva učenja.

U takvom okruženju nastavnici učenike trebaju više doživljavati kao proizvođače znanja nego kao potrošače što podrazumijeva aktivno sudjelovanje učenika u procesu učenja, a upravo to je i karakteristika interaktivnih učionica koje koriste interaktivne alate. Tako učenici funkcioniraju na najvišoj razini vještina razmišljanja višeg reda pri čemu znanje analiziraju, sintetiziraju i prezentiraju kolegama lokalno i globalno. Danas dostupni alati imaju moć osigurati potrebne preduvjete za takav novi model interaktivnog poučavanja i učenja.

Značajno je i to što upotreba interaktivnih tehnologija provjereno vodi ispunjenju cilja povećanja motivacije učenika za učenje. Ipak, iako su trenutno dostupne mnoge komercijalne aplikacije, potencijal novih medija u odnosu na mogućnosti kreativnog i interaktivnog rada još nije iskorišten.

Nadalje, za nastavu i nastavni kadar trebalo bi zbog uvođenja novih medija razviti program usvajanja medijskopedagoških i didaktičkih koncepata. Nastavno osoblje moralo bi biti školovano za ovu svoju novu ulogu. Umjesto pukog prijenosa postojeće na nastavnika usmjerene nastave, treba medijski pripremiti sadržaje i ostvariti interaktivnom tehnologijom podržanu okolinu učenja.

LITERATURA

[1] Pitler, H., Ross Hubbell, E., Kuhn, M. : **Using technology with classroom instruction that works**, 2nd ed, McREL, 2012.

[2] Stoeger, S., Barth D.: **Exploring mobile tehnologies to improve student learning**, 2012.

[3] Türel, Y. K., & Johnson, T. E. (2012). **Teachers' belief and use of interactive whiteboards for teaching and learning**, Educational Technology & Society, 15 (1), 381–394.

[4] http://www.digital-lernen.de (posjećeno, 1/2013.)

[5] Groebel, J., Wiedermann, J.: **Digitale Lernwerkzeuge**, Metastudie der BSP Business School Berlin Potsdam, 2012.

[6] Sessoms, D**.: Creating interactive learning environments through tomorrow's teachers**. *International Journal of Technology in Teaching and Learning, 4*(2), 86-96. (2008) Interactive instruction.

[7] Seels, B. B., Richey, R. C: **Instructional technology: The definition and domains of the field**, Washington, DC: Association for Educational Communications and Technology, 1994.

[8] Newby, T. J., Stepich, D. A., Lehman, J. D., Russell, J. D., Ottenbreit-Leftwich, A.: **Educational technology for teaching and learning**, Pearson, 2011.

PC and microcontrollers applications in the laboratory exercises of the electrical engineering

Olivera Tasić*, Viša Tasić**, Darko Brodić***, Vladimir Despotović***, Marijana Pavlov**, Dragan R. Milivojević**

*Mechanical and Electrical High School, Bor, Serbia
**Mining and Metallurgy Institute, Bor, Serbia
*** Technical Faculty in Bor, University of Belgrade, Bor, Serbia
e-mail: visa.tasic@irmbor.co.rs

Abstract - Development of the new technologies in telecommunications, electronics and electrical engineering is a very dynamic process. It creates the need of constantly improving teaching programs in the field of electrical engineering. That includes exercise-related items like the use of computers and microcontrollers in automatic control systems. Accordingly, the presented examples of laboratory exercises can significantly contribute to better understanding of the planned curriculums.

I. INTRODUCTION

Non scholae, sed vitae discimus (not for school but for life we learn). The old Latin phrase describes the goal of any educational system. Mechanical and Electrical High School Bor was established in 1990 by the decision of the Municipal Assembly of Bor, on the basis of Secondary Education Low, as an independent institution for education pupils in mechanical and electrical engineering [1]. School covers three education profiles: electrical engineering, mechanical engineering and transportation.

The educational scope in electrical engineering covers several educational profiles, such as electrical technician for: multimedia, computers, computer networks, automation, and electronics and electrical energy. Rapid development of the new technologies in electronics, telecommunications and electrical engineering creates the need to constantly improve teaching Practical experience that pupils gain in the laboratory is of particular importance for their understanding of the applicability of theoretical knowledge. Simple, clear, and interesting experiments and exercises are essential to motivate them to expand their knowledge.

Selected examples of laboratory exercises based on the application of microcontrollers and computers in automatic control systems are presented in the paper. Due to well cooperation established between educational institutions and commercial enterprises, our pupils perform their practical exercises not only in laboratories within the school, but also in laboratories of other scientific institutions and enterprises within the town [2].

II. COMPUTER AND MICROCONTROLLERS APPLICATION IN AUTOMATIC CONTROL SYSTEMS

In this part of paper, some examples of laboratory exercises that use Intel, Motorola, Siemens, and Microchip microcontrollers and microprocessors will be shown. It is well known that microcontrollers are designed for embedded applications. A microcontroller is a single chip that contains the processor (CPU), non-volatile memory for the program (ROM or flash), volatile memory for input and output (RAM), a clock and an I/O (Input/Output) control unit.

A. Traffic Light

The first example is a simple control application for programming the traffic light using different microcontrollers and PC software.

The first control application is realized using Siemens Logo! 12/24RC logic module. In its basic configuration it has integrated display and operator panel, 8 digital inputs, of which 2 can be used in analog mode and 4 relay outputs. Number of inputs/outputs can be increased using the extension modules. The Logo control system is perfectly suited for educational purpose, as it simplifies the automation process by replacing many time switches and relays, counters and protective relays with logic modules. Well designed exercises can encourage pupils to realize their own small-scale automation projects.

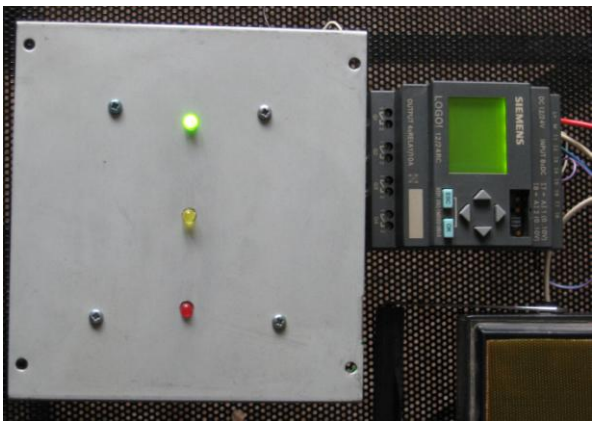


Figure 1. Simple model of traffic light realizes with Siemens Logo! 12/24RC logic module

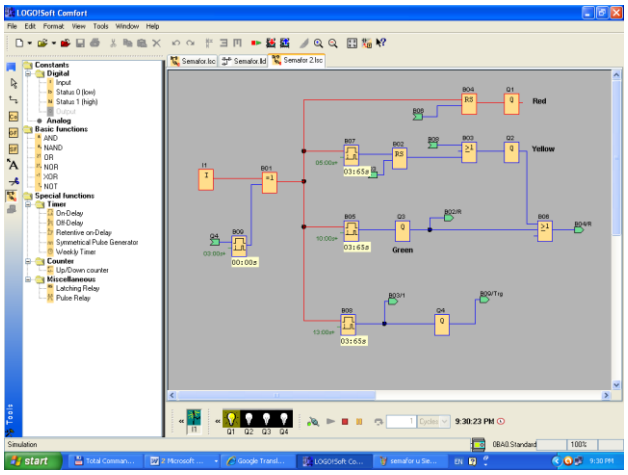


Figure 2. Programming the traffic light in Logo! Soft Comfort software using the functional block diagrams

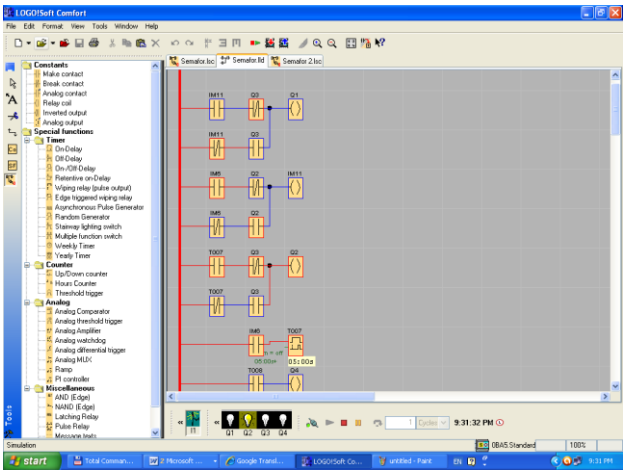


Figure 3. Programming the traffic light in Logo! Soft Comfort software using the ladder diagrams

Figure 1 presents a model of traffic light, realized with Siemens Logo logic module [3]. Figure 2 shows an example of the program created with Siemens Logo! Soft Comfort software. This programming software enables control programs to be created, tested, edited, archived and printed out on a PC. The process of creating a program involves positioning and linking up program components from the predefined library (see left side in the Figure 2) on a "drawing board". One particularly user-friendly feature is the offline program simulation facility which enables simultaneous display of multiple function statuses. Except in simulation mode programs can be also uploaded into the Logo logic controller and executed in real-time. Logo! Soft Comfort provides two options for creating circuit programs: a) functional block diagram (see Figure 2) or b) ladder diagram (see Figure 3).

Ladder logic is a programming language that represents a program by a graphical diagram based on the circuit diagrams of relay logic hardware. Ladder diagram programming is important when a programmable logic controller (PLC) is used primarily to replace relays, timers, and counters [4]. Analog quantities and arithmetical operations are unsuitable to be expressed in ladder logic, and each manufacturer has different ways of extending the notation for these problems [5].

Programming the same application in many different programming environments is very rewarding experience for pupils and students. For this reason, the same exercise was realized using the Microchip PIC16F877A CMOS FLASH-based 8-bit microcontroller (see Figures 4 and 5). MPLAB Integrated Development Environment (IDE) is used to develop the embedded application for PIC microcontrollers. It runs under Microsoft Windows operating systems and includes code editor (supports assembler and C languages), simulator with the debugger, and project manager [6].

However, a special development environment, named DLadder, is developed in Borland Delphi 7 programming language, for programming the PIC16F877A microcontroller. It allows programming using ladder logic, interpretation and compilation of programs into native code, debugging and the real-time simulation. It also allows the contents of the microcontroller ports and

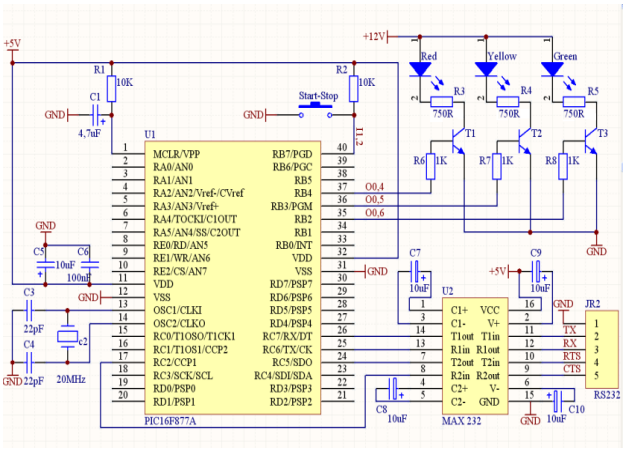


Figure 4. The traffic light electric scheme realized with the PIC16F877A microcontroller [5]

memory locations real time presentation [5]. The primary motive for DLadder IDE developing is a desire to develop programs for PIC microcontrollers using the concept of visual programming. Although there are a number of development environments for PIC microcontrollers on the market, few of them allow writing the programs in ladder logic. Writing a ladder diagram is performed by selection the required elements from the object toolbar (shown in Figure 6) and setting them in the appropriate position on the screen.

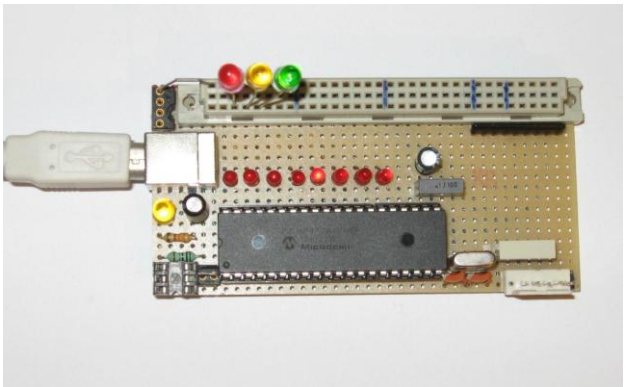


Figure 5. Model of traffic light realized with the PIC16F877A microcontroller

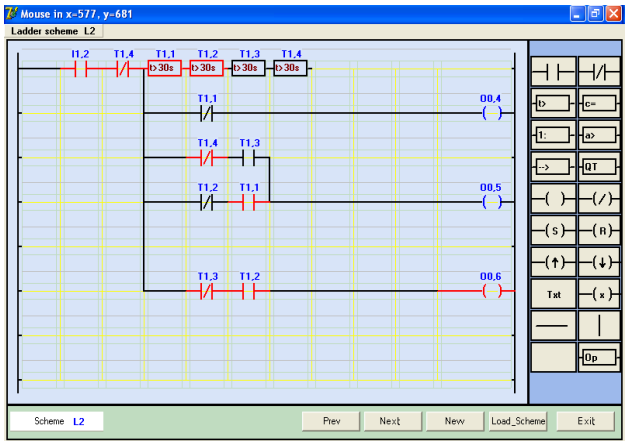


Figure 6. Programming the traffic light in DLadder software

The function objects support the arithmetic operations, logic operations, etc. An example of a simple ladder program, which simulates operation of the traffic light, is shown in Figure 6. Once the code has been built and checked from the syntax point of view, it should be tested. When SIMULATOR option in DLadder is selected, the program appears on the screen with the energized (true) branches highlighted (red color), as shown in Figure 6. The debugging process is simplified in this way.

Furthermore, the traffic light program can be written in assembler language for 8086 microprocessor and executed under the assembler and microprocessor emulator EMU8086 [7], as shown in Figures 7 and 8. EMU8086 supports user-created virtual devices that can be accessed from assembly language program using **in** and **out** instructions. The traffic light lamps are controlled by sending data to I/O port 4 [8]. There are 12 lamps: 4 green, 4 yellow, and 4 red. One can set the state of each lamp by setting its bit:

1 - the lamp is turned on, 0 - the lamp is turned off.

Only 12 low bits of a word are used (0 to 11), last bits (12 to 15) are unused. For example:

```
MOV AX, 0000001011110100b
OUT 4, AX
```

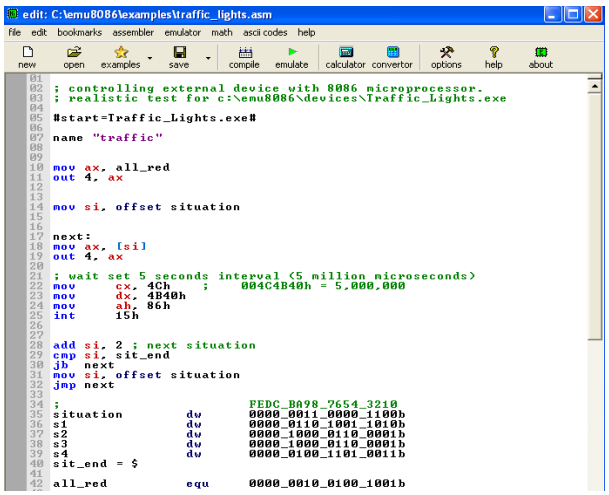


Figure 7. The semaphore example: assembler language code in the emulator EMU8086 editor window [7]

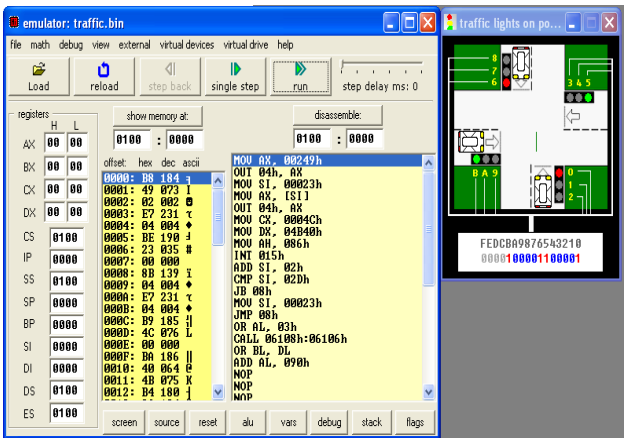


Figure 8. The semaphore simulation in the emulator EMU8086 [8]

B. *Stepper motor*

Next exercise is related to the development of the program to control a stepper motor. Stepper motor is an electric motor that can be precisely controlled by signals from a computer. The motor turns through a precise angle each time, when it receives a signal. By varying the rate at which signal pulses have been produced, the motor can be run at different speeds or can be turned through an exact angle and then stopped [8].

The first part of this exercise refers to the use of a stepper_motor.asm example in the emulator EMU8086 [7]. Figure 9 shows a basic 3-phase stepper motor. It has 3 magnets controlled by bits 0, 1 and 2. Other bits (3..7) are unused. The stepper motor is controlled by sending data to I/O port 7. When a magnet is working, it becomes red. The arrow in the Figure 9 shows the direction of the last motor move. In the stepper motor example, the code below will do three clock-wise half-steps:

```
MOV  AL, 001b; initialize.
OUT 7, AL
MOV AL, 011b; half step 1.
OUT 7, AL
MOV AL, 010b; half step 2.
OUT 7, AL
MOV AL, 110b; half step 3.
OUT 7, AL
```

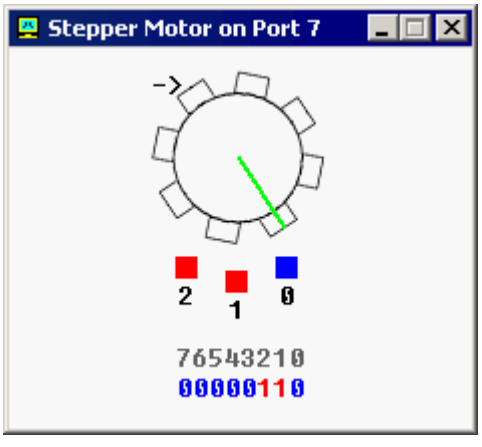


Figure 9. The stepper motor simulation in the emulator EMU8086 [8]

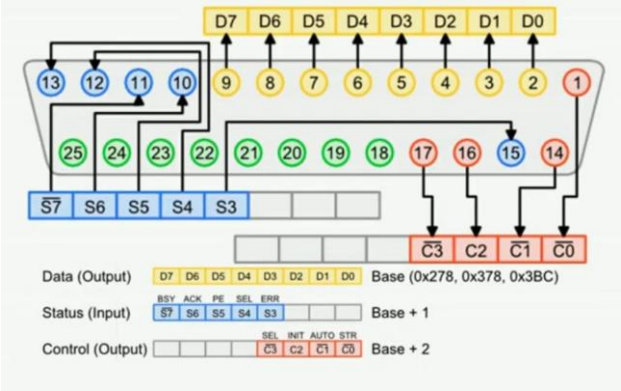


Figure 10. The pin assignment of PC parallel port connector

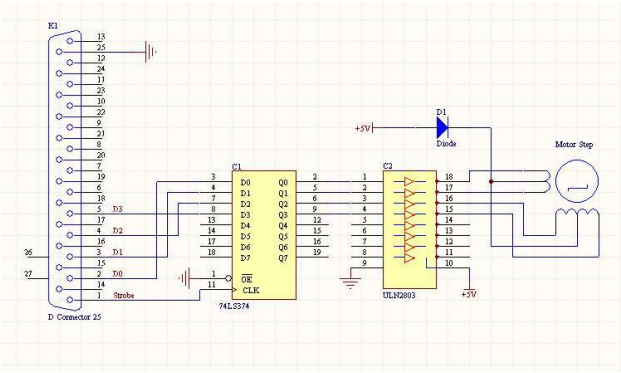


Figure 11. The electric scheme realized to control the stepper motor by use of the PC parallel port

The second part of this exercise refers the use of the PC parallel port (shown in Figures 10, 11 and 13) to control the stepper motor. For this purpose, an appropriate program is created in Borland Delphi 7. The input parameters of the program are: number of steps, direction of rotation, and delay between steps, coil to be switched on, and parallel port address, as shown in Figure 12.

C. Temperature Control

This example is also very simple, but very practical. The short program, shown in Figures 14 and 15, written in the assembler language for Intel 8086 microprocessor presents how to keep a constant temperature using the heater and the thermometer [7].

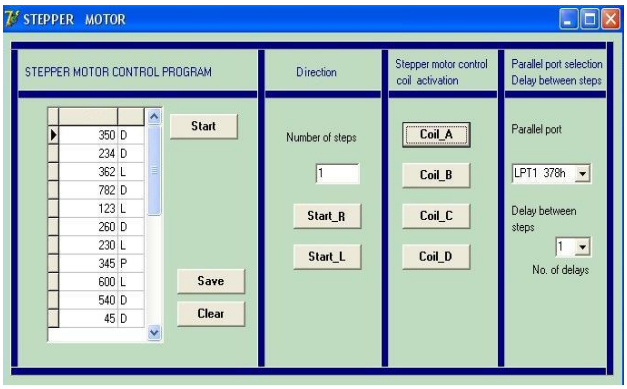


Figure 12. The main window of the program to control the stepper motor with four coils

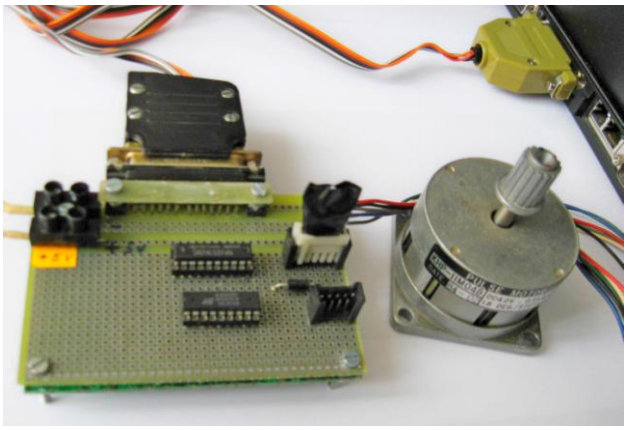


Figure 13. The interface board realized to control the stepper moto

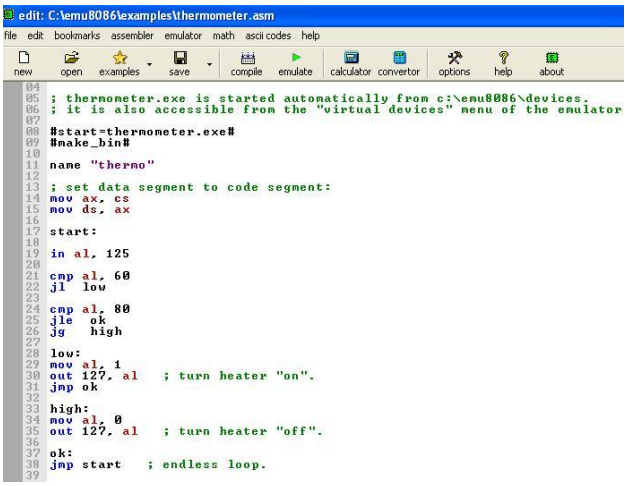


Figure 14. The temperature control example: assembler language code in the emulator EMU8086 editor window [7]

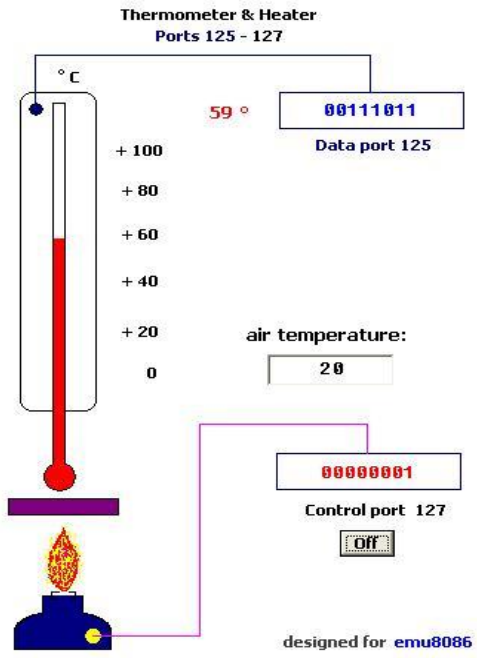


Figure 15. The temperature control example [7]

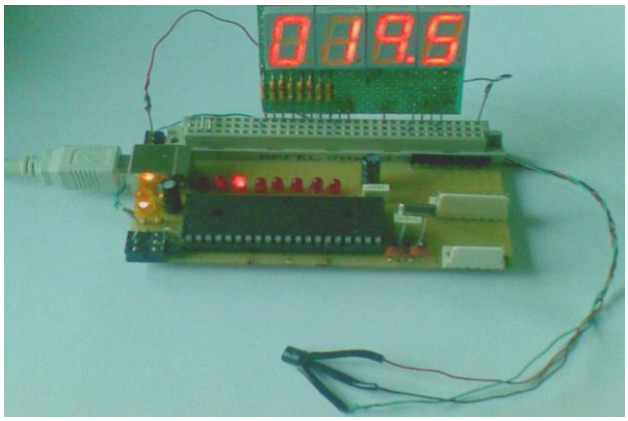


Figure 16. The interface board realized to control the temperature by use of the PIC18F4550 with the LM35 temperature sensor

The second part of this exercise refers to the use of the microcontroller PIC18F4550 to measure the temperature. The LM35 integrated temperature sensor is used [10] (linear scale factor is +10 mV/°C). It does not require any external calibration or trimming to provide typical accuracies of ± 1⁄4 °C at room temperature. Appropriate program is created in DLadder environment and uploaded to microcontroller. While the program is running, the actual value of temperature has been displayed on LED display, as shown in Figure 16. The sensor output is connected to one of the microcontroller analog inputs, as shown in Figure 17.

D. Other Lab Exercises

Laboratory exercises in the field of process control and distributed control systems on the various types of industrial microcontrollers are also carried out. MMS (Microprocessor Measuring Station) is the representative of such equipment, which is still in use [2]. Pupils learn how to configure the MMS and to adjust and verify its validity. The dedicated software for real time operation is executed at the PC workstation, with standard SCADA functions, designed for the use in network environment [9]. Pupils learn and practice how to set the monitoring properties of the program related to the configuration of the input parameters. Furthermore, MMS with power transducers serves to introduce pupils to the principles of measurement of electrical power and control of power consumption [2]. Power transducer output signals are led to the MMS analog inputs. Measurement procedure consists of input signal A/D conversion, and processing of obtained numerical values. The power transducer gives the standard voltage signal (0 - 5 V DC) as an output, proportional to the active and reactive power. Laboratory exercise consists of setting of the output signal from the power transducers by setting the value of consumer's load and setting of MMS analog inputs based on A/D conversion results from readings [2].

III. CONLUSION

Interesting and simple experiments and exercises are essential to motivate pupils to expand their knowledge. Shown exercises can be equally used in the courses that deal with programming languages, as well as in the courses that deal with the elements of automatic control systems. One should note that good cooperation between educational institutions and commercial enterprises is essential to the quality of the teaching process, which enables the use of external laboratories as education bases.

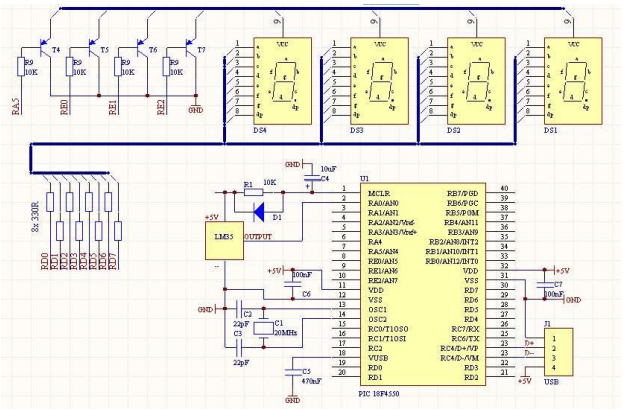


Figure 17. The electric scheme realized to measure the temperature by use of the PIC18F4550 with the LM35 temperature sensor

ACKNOWLEDGMENT

This work is supported by a grant from the Ministry of Education, Science and Technological Development of Republic of Serbia, as a part of project No. TR33037 "Development and Application of Distributed System for Monitoring and Control of Electrical Energy Consumption for Large Consumers."

REFERENCES

[1] http://www.mesbor.rs/ (accessed 10 January 2013)

[2] V.Tasić, D.Brodić, D.Milivojević, M.Pavlov, "Practicing of Lab Exercises at the Laboratory of Applied Electronics and Computer Engineering", Proceedings of 6th. International Scientific Conference Computer Science 2011, 01.-03.09.2011. Ohrid, FJR Macedonia, pp. 429-434, ISBN 978-954-438-914-7.

[3] http://www.automation.siemens.com (accessed 10 January 2013)

[4] W. Bolton, Programmable Logic Controllers, Fifth Edition, Newnes, 2009.

[5] V.Tasić, D.Milivojević, V.Despotović, D. Brodić, M.Pavlov, V.Miljković, "Dladder - an Integrated Environment for Programming PIC Microcontrollers", Proceedings of XLVII International Scientific Conference on Information, Communication and Energy Systems and Technologies ICEST 2012, 28.6.-30.6.2012, V.Trnovo, Bulgaria, vol. 2., pp. 577-580.

[6] Microchip Inc., MPLAB ICD 2 In-Circuit Debugger User's Guide, Microchip Technology Inc., 2005. http://ww1.microchip.com/downloads/en/devicedoc/51331b.pdf (accessed 10 January 2013)

[7] http://www.emu8086.com (accessed 10 January 2013)

[8] http://www.ziplib.com/emu8086/virtual_devices.html (accessed 10 January 2013)

[9] D.Milivojević, V.Despotović and V.Tasić: "Process Control Program as an Element of Distributed Control System", Information Technology and Control, vol.39, pp. 152-158. No. 2, 2010

[10] www.ti.com/lit/ds/symlink/lm35.pdf (accessed 10 January 2013)

On-line E-learning platform supporting education and practice of accountants in EU space

D. Paľová, M. Vejačka
Department of Applied Mathematics and Business Informatics
Faculty of Economics, Technical University of Košice, Slovakia
dana.palova@tuke.sk, martin.vejacka@tuke.sk

Abstract - Accountancy presents very important part of the economy/economic strategies of the states, enterprises or ordinary people, because it allows monitoring and also more effective planning of spending available sources or searching for possibilities how to cover present operations but also further development of mentioned entities. At our country, but also in all EU, the accountancy is educated not only at specialized high schools, universities, but also at institutions providing lifelong learning for adults. The common and long-term lack of provided education is concentrated mostly on theoretical background of the students, but less emphasis on achieving practical skills. Missing practical skills (together with many other factors) lead to high unemployment rates of young people, mostly of new graduates. This problem has lead us to an idea to propose and develop tool, which helps people at risk to achieve skills needed already during their study and ease the entering to labor market to them. Mentioned tool is developed by European universities from five states (Greece, United Kingdom, Bulgaria, Poland and Slovakia) and thank to this international cooperation it could be said, that final tool will present support tool not only for students from Slovakia, but also for students across the European Union at all.

I. INTRODUCTION

In last decades the whole world including European Union (EU) is marked by transformation from industrial society over information society to declared competitive knowledge society. Replacing the Lisbon strategy [1], by the Europe 2020 strategy [2], governments of EU member states aims to guide Europe's economy out of the economic recession (which began in 2008). By this document EU accredits the crucial role to education and training to ensure Europe's future prosperity and social cohesion. It identifies five crucial areas to be reformed: (1) education, science and innovation; (2) employment and social inclusion; (3) business environment; (4), transparency and rule of law; and (5) health.

Making lifelong learning and mobility a reality, improving the quality and efficiency of education and training, promoting equity, social cohesion and active citizenship and enhancing creativity and innovation, including entrepreneurship, at all levels of education and training are the objectives of the European Strategic Framework for Education and Training [3].

Accountancy presents very important part of the economy and economic strategies of the states, enterprises or ordinary people, because it allows monitoring and also more effective planning of spending available sources or searching for possibilities how to cover present operations but also further development of mentioned entities.

Acceptable practical experience is extremely important in all areas of accounting. It is this experience which distinguishes graduates with any accounting qualifications in the job market. However, gaining acceptable practical experience remains a major issue across Europe.

Accounting standards in each country differs, complicating the work of the accounting practice, especially in cases of various international companies and cooperation projects. However education in accounting is dealing with this problem frequently, only occasionally we can found accounting practices included in educational institutions curricula. Due to globalization of the world, accounting principles are frequently used not just by the accountants, but knowledge of this area is also needed for the people with an education from other areas. Search for information for non-accountants in these cases is often time consuming and the result can be uncertain and unclear. Project FASTER has the ambition to develop supporting tools to help different target groups (e.g. students of accounting, non-accountant professionals, or for any person interested in accounting.

II. EU LABOUR MARKET DEMAND ANALYSIS

Based on [1], [4] the key competences for knowledge society include ability to engage the new technologies and creativity, problem solving and team work. That is why the basic digital competence covers:

1. Basic operating competences related to monitors, computers and any other mediums,
2. Psycho-cognitive competences related to digital signs,
3. Basic computing competence (as a user),
4. Interactive media and network competence,
5. Overall socio-communicative competence associated with cyberspace.

One of the declared objectives of initiative eEurope is to "*bring every citizen, home and school, every business and administration, into the digital age and online*" [5]. Training teachers to adapt their curriculum to new technologies and educate digitally literate students and

continual increase of the quality of education in terms of achieving connection between academia and practice are challenges of present. To fulfill above mentioned it is important to know the needs of potential future employers and their demands on the graduates of the educational institution.

The survey realized by website profesia.sk and ARRA (Academic Ranking agency) follows [6], that the requirements on jobseekers are often much more complex, for example the graduates of economics (especially the area of finance, banking, insurance, management) are expected to be masters at area of economy that they graduates, but also they are expected to have active knowledge of a foreign language, leadership and organizational skills, strong team spirit, flexibility, the ability to analyze, ambition, desire to learn, to self-developing, advanced level of work with MS Office package, driving license B.

Based on results of TREND Analyses survey, the employers are complaining about:

- the lack of preparation of graduates for practical problems solving (76% (chosen possibilities always and often together) where 24% of employers chose possibility always),
- the lack of ability to work independently (72%),
- the lack of people with required qualification (64%),
- followed by weak foreign language skills (58%) and excessive wage claims (57%) [7].

Given the above facts, it is necessary to transform provided secondary and higher education in context of more emphasis on acquiring skills and abilities of students to apply their knowledge to solve practical problems and at the same time develop different supporting ICT tools which help students to achieve also missing practical skills.

III. PROJECT FASTER

A. Background and motivation

As was mentioned previously, theoretical knowledge generated in higher educational systems is often not directly linked to the practical needs met in a real-time job environment. EU Priorities on Knowledge Society and Innovation with objectives such as improving the quality and intensity of investments in human capital for the upgrade of the education system, investing in the future, promoting reforms in the education system and improving access and basic skill levels for all [8]. In that context the emphasis is given to linking the educational system with the labor market with focus on the adaptability of employees and businesses and the gradual increase of the participation in education and training so as to be equipped with the skills and abilities needed to cope with the ever-changing labor market while improving workload quality and productivity.

Five educational institutions from five different EU member states: Athens University of Economics and Business (Greece), Sofia University "St. Kliment Ohridski" (Bulgaria), City University, London (United Kingdom), University of Warsaw (Poland), Technical University of Kosice (Slovakia) decided to promote project FASTER with idea of development of a European e-learning tool supporting European collaboration in the field of Finance and Accounting. This collaboration arised from former participations in relevant research projects on ICT implementation and eservices, the market observation and the tendencies with a specific view to SMEs and public institutions in the new economical context:

- ADAPT: “A Multipurpose Network for Training and Support to SMEs”, Acronym “INTEGRA”, Economic Research Centre of the Athens University of Economics and Business, Greece, 1996 – 1997.
- STRIDE HELLAS: “Organizing Distance Learning in Greece”, “ODILE”, Economic Research Centre of the Athens University of Economics and Business, Greece, 1992–1994.
- Formulating a model of e-learning course for Ministry of National Education, entitled “Managing accounting books”;
- E-Guidance and e-Government Services, Executive Agency “Electronic communication networks and information systems,” (2010-2011)
- Scientific research quality enhancement by providing access to full-text information base and licensed statistical data processing packages, Ministry of Education, (2006-2009)
- Managerial Challenges of ICT-based Learning, Supported by: European Commission, MINERVA – SOCRATES Programme, Duration: 2002-2004
- Finance, Logistics and Production Integration Domain by Web-based Interaction Network, Supported by: European Commission, FP6, Duration: 2006-2008
- E-learning for Introduction and Management of Tele-working, Supported by: European Commission, Leonardo da Vinci, Duration: 2005 – 2007
- IT Education as an Activating Factor of Adults from Areas under Danger of Unemployment, Supported by: European Commission, Socrates/Grundtvig, Duration: 2006 – 2009

Concerning the objectives defined in Europe 2020 strategy, the outcomes of the project will support long-term career advancement, international mobility of employees in accountancy branch, as well as graduates with degrees in disciplines other than accounting, or people employed in business, with the foundation studies to become a professional accountant. Special attention is placed on mature and motivated student seeking a career in accounting for which no prior professional experience will be required. This group of people represents the group at risk due to their high unemployment [8].

B. Aims and objectives

The field of finance & accounting is a profession with a key role in strengthening national economies and the

promotion of entrepreneurship. Based on the analysis of labor market, EC documents and activities done in the field of education and accountancy the project partnership defined main project aims and objectives [8]:

- to produce a modern technological e-learning & knowledge-sharing environment on finance and accounting which allows both students and professionals in pursuit of lifelong learning to gain practical knowledge on their discipline, but not restricted to national borders. The platform will propose besides e-learning environment also a job seeking mechanism.
- to link studies & working environment and widen transnational pursuit of jobs in finance & accounting.
- to offer new incentives and produce a prototype ICT environment expandable to all languages and to all disciplines.

These objectives help consortium to support lifelong learning. The developed platform improve the quality and attractiveness of vocational education and training through the review of its primary structure and curricula qualification, setting professional rights and establishing a system of formal accreditation of learning at a Pan-European Qualifications Framework. Furthermore, FASTER aims at fostering long-term career advancement and international mobility, ensuring that the professional development of certified general accountants remains a lifelong experience which distinguishes graduates with any accounting qualifications in the job market while at the same time providing graduates with degrees in disciplines other than accounting, or people employed in business, with the foundation studies to become a professional accountant [9].

C. Target groups

As is presented previously, the project ambition is via developed FASTER platform to address not only educational institutions (academics and students) but also public institutions and private companies, jobseekers and consultants operating in the field of accountancy. The platform could provide benefits to following target groups:

- undergraduate and graduate students in fields of accountancy,
- people interested in vocational training,
- academics,

  by providing material relevant to the universities curricula and the national context of finance and accounting, supported by self-assessment tools, practical examples and exercises, use-cases, etc.,
- employers, potentially looking for employees in area of accounting.
- jobseekers,

  by providing tools for searching new jobs in the field of accountancy, uploading CV, etc.,
- practical accountants by providing interesting links to books, journals, legislation and vocabulary connected to information usable in practical accountancy.

All platform parts will be designed in 5 different languages (languages of project consortium – partners from Bulgaria, Greece, Poland, Slovakia and United Kingdom). The multilingual character of the platform enables to reach an even larger target group and could help them widen their operating scope/area [9].

D. Tools and methods used for ensuring the expected quality of final platform

To be sure, that designed final platform will reach expected quality and fulfill the requirements of further exploitation in real life; we use more tools to evaluate particular sub-processes done during design phase:

- analysis of future user requirements via on-line questionnaires and interviews,
- analysis of feedbacks of users involved in pilot testing,
- development of different supporting groups of people:
  - Accounting Network – group of people from members the FASTER target groups created with aim to serve the project as the initial international pool of exploitation and dissemination.
  - National Support Groups formed in each country with intention to provide support to the project and to represent the interests of each key player and country. These support groups have to follow the project team activities and outcomes and to give an advice how to improve it.
  - External Evaluation Committee represents an external mechanism of results and outcomes evaluation throughout the project implementation.

IV. FASTER PLATFORM USER REQUIREMENTS ANALYSIS

During early stage of FASTER project was conducted survey regarding user requirements on such platform. Survey was conducted at the same time in four from five cooperating countries – Slovakia, Greece, Bulgaria and Poland. It was realized in form of electronic questionnaire and structured interviews. Potential users from four target groups (stated above) were addressed in this survey as respondents. Aim of the survey was to find out user requirements on such platform and how those potential users evaluate usefulness of such platform and its planned features and job seeking environment. Additionally we recorded any supplementary requirements of potential users of the FASTER platform [10].

Group of students (undergraduates and postgraduates) in finance and accounting in Slovak part of survey compromised from 95 students (89 undergraduate students and 6 postgraduate students). This group contained from 68 women and 27 men what corresponds with structure of students of our faculty. In group of people interested in

vocational training 20 people interested in vocational training, there were 18 people (12 women and 8 men) currently working in area of finance and accounting. In academics' group was 21 people (14 women and 7 men), from which 12 have practical experience in accounting. Group of employers contained 20 managers (11 women and 9 men) and executives from accounting firms [10].

Outcomes of survey was average evaluation from four groups of respondents regarding of both parts of FASTER platform (educational and job seeking). It compromised from ranking of educational platform features and ranking of job-seeking features, both based on the average evaluation of the groups' answers. Respondents assigned their evaluation in 5 level Likert scale (scale from level 1- not important to level 5 standing for very important) for all basic features of future FASTER platform. Further they had possibility to recommend and suggest any other features for further development of platform [10].

Basic significance of educational platform and job-seeker part of platform was investigated by question: "Do you believe that the existence of:

1. online internet-based educational platform with theoretical courses on accounting and practical accounting applications in the teaching of accounting

2. and the platform operated as a job seeking environment on the field of accounting would be useful?

Following table presents results of significance evaluation of educational and job-seeking parts of FASTER platform in Slovakia. Both parts were evaluated very positively (positive is every evaluation above 2.5), job-seeking part was evaluated by academics and employers actually with highest grade (5.00) as very important in such platform [10].

TABLE I. SIGNIFICANCE OF EDUCATIONAL AND JOB-SEEKING PARTS OF PLATFORM IN SLOVAKIA

Slovakia	***Students***	***Academics***	***Employers***	***People interested in vocational training***
Educational part	4.08	4.83	4.95	4.55
Jobseeking part	3.95	5.00	5.00	4.45

1-5 Likert scale (1: not important – 5: very important)

Features of FASTER platform should include in its educational part concise textual and slide-based presentation of the theory, practical applications in various accounting topics, case studies, links to various electronic sources (like accounting databases, national and international accounting sites, accounting journals, E.U. and national regulations and legislations), suggested further study materials for various accounting topics, glossary in all five languages of countries participating in FASTER project and self-assessment tests in form of multiple choice and true or false questions [10].

In job-seeking environment of FASTER platform should be possibility to upload announcements of job offers in accounting and finance for firms from all countries of project partnership, job seekers could upload job-seeking advertisement, their CVs or complete predefined CV template. Also direct communication between employers and potential employees should be enabled.

Following table contains average evaluation of significance of all proposed features of FASTER platform from all target groups in relevant partner countries. Again Likert 5 level scale was provided to respondents to assign their evaluation to each of proposed features of platform.

TABLE II. SIGNIFICANCE OF EDUCATIONAL AND JOB-SEEKING PARTS OF PLATFORM IN PARTNER COUNTRIES

Students	***Slovakia***	***Greece***	***Bulgaria***	***Poland***
Educational part	4.08	3.87	3.24	4.42
Jobseeking part	3.95	4.27	3.92	4.49
Academics				
Educational part	4.83	4.39	4.14	3.94
Jobseeking part	5.00	4.27	4.29	3.43
Employers				
Educational part	4.95	4.33	3.65	4.44
Jobseeking part	5.00	4.27	3.20	3.48
People interested in vocational training				
Educational part	4.55	4.03	3.40	4.65
Jobseeking part	4.45	4.20	4.00	4.06

1-5 Likert scale (1: not important – 5: very important)

Respondents evaluated both parts of platform and their features very positively. As the most important features of FASTER platform respondents identified practical examples and applications in various accounting subjects and case studies, links to EU and national regulations and legislations, concise presentation of the theory in the form of text, glossary of accounting terminology in Slovak and in English and proposed further reading for various accounting subjects [10].

Respondents further proposed to add to the educational part of platform answers to frequently asked questions, possibility to send feedback from users to the platform administrator, consultations with accountancy lectors, online communication between users, instructional videos, online advisory and consultancy in the area of accountancy legislation [10].

In job-seeking part of FASTER platform respondents considered as the most important features possibility to upload job offers by domestic and abroad firms, possibility to upload job-seeking ad by job-seeker and the possibility to directly communicate with potential employers. They would further welcome tips for job interviews, information on upcoming seminars in the field of accounting and advices about form of CVs [10].

Respondents requested more practical examples of accounting, more interactivity and better information and communication capabilities of the platform. Job-seeking opportunities of FASTER platform have according to their

opinion great importance and it can satisfy different needs of various groups of potential users [10].

V. FASTER PLATFORM DESCRIPTION

Developed eLearning FASTER platform will be based on architecture of common used Content Learning Management Systems (CLMS), which employs a System Semantics Engineering Methodology to develop Learning management solutions that conform to the MDA paradigm through the endless process of change. The core component of the platform will be the learning service and the learning object (Fig. 1). Systemic viewpoints of learning objects and paths are exposed in terms of web-services and the corresponding WSDL descriptions can be employed from third-party standardized SOA products. The semantics of the Learning system are then translated into adaptable Learning Management models executed by the Learning Management Engine of the platform [8].

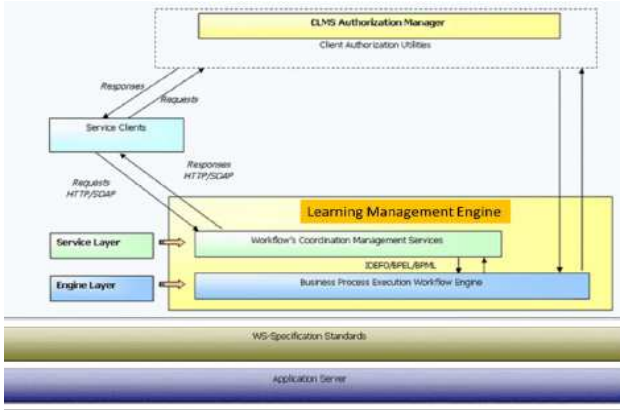


Figure 1. The architecture of XLMS Server

From the content point of view, the eLearning platform itself will be available via project official web-site (www.fasterproject.eu) with possibility to register users and login to non-public education and jobseeker environment (Fig. 2)

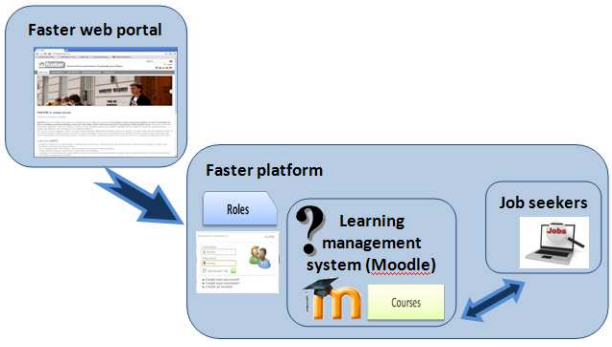


Figure 2. FASTER official web-site and FASTER platform connection scheme

The FASTER platform could be described by three parts:

- *Part A – Laboratory* containing:
  - Resources with 400 different items connected to field of accountancy (bibliography, electronic books, journals, all types of publications, legislations, etc.).
  - Glossary with the collection of 500 terms translated in all partners' languages.
  - At least 250 various links to related websites on accounting systems, national and international sites, European or national regulations and legislations.
- *Part B – eLearning courses* covering 8 subfields of accountancy:
  - *General accounting* containing fundamentals of Accounting.
  - *Accounting plan* describing a concept and operation of the accounts of General Accounting Plan, accounting cycle, accounting errors.
  - *Accounting for public limited companies (S.A.) - limited liability companies (Ltd) - Groups of companies &Code for accounting books* concerned especially on legal and organizational structure of enterprises, business foundation process, business profit, etc.
  - *Tax Accounting* with description of tax system basics, different types of taxes, their calculation and filling in statements to relevant forms.
  - *Balance sheet issues*
  - *Analysis of financial statements* - business valuation where are the Action Plan write-up, Decision making techniques, Problem resolving presented.
  - *Accounting ERP* module with aim to help users familiarize with accounting packages available on the market.
  - *International accounting standards* presenting intercommunity trade, monetary systems (European – International).

Within this part besides study material user could find FAQ, practical examples, use-cases, self-assessment questions and tests. These results will be available to employers announcing in part C to make easier find really theoretical and also practical skilled employee.

- *Part C – Jobseeker in Accounting* - part specialized on announcement of job offers, with uploading curriculum vitae, direct communication between employers and job seekers, arrangement of interview, upload of videos options.

As follows from user requirement analysis done by all project partners and presented above, there is not necessity to have access to all parts of platform to all user target groups. Fig. 3 describes our idea of different access permission.

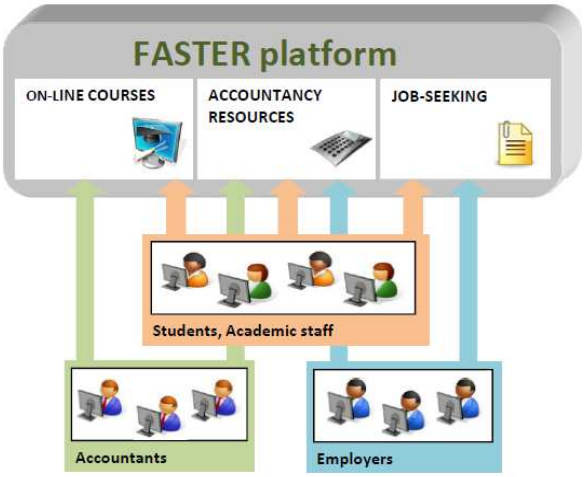


Figure 3. FASTER platform parts and their accessibility for different target groups

VI. CONCLUSION

All project consortium countries (i.e. United Kingdom, Greece, Bulgaria, Poland and Slovakia) are countries with different types of economy and different taxation and accounting legislation systems. Therefore, engaging partners from these countries, guarantees that FASTER outcomes will meet multiple EU needs in accountancy education and employment. Another parameter the consortium meets is the geographical coverage of EU (North, Middle and South-eastern Europe), so it has potential in increasing employment mobility and integration of job markets within involved EU member states.

Dissemination and exchange knowledge will be significantly easier and effective, while partners with a good geographic and lingual distribution are engaged in the project. Each partner organization of project consortium brings to the project the national level of knowledge in finance and accounting and also understanding of the needs of professionals in each national context. FASTER has potential to homogenize the differences in educational systems and provide information on the highest level of knowledge that professionals receive in the best possible way. Further it provides valuable practical examples and case studies in accounting topics to improve practical skills of persons dealing with accountancy. The FASTER platform fulfills the task of providing fast and accurate education for a professional accountant to work in a foreign country affected by local conditions. So the FASTER platform aims to be a modern e-learning platform for numerous users and at the same time a valuable job opportunity seeking tool for interregional exchange of professionals.

ACKNOWLEDGMENT

This work has been kindly funded by the European Commission through the LLP Project FASTER: Financial and Accounting Seminars Targeting European Regions (No. 518894-LLP-1-2011-1-GR-KA3-KA3MP). I also wish to acknowledge my gratitude and appreciation to all the FASTER project partners for their contribution during the development of research presented in this paper.

REFERENCES

[1] EurActiv, "Lisbon agenda," [online], 2011, Available at: <http://www.euractiv.com/en/future-eu/lisbon-agenda/article-117510>. [Accessed 1. February 2013].

[2] European Commission, "Europe 2020 - EU Strategy for smart, sustainable and inclusive growth," [online], 2010, Available at: <http://ec.europa.eu/commission_2010-2014/president/news/documents/pdf/20100303_1_en.pdf>, [Accessed 30. January 2013].

[3] European Commission, "Education and Training Strategic framework for education and training," [online], Available at: <http://ec.europa.eu/education/lifelong-learningpolicy/doc28_en.htm>. [Accessed 7. January 2013].

[4] L. Révészová, "Education via modeling and simulation," 2012. In MIPRO 2012: 35th International Convention on Information and Communication Technology, Electronics and Microelectronics – Proceedings 2012, pp. 1161-1165. ISBN 978-953-233-069-4.

[5] Council and the European Commission, "eEuropa 2002 Action Plan," [online], Available at: <http://ec.europa.eu/information_society/eeurope/2002/action_plan/pdf/actionplan_en.pdf>. [Accessed 14. January 2013].

[6] The Academic Ranking and Rating Agency (ARRA), "Rating higher education institutions," 2012. [online], Available at: <http://arra.sk/hodnotenie-vysokych-skol>.[Accessed 18. January 2013]. (In Slovak).

[7] J. Barta, "On Origins of Diplomas," 2010. [online], Available at: <http://www.etrend.sk/trend-archiv/rok-2010/cislo-48/o-povode-druhov-diplomu.html>.[Accessed 18. January 2013], (In Slovak).

[8] FASTER consortium, "Project official web site" [online], Available at: <http://www.fasterproject.eu>. [Accessed 12. January 2013].

[9] FASTER consortium, "Detailed action plan," internal document of project. 2012.

[10] FASTER consortium, "User requirements on FASTER platform: Survey outcomes," [online], Available at: <http://www.fasterproject.eu>. [Accessed 31. January 2013].

Choosing Free Library Software: Experiences of the Faculty of Humanities and Social Studies in Zagreb

Tea Čonč

Faculty of Humanities and Social Sciences/Library, Zagreb, Croatia

tconc@ffzg.hr

Summary

The Faculty of Humanities and Social Sciences Library relocated in 2009. The new way of conducting business meant the change of software used, as the application CDS/ISIS no longer suited the Library needs. The library needed a modern integrated program solution and had no time to wait for the implementation of both Voyager and later Aleph which was organized by National and University Library. At that point it was decided to implement the free integrated library software, Koha.

The paper brings a brief overview of the most used library programs in Croatian libraries, analyses the free software concept and discusses the possibilities of using the free software solutions in libraries. The surveys show that free software solutions are, in terms of quality and comprehensiveness, no longer behind their commercial counterparts. And as such, among the available free library software, Koha seems to be the most competent one.

The operation of The Faculty of Humanities and Social Studies Library confirms this statement. Koha provides quality support in the everyday work of the Library and enables further development of its services.

Keywords: Faculty of Humanities and Social Sciences Library, free integrated library software, Koha, library software implementation

I. INTRODUCTION

Choosing the right integrated software for business support is a decision that deeply affects the organization and its development. It is expected that the software have a set of necessary functionalities to support work processes and good capabilities for further upgrades as well, to be from a trusted manufacturer and within an acceptable price range.

The Faculty of Humanities and Social Sciences Library relocated into new specialized building in March of 2009. Doing so united all 23 separate libraries along that were scattered around the faculty classrooms. [1]

A new, modern way of conducting business in a new space demanded a change of software. Voyager, the software which was obtained by the National and University Library with intent of implementation into all libraries of University of Zagreb and later other universities in the country for the needs of integrated university library system, was not ready for use. As the Library did not have time to wait for a solution, they decided to use the free library software Koha.

II. THE NEED FOR A NEW LIBRARY SOFTWARE IN THE TIME OF RELOCATION

The automation of Library's business activities at the Faculty of Humanities and Social Studies started in 2001 with the use of CDS/ISIS software for DOS and later for Windows operating system which only provided support for cataloguing library material and lasted until 2008. To have an online public access catalog (OPAC), free software solution Web PAC, needed to be developed and implemented. [2]

That was not enough in order to make the Library with its 6 floors and 12 working hours of open access a day to function successfully. Check out needed to be automated and the library material that was to be placed in the open access needed to be computerized. That included adding new signatures according to the newly formed system of open access positioning to every item, printing signature stickers and programming RFID microchips[1] for every item as well. The total was approximately 300.000 catalogue entries into 20 separate databases. There were at least one third more library items and librarians had to

[1] Technology that uses wireless radio frequency identification.

prepare about 100.000 volumes so that the open access could function normally from the first day. [3]

The Library needed software that supported check out of material and also integrated the RFID system which is used to identify the items as well as a security system to keep track of the items and prevent theft. And the most important part was that it had to be available immediately. The Library didn't have time to wait as the relocation took place in the middle of a semester and the prolonged unavailability of the library system would have had negative consequences on the scholarly and educative functioning of the Faculty.

The National and University Library already purchased the software Voyager earlier for the needs of supporting the integrated university library system which was not fully developed to this very day. As a part of that project, the Library of The Faculty of Humanities and Social Sciences in 2006 and 2007 prepared for implementation of the Voyager system. However, at the beginning of 2008 not only was the Voyager not ready for implementation but due to its lack of support for Croatian diacritical marks, it was dropped all together by the National and University Library. At that point, the National and University Library decided to take Aleph, a substitute software from the manufacturer that in the meantime bought Voyager's manufacturer. The other Croatian university libraries in June of 2008 joined to express their dissatisfaction with insufficient transparency and bad informing regarding purchasing, implementing and later on switching the software they were to implement.

Under such circumstances, in the spring of 2008, with the suggestion of system librarians,[2] the Faculty administration decided to implement the free integrated library software Koha for the needs of the Library.[3]

Koha, the first integrated library open code software was developed in 1999. By the time the Library decided to implement it, the software was already in its 3rd version, the first one tailored to the needs of academic libraries (earlier versions were strictly developed for public libraries where Koha's development in the first place begun). Furthermore, Koha is web based software which enables simple use without installation of client software on librarian's computers which means that only requirements needed to use the software is a computer with a web browser. Thirdly, Koha stores its records in MARC21 format (same as Voyager and Aleph!), enables the exchange of records with other systems by using protocol Z39.50 and integration with other systems by using SIP protocol which enabled connection of user database and the self-check system. [4] [5]

[2] Larger libraries usually employ specialized experts for library software maintenance.

[3] Both The Library and The Faculty have good experiences with using free software. For instance, Linux (operating system), Moodle (e-learning program), EPrints (digital repository support) and WebGui (used for developing WebPages).

III. SOFTWARE IN CROATIAN LIBRARIES AND AVAILABLE OPTIONS

This chapter brings a brief review of the most widespread library software in Croatia as a part of the bigger picture that yielded the Library's decision to implement. There will also be some mention of COBISS, the most widespread software in the region.

CDS/ISIS [6] is nonprofit library software developed by UNESCO. The software itself is free of charge but it is not an open source software.

The libraries of the Faculty of Humanities and Social Studies used this software since the beginning of automation back in 2001, all the way to 2008 when they switched to Koha. Although it only has cataloguing capabilities, many libraries use it as it is free of charge. In Croatia, it is used and developed, among others, by the Croatian Academy of Sciences and Art Library which however considers replacing it, and also by the Croatian Information-Documentation Referral Agency Library.

Library software **Crolist** [7] was designed and developed in 1992 by a Croatian company based in Zagreb called Unibis, Inc., for the needs of the National and University Library. It was abandoned in 2007 when the National and University Library switched to Voyager, but still remained the software of choice for many libraries of all types, especially academic, in total more than 50 of them. Catalogue records can be searched via union catalogues created by cooperative cataloguing. The problem lies within individual cataloguing on lower levels which leads to more fractioned databases. Today, there is an international version of Crolist called Unilib and is still developed and upgraded by the Unibis company by adding new features.

Software called **MetelWin** [8] is a product of a company from Varaždin called Point, Inc., which was designed in 1993 to suit the needs of the Metel Ožegović City Library in Varaždin. Today it is the most widespread library software in both Croatia with almost 300 libraries using it as well as in the region with about a 1000 libraries using it. It is mostly used by public libraries, but also school libraries and, in somewhat smaller percentage, academic and special libraries (for example, the Library of Faculty of Organization and Informatics in Varaždin). Software has numerous modules and is constantly being developed, especially in the field of digital material database support.

The libraries are satisfied with its simplicity and good customer support.

Catalogue records of Croatian libraries that use Metelwin are searchable through a metasearch engine, from a single access point. However, this is not a real union catalogue as it is with the Crolist.

Aleph [9], software made by an American company Ex Libris, is one of the most popular and widespread commercial library software in the world. It possesses modules for a complete support of library operations as well as the ability to link up with other systems.

At the time of writing this article, in January of 2013, the software Aleph has been adopted and implemented by 33 academic libraries within the Ministry of Science, Education and Sports of the Republic of Croatia. [10] The implementation started in 2010. Every library has its own searchable database searchable via metasearch engine, but union catalogue is still nonexistent.

Koha, the free integrated web based software solution for libraries, is developed by user community. Library of Faculty of Humanities and Social Studies is the second,[4] and the largest Croatian library to adopt Koha. Today it is used by several other libraries as well, such as the Library of the Ruđer Bošković Institute and the Library of the Faculty of Electrical Engineering and Computing in Zagreb. System librarians at the Faculty of Humanities and Social Sciences formed a Koha expertise group who organizes occasional user meetings to discuss the experiences.

ZaKi software is a product made by a company from Zagreb called Viva Info, Inc and was designed in the early 1990s for the needs of Zagreb City Libraries network. It is also used by many other public libraries and some academic libraries (for example, the Library of The Faculty of Economics and Business) which adds to a total of 150 libraries. There is a union catalogue within the Zagreb City Libraries network and there are separate catalogues for other software users.

The former Yugoslav countries (Slovenia, Serbia, Bosnia and Herzegovina, Montenegro and Macedonia) and also, since 2006, Bulgaria, use **COBISS** (a total of 650 libraries), a software developed in 1987 as a joint project of the former Yugoslav republics. Today, it is being further developed by the Institute of Information Science in Maribor (IZUM). Each country has its own separate union catalogue with a high participation percentage of libraries in the country. Slovenia has already delivered its catalogue records to the world's largest catalogue database, the WorldCat, with Serbia next to follow. [11]

Of all the mentioned software, the CDS/ISIS has by far the weakest functionality. Metelwin, Crolist and Zaki are constantly being developed and updated by Croatian companies with its highly developed user network. And COBISS has an even larger and more structured regional network.

However, these are all software products that have been around for 20 or so years and have been developing on the same principles. The question remains can such a software be successfully transformed into a new generation software despite the, for instance COBISS, great progress in functionality and interlinking and the fact that both Crolist and ZaKi have their own, progressive OPAC solutions. The only truly new generation software products are Koha and Aleph, and the web based software is, however, only Koha.

Therefore, the selection for the Library software in 2009 was very limited: none of the software used in Croatian libraries was adequate and Voyager or Aleph was not yet ready for implementation. There was no union catalogue and there was no money for untested commercial solutions. Because of the relocation, the Library had to take the risk and find a solution outside the existing box. Although at that time Koha was the best free integrated library software with good chances to stay the best solution in the years to come, there were no guarantees that Faculty's resources would successfully handle its implementation and maintenance for the next years. One system librarian was not enough.

For that matter, few months after implementation, Faculty approved employment of another system librarian.

IV. FREE LIBRARY SOFTWARE

Every new advance in the development of information and communication technology and every new informatics breakthrough bring new questions and discussions to the expert community. Speaking in general, since it appeared more than a decade ago, the concept of free software[5] has made substantial progress. Beginning with a few, separate, enthusiastic projects it has developed into a worldwide movement that has several ground rules. Specifically, the software that survived is still being developed and has proven to be a comprehensive and quality alternative to the proprietary software. Authors have been evaluating and grading free software for years, conse-

[4] First was Library of Evangelical Theological Seminary in Osijek (library catalogue available at http://opac.evtos.hr/).

[5] Today there are several terms used to describe this software, with the most accurate being *free software* and *open source software*. The English language also coined the term *free and open source software* or in short *FOSS*. This paper uses the term *free software*.

quently leading to intense debates among experts about their advantages and disadvantages.

According to the philosophy of the GNU Project, free software means that the software's users have freedom. (The issue is not about price.) Specifically, free software means users have the four essential freedoms: (0) to run the program, (1) to study and change the program in source code form, (2) to redistribute exact copies, and (3) to distribute modified versions. [12] Therefore, the source code is public and accessible at all.

The user community, a collection of individuals and organizations interested in upgrading and maintenance takes care of the software ensuring constant further development. You can find out a lot about the software and its development status from the way its community functions and the state of the community should be one of the key elements when choosing software. [13]

Besides the libraries, community can consist of companies involved in commercial software development and user support. That widens the user support network and also makes the community more stable.

Academic libraries, which have always operated with limited space and budget on the fringes of larger organizations, soon saw the benefits of free software. The number of libraries that use free software, were it integrated software solutions or individual task and services solutions, increases every day. Along with that, the number of free software solutions for operating libraries and similar organizations is also rising.

The most well known software solutions for libraries, beside Koha, are Evergreen and PhpMyBibli (PMB). A survey conducted by T. Mueller in 2011 awarded these programs the highest grades among 20 other free library software solutions. The purpose of the survey was not only to create a reference point for choosing and obtaining free integrated library software but also to correct some previous misconceptions regarding free software such as implementation and sustainability risks. After filtering through 17 software programs in the first two steps, the best three programs left were evaluated based on 800 criteria points. In general, and by judging the number of criteria points satisfied, Koha proved to be the most competent software. Besides that, Koha's community was graded as the best organized and the most active one. [14]

An elemental contribution to the community is provided by librarians-system users who present their ideas about improving work functionalities to software developers who then transform those ideas into the programming code. Librarians are also included in all the other activities of the community: testing, running the documentation together with the developers (*Koha developer's wiki*), maintaining Koha's WebPages, news lettering and maintaining mailing lists, translating, organizing conferences and workshops, helping to start companies that provide support for Koha, etc.

This sort of cooperation ensures constant system maintenance and development as well as better understanding between developers and users. These changes happen faster than with the proprietary software [15], where it takes longer to get and answer user feedback.

Besides the comparison between each other, the survey also showed that all three best rated software are solid solutions with a vast array of functionalities that are by no standards behind their proprietary competitors. These are the, so called, second generation library software programs, a very accepted concept lately. Innovations in functionalities that appeared two decades after first informatization of libraries mark the possibility of managing all kinds of library material, irrelevant of the type, format and location, based on a service based architecture (applications such as interlinked Internet services), the ability to track new ways of conducting library business and a new (third) generation of publicly available online catalogues (OPAC) with advanced search options. [16]

In 2010 Yang and Hofmann compared, through 10 parameters, the functionalities of three new generation OPAC's: free software Koha and Evergreen and the proprietary software Voyager. The results indicated that Koha's OPAC is the most advanced and most innovative solution. [16] Macan, Fernandez and Stojanovski bring an even more in-depth analysis by comparing Koha and ABCD (a free software based on an older program ISIS), concluding, based on their attributes, what software is best suited for a certain type of library. [17]

As being interoperable and service orientated, these modern programs are capable of international protocol communication with other systems inside and outside of the library. They enable better user accessibility as well as better traditional and modern library service provision.

V. THE PRICE OF FREE SOFTWARE

A very important aspect of software is its price, or the financial resources the institution needs in order to acquire, implement, maintain and upgrade it. Those payments refer to either one-time payments (basic packs, implementation, bonus packs or modules and possible unscheduled interventions) or regular monthly payments (monthly maintenance fees or maintenance employee's wages).

With commercial software prices vary depending on the number of licensees a library needs. Furthermore, every extra module not included in the original package as well as any alterations to the software the library requests has to be purchased. Adding to that, purchasing and implementing software doesn't diminish the need for system administrators that will modify the system to fulfill the library needs and be at service for system users. The specificity of new generation software is their high level of modularity; the library is expected to skillfully optimize the received package to its fullest extent. This applies to Koha as well as to Aleph.

On the other hand, having free software does not mean having costless software. When it comes to free software, the institution does not pay the software license price, but it does pay the experts to implement, adjust and maintain the system. Those experts are mostly system librarians whose work is considered as regular expense. In case the library does not have a system administrator on its payroll, it can hire an outside company. Luckily, the number of companies providing free software implementation and maintenance services is on the rise.

The need for individual source code intervention diminishes with better community activity as the commonly needed functionality is bound to appear in the next software version. To be more precise: libraries that experience the need for an extra functionality can develop it by themselves or pay some company to develop it. They give back this development to community and gain something else from the community. This concept already proved its economic sustainability in the Library of the Faculty of Humanities and Social Studies. The Faculty approved the initial budget for buying software tools for preparing and conversion of the data needed before the implementation of the Voyager in 2005, but at the time of switching to a new system in 2008 and implementing Aleph in 2010, further funds were not assigned. [18]

In conclusion, it is worth mentioning another three financial aspects regarding free library software. The first one is the possibility of selecting the cheaper software solution and independence when choosing the supplier [19], especially in the time of the financial crisis that heavily affects libraries. During such times, financial investments in libraries' infrastructure are pushed aside as the libraries' primary investments are expansion of the informational sources and databases in order to retain the quality of its services despite the financial cutbacks. The other aspect refers to increase in competition on the market with the use of free software that can ultimately lead to price reduction of commercial software [20], as well as the upgrade of its functionalities (for instance the development of the new generation OPACs [21]). And finally, by choosing free software the main library mission is empowered: to serve the public by advocating free flow of information in today's world of twisted values.

VI. DEVELOPMENT OF KOHA IN THE LIBRARY

Implementation of Koha required significant time, effort and most importantly enthusiasm from the employees, among whom especially from the system librarians.

The transfer of data from 20 CDS/ISIS databases into Koha database was performed during October and November of 2009. Libraries began operating immediately, without delay, after the transfer of the data into a new single database.

Intense work on adjusting the records for relocation (new signatures, microchips, and labeling) finished in February of 2009. During that period, library services were unavailable for only three weeks. The new Library was officially opened on March 11th 2009, and the automated enrolment of new users and checkout started on March 16th. All services were immediately available to the users with a 110.000 volumes in open access. The rest of the material was situated in closed stacks and were available at request.

The user information is obtained from faculty systems, the users sign in into the library system with their unique electronic identity, the AAI@Edu.Hr. With that, they can see their overdues, extend the loan period, tag certain books and create record lists. Koha supports SIP 2 protocol used to connect with RFID technology and communication with self check stations, copying machines which are located in reading rooms, and Zebra printers for printing signature labels. Intervention in the source code executed by system librarians made all this features, later given back to the community, possible.

Koha supports protocol Z39.50 both as a client, which enables downloading library records from other libraries, such as the National University Library and the Congress Library in Washington, but also as a server which enables other libraries to download records from the Library. [22]

Using the report system that is based on SQL enquiry principle every business record is available, not only to system librarians, but also to other librarians using a user module that operates on a point-and-click principle.

The system shows that as of September of 2012 The Library has approximately 400.000 record units. With unprocessed books and bundles of journals not yet individually processed in Koha, the number of volumes is

closer to half a million. There are around 200.000 items in open access while the rest is in closed stacks, available at request. After the data were transferred from CDS/ISIS-a to Koha, approximately 3/4 of the records were given new signatures which is a direct preparation for the library material revision the library is obligated by law to do after relocation.

The freedom of adjustment of Koha's source code was also used in automatic allocation of several chains when giving new signatures in closed stacks. Code intervention enabled that development without raising extra cost.

The Library currently has over 11.000 registered users that conducted over 1.300,000 individual transactions (check in, check out, loan period renewal) over the course of three and a half years. On average, over 1.000 volumes are either loaned or returned on a daily basis. Use of journals at reading rooms and use of old and valuable books, however, are not documented.

At the beginning of 2012, the acquisition module was released and, with accordance to the Law on Fiscal Responsibility, a new system of ordering and receiving library material was formed that immensely eased the work of accounting, but also the librarians. Also a computer inventory book was introduced (again the new feature created by system librarians intervening in the source code) which eliminated the last remains of the paper based procession work in Library.

The functionality of reserving material is also ready for use and will be released shortly when all organization requirements are met.

Although Koha does not have a special module for interlibrary loan, the Library is currently working on adjusting Koha to automate this service. The solution will include linking Koha as a system for records tracking with Request Tracker, free software for monitoring requests and smoothening communication with users.

Koha's interoperability and flexibility are the most common advantages of open source software which gives hope for further development and successful linking with other software and services.

VII. CONCLUSION

This paper discusses the differences between free and proprietary library integrated software. The experience of the Library of the Faculty of Humanities and Social Studies described here can be used as a reference for libraries that find themselves in similar situations.

When choosing between free and commercial software, two key criteria emerge: price and freedom. With the purchase of commercial software, the library commits to further payment for all functionalities and modules not included in the original package as well as all modules needed later on and also pays the cost of scheduled upgrades and maintenance. The purchase of proprietary software does not even lower the need for system administrators that will adjust program to met the specific needs of the library and provide support to the system users. Free software, on the other hand, implies bigger knowledge and engagement, and taking the responsibility of the employees in charge of choosing, implementing and maintenance of the software.

As far as the freedom criterion goes, the institutions should be free in choosing what kind of technology and what specific software they use and should be able to make decisions based solely on expert criterion without any outside pressure. At the moment, most interested parties in Croatia feel that the national and university strategies of library interlinking are not conducted properly. Therefore, each library has to consider only its own interest as the real system level links (not only software level!) are nowhere to be seen. In that case, the priority is the financial interest of the institution, or in other terms, avoiding spending money on uncertain projects. The two things that matter the most is to choose format (MARC 21 for example) that will provide export of data and, some day, successful migration to the another software and secondly, to choose a software that can be linked in the future, when all requirements are met, with other systems into a unique university (and later on the national) library system.

Free software provides this option as they have a high rate of interoperability and come with support for linking with other systems. And even if there is no such support, certain functionalities, even the ones for communication with other systems can be obtained with some intervention into the source code. With proprietary software the manufacturers, for their own interest, often dictate libraries the terms of use even for other specified software. Because of that, libraries are often unable to link their system with other preferred systems and services, so they agree to manufacturer's terms.

Today, the best free library software (Koha) doesn't differ in terms of quality from its best proprietary competitors (Aleph), only that Koha is a web based software whereas Aleph is a client/server designed application. Manufacturer reliability, for several top quality products, is on similar level. Free software is being more and more backed up by companies that provide commercial development and support services so that even

institutions without any informatics knowledge can decide for acquisition of free software.

In spring of 2008, the Library of the Faculty of Humanities and Social Studies chose, out of several available options, software solution that was only in its first version adjusted for academic libraries and which was still not tested enough in Croatia. By doing so, the library took a certain risk as there was no time, because of the relocation, for detailed analysis and projections. Luckily, Faculty management understood the need of another system librarian workplace.

This way the Library showed that even in a small library (and informatical!) community such as the Croatian one, one should always test new possibilities, think globally and find new solutions that surpass local horizons. That kind of attitude yields success.

ACKNOWLEDGMENT

The author wants to thank Marijana Glavica and Dobrica Pavlinušić, system librarians at Library of the Faculty of Humanities and Social Sciences in Zagreb. They helped a lot with the article and, altogether, made this story possible.

REFERENCES

[1] For more information about relocation and the new library building see M. Jurković (ed.), Knjižnica Filozofskog Fakulteta□: 2009., Zagreb: Filozofski fakultet Sveučilišta, 2009. In this publication are available some short information about choosing and implementing Koha in 2008.

[2] M. Glavica, and D. Pavlinušić, "Building WebPAC for Faculty of Philosophy Libraries: experiences and lessons learned," presented at the Carnet Users Conference, Zagreb, 2002, unpublished. Available at http://cuc.carnet.hr/cuc2002/cuc2002cd/pdf/h3.pdf (accessed February 4, 2013)

[3] M. Glavica, and D. Pavlinušić, "Sveučilišni knjižnični sustav i Knjižnica Filozofskog fakulteta", 2011, unpublished. Available at http://www.scribd.com/doc/72909072/Koha-FFZG (accessed February 4, 2013)

[4] Glavica, Pavlinušić, 2011. (accessed February 4, 2013)

[5] More about Koha's life journey and present issues see at webpage: "History," at Official Website of Koha Library Software. Available at http://koha-community.org/about/history/. Also, see article J. Ransom, C. Cormack, and R. Blake, "How hard can it be?□: developing in open source". The Code4Lib Journal 7(2009). Available at http://journal.code4lib.org/articles/1638. Best online source for Koha is "Official Website of Koha Library Software". Available at http://koha-community.org. (all accessed February 4, 2013).

[6] Acronym CDS/ISIS stands for "Computerized Documentation System – Integrated Set for Information Systems". For more information see "CDS/ISIS | United Nations Educational, Scientific and Cultural Organization." Available at http://www.unesco.org/new/en/communication-and-information/access-to-knowledge/free-and-open-source-software-foss/cdsisis/ (accessed February 4, 2013)

[7] For more information on Crolist see Ž. Rajter, and L. Jelenković, "Od konceptualnih modela preko OPAC-a treće generacije do sljedeće generacije knjižničnoga sustava = From conceptual models over the third generation OPAC up to the next generation of library system," Arhivi, knjižnice, muzeji□: zbornik radova, 2013., in press. Available at http://www.unibis.hr/dokumenti/OD%20KONCEPTUALNIH%20MODELA%20PREKO%20OPAC.pdf (accessed February 4, 2013).

[8] More information available at "Metelwin programski proizvod za knjižničarstvo." Available at http://library.foi.hr/digi/metelwin/ (accessed February 4, 2013).

[9] For more information on software see "Aleph: kratki pregled," 2011, unpublished. Available at http://iks.nsk.hr/download/ALEPH-kratki-pregled.ppt (accessed February 4, 2013).

[10] List of the member libraries see at „Članice integriranog knjižničnog sustava," 2012, unpublished. Available at http://iks.nsk.hr/clanice-integriranog-knjiznicnog-sustava (accessed February 4, 2013).

[11] More information about COBISS see at „COBISS.net." Available at http://www.cobiss.net/ (accessed February 4, 2013).

[12] „Philosophy of the GNU Project" at „GNU Operating Systems." Available at http://www.gnu.org/philosophy/philosophy.html (accessed February 4, 2013).

[13] T. Mueller, "How to choose a free and open source integrated library system," OCLC Systems & Services 27 (2011), no. 1, pp. 60 and K. Yuen, "Can this economy support library changes to an open source software system?" LIBR 287 (2009). Available at: http://eprints.rclis.org/13755/1/YuenOSS287Paper.pdf (accessed February 4, 2013).

[14] S. Q. Yang, and M. A. Hofmann, "The Next Generation Library Catalog: A Comparative Study of the OPACs of Koha, Evergreen, and Voyager," Information Technology & Libraries 29 (2010), 3, pp. 141.

[15] Y. Wang, and T. A. Dawes, "The Next Generation Integrated Library System: A Promise Fulfilled?" Information Technology and Libraries 31 (2012), 3, pp. 76–84.

[16] Yang, Hoffmann, 2010, 149.

[17] B. Macan, G. V. Fernandez, and J. Stojanovski, „Open source solutions for libraries: ABCD vs. Koha", Program (2012), in press.

[18] Glavica, Pavlinušić, 2011.

[19] Mueller, 2011, 73.

[20] Yuen, 2009.

[21] Yang, Hoffmann, 2010, 141.

[22] Glavica, Pavlinušić, 2011.

Development of Information Competencies via Process Modelling and ICT Integration

L. Révészová
Department of Applied Mathematics and Business Informatics,
Faculty of Economics, Technical University of Košice, Slovakia
libusa.reveszova@tuke.sk

Abstract - Today it is very important to prepare students for challenges of knowledge economy. It is necessary for all graduates to understand functions, possibilities, advantages and disadvantages of the information and communication technology, modelling and simulation. However informatics, modelling and the use of ICT create a very extensive field for studying. How can we integrate the above mentioned targets into the education? The article summarizes the results of the research of information competencies of secondary school graduates who entered the first grade at the Faculty of Economics, Technical University of Košice. The main issue we are dealing with is how to develop information competencies of our graduates - the future advanced users - via integration many aspects of the ICT use and process modelling into the education. We developed an innovative manner of education based on modelling, simulation and on a well known opinion that "no kind of teaching can replace the personal experience". In the proposed innovative way of teaching we are trying to transform problem solving process into an active process of cognition. The paper deals with the experience in teaching and it also discusses the base of knowledge and methods which have been implemented into the subject Informatics II.

I. INTRODUCTION

As we work at the Faculty of Economics, we try to gain inspiration for innovation of education look at our problems from the perspectives of economists and managers. Many times it's interesting and useful.

According to the development and current trends in knowledge society it is necessary for educational institutes not only to be "educating" ones but also to be "learning" ones. The ability to work with information and its transformation into knowledge can massively effect further development. Anyone who works in an educational institute ought to have enough information and intellectual capital, moreover accept technologies development for prompt reactions to "customers needs", modify their "products and services" due to the latest demand and requirement.

As well as in the economic sphere also in the school environment it is necessary to count with a minimal stability and certainty. The only thing that is certain in today's world is a permanent change. For a successful existence in this turbulent environment it is very important to learn how to cope with both changes and the quickly transforming world. If anyone wants to compete on the market nowadays, amongst so fast growing competition, the ratio of their change has to be higher than the ratio of changes in the environment around them. A good position can only be the result of a quick adaptation to current needs and changing requirement [20].

II. SELECTED RELEVANT MANAGERS STRATEGIES

In managers' practice there is the ability of effective usage of knowledge called knowledge management. We implement knowledge management into educational process as a complex, consisting of three basic pillars - human's abilities, processes and technologies in their mutual relations and connections. In our attitude we try to use a so-called holistic model of knowledge management. This tool allows us to profit from synergic effect.

The application of this model consists of the following basic steps: learning before, learning during and learning after a task or action.

In this paper our point of view is based on previous three steps and on a well-known strategy Business Success Rules - 6Ws of corporate growth [13]. It helps us find an unlimited way be inspirited and to innovate!

The success of a modern institution, organization enterprise, according to [13], is based on clear definition of the six following areas, shown in Fig. 1.

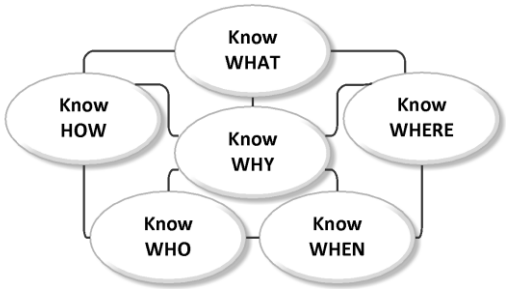


Figure 1. 6Ws with relationships. Source [13].

We consider these parts in a wider context, not only for enterprises, as follows:

1. Know WHY seems to be the most important part. It represents vision and values, knowledge of the context and connections of specific initiatives and activities in the organization or community. It is reasonable to start "doing something" only when we have a passion for something and want to create something that we can be proud of. It is

necessary to inspire yourself and other people with a clear vision and define shared values. Then we can build our distinctive capabilities to achieve our sustainable competitive advantage by continuous developing the existing and creating new resources in response to rapidly changing conditions. It helps us focus on strengthening distinctive capabilities, processes, tacit knowledge and teamwork which cannot be replicated. The most important question is: What is the real purpose of our work/effort/business?

2. Know WHAT in the company represents the control and review of data in databases, applications, technologies, classification, test specifications and standards. In educational institution it means control and review of educational content and processes. It presents also work with important models, identifying trends and design and implementation of continuously modified solutions. It allows to gain the prolonged benefit of implementing some unique value-creating strategy based on unique combination of internal organizational resources and capabilities. It enables our institution/business to survive against its competition over a long period of time.

3. Know WHERE presents the ability to create fitted strategies and to identify suitable place for launching new or existing products, services (also in the field of education), understanding of needs of the marketplace and their translation into offers that provide needed benefits.

4. Know WHEN it is change management, it is the timing of the new "products" development, managing of preparation time, closing old products and market access. In some situation timing is everything. The customer value derives from timely delivery. Change is unavoidable, but if we can anticipate it and understand important (not only business) cycles, we can ride with change instead of being run over.

5. Know WHO identifies people with information inside and outside the institution/organization/enterprise, information network that helps create things. These are community experience and knowledge networks that are designed to be an innovative and knowledge-based engine for the institution/organization/company. The corporate vision is worthless, strategies are powerless and shared values are corrupt without the right people to execute.

6. Know HOW refers to practice/business model, ability to develop over time. It includes the best practice, trade secrets, engineering standards,... .Today it is more profitable to manage processes than people. It is better to focus not on what people do, but on how they do it. It helps establish a synergistic enterprise-wide cross-departmental, cross-company coordination of work activities which create and deliver ultimate value to customers. In educational institution it means modern attitudes, methods and manners of education [13].

III. IMPLEMENTATION OF THE SELECTED MANAGERS STRATEGIES INTO THE EDUCATION PROCESS

We draw a special attention to the preparatory phase of education – we "learn before" carefully and in detail. Then during education process we notice and register our experience - we "learn during". At the end of the courses we change the content and methods according to gained experience for the next years.

Annually we search for answers to these questions:

1. What is the set of information competencies of our graduates needed for success in the knowledge economy? What are the needed skills in the field of enterprise modelling for our students, future different level managers?
2. What are the real information competencies and students knowledge base gained during their secondary school education? How can they use these skills?

The solutions of the mentioned problems present for us core answer to the most important question "WHY".

A. Know WHY

The development of our society depends on the development of a new paradigm - knowledge economy. Knowledge economy lies in creation of added value which is based on reaping knowledge, not only on the manufacturing process. Economic success is based more on an effective usage of immaterial assets - knowledge, skills and innovative potential, than on the key source of the competitive advantage. The importance of education and usage of scientific results is increasing in all areas of life. Education and learning are becoming the key interest. Nowadays in most companies the investment into immaterial capital is an important issue.

To develop a new knowledge society in Europe it is necessary to equip young people in a better way for the labour market, improve all levels of education and training. According to Council conclusions concerning the role of education and training in the implementation of the 'Europe 2020' strategy [8]: Education and training have a fundamental role to play in achieving the 'Europe 2020' objectives of smart, sustainable and inclusive growth, notably by equipping citizens with the skills and competences which the European economy and European society need in order to remain competitive and innovative, but also by helping promote social cohesion and inclusion. Higher education institutions should be encouraged to improve the quality and relevance of the courses they offer.

The changes of the economic environment started by scientific and technical progress in the first half of 20th century and followed by the continuous speed development, massive spread and the extent of computers use as well as the extensive conglomerate of technical means and technologies, especially information and communication technologies (ICT). These are necessary in today's world mostly in more developed countries. ICT participate in making huge amount of information available, which is very important in getting new knowledge. They enable effective and targeted analysis of enormous databases. They support and enable quick, cheap and quality communication between people, institutions; they help decentralize research and manufacturing activities. They enable to use new reproduction of knowledge [12].

Based on a sharp development in ICT area and the globalization of the market the companies and organization are made to make a continuous improving of the management systems using the latest ICT. The evolution of the new products is happening daily, each day in a shortened life cycle [19].

We assume that the importance of information competencies will grow, because ten years ago the "information" work has participated on GDP in developed countries with more than 50-60 % [4]. It is expected that in year 2020 the ratio of manual workers will be only 10 - 12.5 % of the whole [5].

The reasons why we consider education in the field of information management and business process modelling for necessary are the facts given in [16]. Based on studies in more than 150 companies, more than a half of the projects of reengineering business processes either failed or were not able to reach even the minimum of the planned results. The typical obstructions and shortcomings are reluctant towards changes in organization, inertia of employees and management, incorrect choice and insufficient preparation of project team, the lack of communication, time for a precise planning, not a detailed knowledge of essentials of particular problems, estimating of unclear aims, disability to estimate an extent of projects, incorrect choice and preparation of methodology.

On the other hand, as critical factors of success can be considered the following ones: direct connection of the management, active support of the leading management, support of employees, ensuring a high rate of connected people, match of improving projects with company's strategy, strict business case study, good methodology, clear understanding of strategy not only technical aspect of changes.

That is why our effort is focused on creating modern education system for future advanced users/managers in the area of "understanding and using ICT" and "business process modelling".

B. Know WHAT

This means to provide education in accordance with practice demand, based on current knowledge basis of our students – graduates from secondary schools. As we mentioned, the important query of our research is: What are the students´ information competencies gained during their previous education?

This is the question we have been dealing with since the academic year 2003/2004. We created the questionnaire entries based on current legislation, pedagogical documentation and standards e.g. [17], [7].

Using the questionnaire administrated in the first semester on the first seminars of subject Informatics I we searched for the extent and content of compulsory education and students´ real information competencies. The questionnaire contains except the closed items (with possibility to choose answer "yes" or "no") also open items concerning information systems (IS), type of IS, life cycle of IS, modeling and the work with IS.

The total number of our respondents over the here presented period was 500 graduates of secondary schools - grammar school graduates: 455 – 91 %, business school graduates:45 – 9%; mainly from the region of Eastern Slovakia, Faculty of Economics first-year students.

Figures 2, 3 and 4 present the results of the last three years in average percentage of respondents, who worked with mentioned terms, programs or applications.

As we can see in Fig. 2 and Fig. 3 in the fields of working with text editor, operating systems, using internet services and search machines there are nearly 100 % respondents familiar with these areas. In the field of working with spreadsheet program and software for presentation also in the fields of informatics elementary terms, basics of hardware and architecture of computer, internet protocols programming languages the percentage is decreasing, in spite of the fact that these fields are a part of compulsory informatics curriculum. The surprising fact is that gradually fewer students are familiar with algorithms, basics of computer networks (e.g. topology) and databases systems.

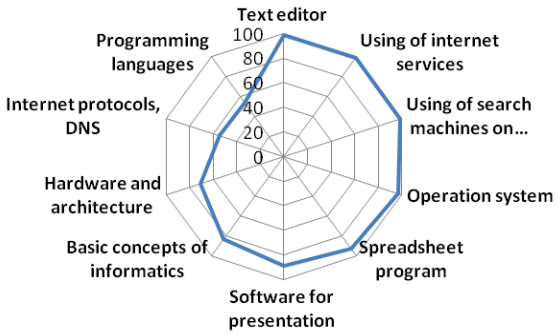


Figure 2. Percentage of respondents (more than 50%) who deal with mentioned applications, programs and terms.

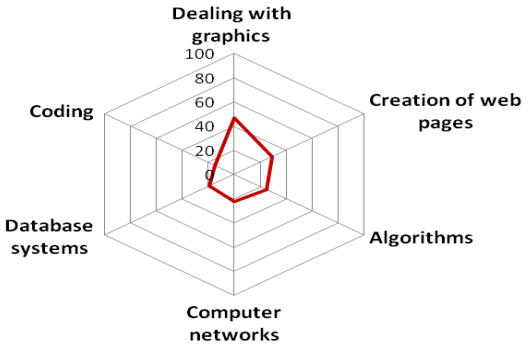


Figure 3. Percentage of respondents (less than 50 %) who deal with mentioned aplications, programs and terms.

The ability to communicate via computer network also belongs to information competencies. It means using internet services and understanding the basic rules of internet communication. Fig. 2 and Fig. 3 also present the percentage of the students, who used internet services such as email, www, ICQ and Skype. As we can see the last three years almost all secondary school students worked with and used internet services, especially the application for web browsing. Despite of that it is

surprising that the number of students who are familiar with basic rules how Internet works – internet protocols, DNS, … is very law.

As we can see in Fig. 4 the worst situation is in the area of working with IS. The evaluation of the students answers has shown that in average less than 20% of the students have had/presented experience with IS and less than 1% specified working with modelling as a method of simplifying and recognition of real objects. Less than 10% of students recognize the types of IS and less than 1 % recognize other terms.

Despite the fact the knowledge of IS is introduced in the teaching plans of obligatory secondary school subject, despite the penetration of Internet and ICT into every aspect of life - we daily use various IS - libraries, enterprises, offices, entertainment, despite the effort to integrate ICT and computer science into the education. Despite the fact that annually, nearly 100% of the students answer that they have an access to Internet and they use its services, more than 80% of the students cannot give examples of IS they work with.

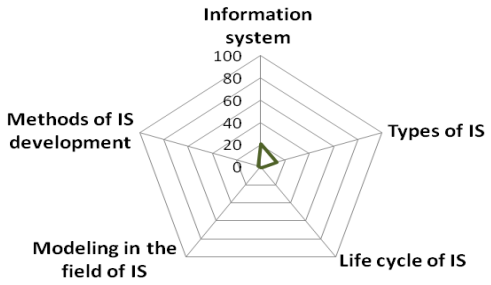


Figure 4. Percentage of respondents who are familiar with the IS field

However despite growing interest and investment in the area of ICT, important gaps remain in the current knowledge base and abilities of all types of ICT users. Today, 60% of managers believe that the current ICT opportunities limit their development and are not able to fulfil their requirements on business changes in the required time, cost and quality. On the other hand, 60% of ICT workers believe that business managers do not know what they want; they do not know what they can demand from ICT and are not able to define their requirements clearly [2].

Our new content and methods of education are based on studying the accessible sources of literature, the latest knowledge and the trend in the field of information management, business process modelling and IS implementation and integration. We also consider published experience of experts from the area of IS development, contents of foreign universities curricula, contents for educating experts in the field of "business informatics" and software engineering, our own knowledge and skills [14], [15], [18]. According to the mentioned facts we think that an advanced user/manager ought to be able to:

- notice where it is necessary to improve IS using modern ICT, precisely define users' requirements in accordance with every user's needs,
- create and control the business process model, control the IS proposal (or its parts) in various forms, also a graphical one by using common standards,
- understand the basic principles of modern technologies (e.g. data mining), understand the need of their application for keeping the ability to compete for the company,
- actively participate on strategic planning of IS - understand such terms as life cycle, architecture of IS, differ particular IS e.g. transactional, decisions supporting, executive.

From our annual detail analysis of knowledge level of the students, considering all the mentioned aspects we have decided for a concept of education in two parallel planes. Considering big differences in ICT knowledge/skills of secondary school graduates it was necessary to keep the classic procedure in teaching the subject Informatics I in a structure very similar to ECDL [7]. This seems to be vital for further correct understanding of possible work in the frame of enterprise informatics.

In the subject Informatics II we included a chapter called Information Systems. We started from the basic theoretical concepts in the Theory of Systems and modelling in general, business processes modelling using ARIS and UML. By solving practical problems, students become familiar with available features helping them work with a lot of available information.

In the second parallel plane the key theme is management of IS in the relation with modelling of the business processes. For teaching in this area we created two study materials "Information Systems for Economists" which consist of the chapters: Development in the information area, Basics terms and their meaning, Information systems, Specification of user requirements, Modelling using the Unified Modelling Language (UML) and the second one "Basics of Business Process Modelling" with content: Information and knowledge society, Business processes, Business process modelling methodology, Business process modelling standards, Business process modelling tools.

C. Know WHERE

According to our experience it is useful to use traditional face to face method of teaching – lectures and practices for teaching theoretical basics in school. In the part where students have to create a project, we use e-learning, learning-by-doing which are focused on self studying outside the school, while we are using the functionality of LMS Moodle environment for communication and consultation.

D. Know WHEN

In order to prepare our graduates for the next subjects at our faculty we have decided to include aspects concerning development of information competencies via business process modelling and using ICT into the basic informatics course in the first and second semesters. The content of the Informatics II is primarily focused on tools

which help develop information competencies needed - facilitate the processing and presentation of information.

E. Know WHO

The effective realization of this way of education is based on involvement of a team of enthusiastic teachers. It is necessary to build connections between different fields of informatics and also to build inter-subject connections among the subjects e.g. Informatics, Management, Accounting, Civil and Industrial Law, Marketing, Finance of the Firms, etc. The preparation for education is more demanding, it takes a lot of time to arrange and process all activities but the manner of the education is interesting and valuable for both – students and teachers.

Using information in the right place at the right time depends on the proper users education and capabilities. It is very important for all university students to be familiar with basic terms from the area of getting accessible information from different sources. They should understand functions and possibilities of the ICT, Internet, advantages of information systems modelling and design, because they will use all these terms in professional and common life.

F. Know HOW

Problem oriented project teaching is used as a dominant form. Students find themselves at a position of a manager of a virtual company, at a project specifying users' requirements they have to think about how to differ from the others, which innovations they should bring into the model of the main process of the chosen company. They have to have or by the method of learning-by-doing get knowledge of IS and their management so that the proposed IS would cope with more than just processing a lot of data.

The recommended structure of the user requirements project consists of four parts as we can see in Fig. 5.

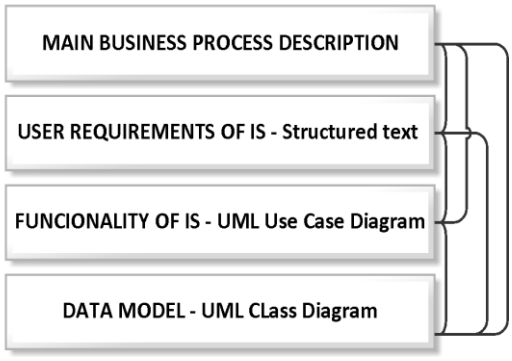


Figure 5. Students´projects structure and relationships

In the first part the students specify the core business process by text and a graphic model. On this basis they create requirements on the IS in the second part. They describe the functionality of IS in the third part using UML - Use Case diagram and the basis for the data model is a proposal of classes and their attributes using UML Class diagram.

In the proposed innovative way of teaching we are trying to transform solving problem process into an active process of cognition. The problem solution can be considered a very valuable skill [3].

The scheme in Fig. 6 shows the particular phases of a systematic problem solution, which we attempt to teach students by working out parts of the project.

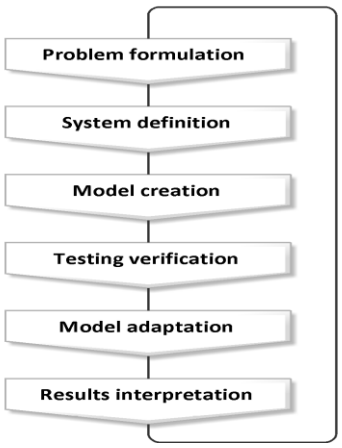


Figure 6. Process of systematic problem solution. Source [11].

The base of creation students' projects is the description of the selected business process in graphic and text forms. In connection with this part students have to work out the next parts of the project. We want students to use modern standards, notation and tool for effective work in the area of business process modelling – ARIS Express, accessible on the web page http://www.ariscommunity.com/aris-express, Fig. 7.

Figure 7. Webpage http://www.ariscommunity.com/aris-express

This free software tool is an excellent environment with a wide functionality for work in the area of business and organization modelling as we can see in Fig. 8. Installing and working with this application is a great chance for students to extend their information competencies and modelling skills. They can use "Organizational chart" to present the organizational structures, to illustrate the relationships between organizational units, roles and persons. "Process landscape" gives an overview of the value-added processes in a company. It also serves to show hierarchies. "Business process" describes the process as a sequence of events and activities. IT systems, organizational element or data may be added. "Data model" illustrates data

structures by means of data units – entities, including their relationships and properties. "BPMN diagram" enables modelling of processes according to the Business Process Modelling Notation (BPMN 2.0) - also across organizations. "IT infrastructure" shows the infrastructure of the organization. There is the possibility to present networks, including hardware and IT systems. "System landscape" shows the IT systems within the organization. "Whiteboard" is used to gather and structure ideas and tasks. It is similar to post-it notes used with flip charts. "General diagram" provides a selection of graphic elements to which any required meaning can be assigned. Here we can illustrate the content which is not covered by other models [1].

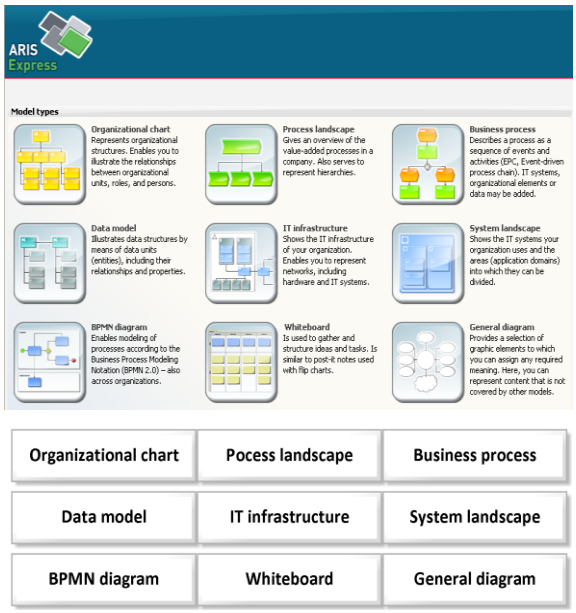


Figure 8. Description of Aris Express 2.4 Model Types Source [1]

IV. CONLUSION

„*Your ability to gain, process and use information decides whether you will belong to winners or losers*" wrote Bill Gates in 1999 [10] p. 19. Since then, the meaning of his words has been constantly increasing.

The information aspects of the management at all levels have very important function in increasing competitiveness and management quality of all economic systems. Nowadays many managers in higher position are expected to be information managers as well [6]. They have to be involved in analysis, modelling and building of the effective information systems of the firms, enterprises and organizations.

We often recognize an attitude that the primary things in development, implementation and effective application of ICT are modern technical devices, computers-hardware and software. These are surely unthinkable part of it, however the key part are people - users, as it is stated in [9]. Firstly, it depends on competent users what they will expect from ICT, information systems and informatics/computer science, what will be their requirements on information, its quality and quantity. The investments into informatization projects and new technology are often very high. The efficiency of these investments depends on users, their motivation, interest and proper education.

REFERENCES

[1] Aris Express 2.4. Available at: http://www.ariscommunity.com/aris-express

[2] J. Balun, "Business process management – pilier inovatívnej spoločnosti", eFocus, Vol. IX, No.2/2009, Digit, s.r.o., Bratislava, p. 45 – 48, (in Slovak).

[3] H. Brožová, H. and M. Houška, "Modelování znalostí", 1. ed. Praha: Professional Publishing, 2011, ISBN 978-80-7431-069-0 (in Czech).

[4] M. Dertouzos, "What Will Be: How the New World of Information Will Change Our Lives", San Francisco: HarperCollins, HarperEdge, 1997.

[5] P. F. Drucker, "To nejlepší z Druckera v jednom svazku", Management Press, Praha, 2002, ISBN 80-7261-066-X (in Czech).

[6] M. Earl, "Are CIOs Obsolete?", Harvard Business Review, Vol 78:2, 2000.

[7] ECDL Foundation, "ECDL Standards", Available at: http://www.ecdl.org/programmes/index.jsp?p=102&n=108&a=0 www.ecdl.org [Accessed 2012-01-15].

[8] EUROPE 2020' Strategy, "Council conclusions on the role of education and training in the implementation of the 'Europe 2020' strategy", Official Journal of the European Union, Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2011:070:0001:0003:EN:PDF [Accessed 2012-01-15].

[9] L. Gála, J. Pour, and Z. Šedivá, "Podniková informatika", Praha, Grada Publishing a.s., 2009, ISBN 978-80-247-2615-1(in Czech).

[10] B. Gates, "Business @ the Speed of Thought. Using a Digital Nervous System", Warner Books, Inc., New York, USA. In Czech: Byznys rychlostí myšlenky – jak uspět v digitálním věku, Management Press, Praha, 1999, ISBN 80-85943-97-2.

[11] J. Habr and J. Veprek, "Systémová analýza a syntéza", 3. ed. Praha, SNTL, 1986, ISBN 80-247-042402, (in Czech).

[12] J. Kelemen et al., "Pozvanie do znalostnej spoločnosti", Iura Edition, Bratislava, 2007, (in Czech).

[13] V. Kotelnikov, "6Ws of Sustainable Corporate Growth", In 1000ventures. Avaliable at: http://www.1000ventures.com/business_guide/business_6w.html [Accesssed. 2013-01-02].

[14] D. Paľová, "E-learning as a tool for improving ICT knowledge of university teachers", In: MIPRO 2012 : jubilee 35th International convention : proceedings : May 21 - 25, 2012, Opatija, Croatia. - Rijeka : MIPRO, 2012, ISBN 978-953-233-069-4.

[15] D. Paľová, "ICT supported education - one of steps to achieving european knowledge economy", In: MIPRO 2011: 34th international convention on information and communication technology, electronics and microelectronics: May 23-27, 2011, Opatija, Croatia. - Zagreb: ISBN 978-953-233-063-2.

[16] V. Řepa, "Podnikové procesy, Procesní řízení a modelování", 2., aktualizované a rozšířené vydání, Grada Publishing, Praha, 2007, ISBN 978-80-247-2252-8 (in Czech).

[17] State educational program; (in Slovak) Štátny vzdelávací program pre gymnázia ISCED-3a, Avaliable at: http://www.statpedu.sk/files/documents/svp/gymnazia/vzdelavacie_oblasti/informatika_isced3a.pdf [Accesssed. 2013-01-02].

[18] M. Vejačka, "Introduction to electronic banking by eLearning methods" In: MIPRO 2012 : jubilee 35th International convention : proceedings : May 21 - 25, 2012, Opatija, Croatia. - Rijeka : MIPRO, 2012, ISBN 978-953-233-069-4.

[19] D. Vymetal, "Informační systémy v podnicích, teorie a praxe projektování", Grada Publishing, a.s., Praha 2009, ISBN 978-80-247-3046-2 (in Czech).

[20] J. Wagner, "Měření výkonnosti Jak měřit, vyhodnocovat a využívat informace o podnikové výkonnosti", 1. ed., Grada Publishing a.s., Praha, 2009, ISBN 978-80-247-2924-4 (in Czech).

Research Assistance Library Service – In between Formal and Informal Learning

B. Marijanović*, S. Šimara* and J. Lisek*
* University of Zagreb, Faculty of Electrical Engineering and Computing/Central Library, Zagreb, Croatia
ferlib@fer.hr

Abstract – The purpose of this article is to present the development of a research assistance library service – online assistance in finding resources on a given topic – in the period between 2006 and 2012. With the implementation of the Bologna process in Croatia, there was a need for a more proactive involvement of the Faculty of Electrical Engineering and Computing Central Library in the Faculty teaching procedure. With its resources available, the FER Central Library sought to contribute to the quality of the way students conduct research for their papers, and therefore launched the research support service, which, over the years, developed into a successful virtual link between given research topics, mentors, students, and the Library itself. Furthermore, the virtual library support proved to be a useful non-formal learning tool. Results of an online questionnaire the service users completed are presented, and the future plans for further improvement of the service are set forth.

I. INTRODUCTION

Environment in which academic libraries operate is rapidly changing, and consequently, the way the libraries organize their services develops accordingly. Services need to be adapted to market demands and user needs, and library mission and objectives need to be developed in this direction. The rapid development of information and communication technology is changing the way users search for and obtain information. More and more information are available in electronic form, which leads to overabundance of sources and the inability to distinguish credible sources from those irrelevant. Abundance of information, resources and ways to get quality literature can often confuse the user. Many users are not familiar with the opportunities provided by information technology, many are not aware of the depth of the information environment; do not govern the skills and knowledge necessary for the selection of available information.

Fundamental task of academic libraries is to support the educational and research work of the members of the academic community. As stated by [1], the first principle of modern teaching becomes learning based on research, and the teaching process is based on preparing students to critically rethink, and on mastering the skills of argumentation. For this purpose, it is necessary to provide easy access to diverse sources of knowledge and information for students, as well as for teachers. According to [2], digital reference services refer to a network of expertise, intermediation and resources put at the disposal of a user seeking answers in an online/networked environment. A digital reference occurs when a question is received electronically and responded to electronically. As outlined by [3], it is important to note that in developed countries an increasing number of universities have established special units to support academic teaching, while "libraries" are increasingly being referred to as "learning resource centers". Although a growing number of libraries in Croatia have over the years developed some kind of online reference service[1], as stated by [3], there is one important aspect in the Croatian academic librarianship to which, unfortunately, insufficient attention is paid - competency of subject experts to organize information related to scientific research projects and activities of teachers themselves. In formal education environment such as a faculty, academic libraries and librarians can serve as a link between formal and informal learning, i.e. as a "service" in the course of studying and life-long learning, and they act as an important component in the process of non-formal learning. According to [4], non-formal learning consists of learning embedded in planned activities that are not explicitly designated as learning, but which contain an important learning element.

II. FER RESEARCH ASSISTANCE LIBRARY SERVICE

Faculty of Electrical Engineering and Computing (hereinafter FER) counts around 3,800 students at Bachelor and Master program lavel, and 450 Ph.D. students [5]. In the academic year 2005/2006 the Bologna Process was implemented in Croatia, and FER managed successfully to apply the guidelines and standards of the program to its curriculum.

FER Central Library, as an organizational unit of the Faculty, finds its calling and mission in providing a high-quality support to scientific-teaching procedure, as an active partner in enhancing the quality of learning outcomes [5]. With the implementation of the Bologna Process at the Faculty, recognizing information needs for new generations of students and scientists has become the

[1] For example, Ask a Librarian (http://www.knjiznica.hr/pitajte-knjiznicare/) by Zagreb City Libraries, and QuestionPoint, offered in Croatia by the National and University Library in Zagreb (http://www.nsk.hr/en/ask-a-librarian/).

primary goal of the Central Library, as the transformation of its traditional role into a new one, directed towards the 21st century study needs, became evident.

The Faculty curriculum includes writing seminar papers[2], participating in a project,[3] and preparing BSc and MSc thesis at the end of the study program. With its resources available, FER Central Library sought to contribute to the quality of the way students conduct research for their papers, and therefore launched a research support service. By the usage of the Library holdings, and the skills of information retrieving and searching the available online databases[4], the Library staff recommends literature and materials to students, postgraduates, corporate employees, etc. For the purposes of entry and processing of data related to the literature search on demand service, a database was developed, as well as a web form[5] for communicating with Library users [5].

After receiving a request, a librarian would begin to search for reference material on a given topic. In doing so, the librarian would primarily use the FER library catalogue[6], available online databases,[7] and catalogues of other libraries in Croatia [8] . The reply including information about the prepared materials (references to literature and Internet-pages or a guidance to other sources), as well as electronic sources (electronic articles and e-books), would be sent via e-mail, while printed materials would be put on hold to be picked up at the Library. Although insufficiently informed about teaching process itself, this segment of the Library indicates its striving to be as actively possible involved in the process of enhancing the quality of studying, and thus learning outcomes as well. By supporting the curriculum and helping students to successfully master the required tasks, the Library becomes a medium through which students are achieving their objectives, and thereby also acquire new knowledge and skills. As stated by [7], although FER Library users are advanced users of information technology, in practice they lack knowledge about information sources and their use, retrieval strategies, as well as ethics of quoting. Therefore, the FER library serves as a virtual place of non-formal learning, which is primarily based on information literacy.

The Library staff was caught unprepared by the first student rush: when creating the software for monitoring of teaching, the Faculty administration did not recognize the Library as a potential user of the software, and librarians did not have access to course details, such as the list of recommended materials for the given course, etc. Furthermore, librarians were not sufficiently familiar with the changes occurring in the teaching process.

Therefore, the first version of the form used for submitting requests for the Library service was an *ad hoc* paper one, and the Library staff was dependant solely on the information provided by students.

Estimating that the number of users will increase, the FER Library developed first of many versions of the online form, which was accessible through the FER Library web page. This type of web form helped librarians to locate exactly what students needed, as well as to keep track of the data relevant for assessing and evaluating the service itself. A great improvement occurred in the third year of the service, when Vice-Dean for Student Affairs initiated the integration of the Library reference service into the Faculty's Information Support Centre system (Quilt CMS[9]) [6], that is, a part of the system which relates to the Seminar and Project course. The application enabled students to choose whether they want "librarian's help in writing a seminar paper, project, etc". The request would then be automatically sent via email at ferlib@fer.hr[10]. This major breakthrough enabled [8] the Library staff to have access to a large number of data…, detailed information about the paper itself, but also the mentor.

III. QUESTIONAIRE RESULTS AND COMMENTS

For the purpose of continuous improvement of the service with the goal of meeting the user needs, and current curriculum as well, the Library has systematically conducted evaluations of the service, as well as recorded quantitative data related to it. The analysis was conducted over the period of six academic years.

[2] The aim of the Seminar course is improvement of presentation, communication and writing skills of students by providing an opportunity to discuss their ongoing work and interests with others. Students work in smaller groups on recent developments in the field of electrical engineering and information technology or computing (http://www.fer.unizg.hr/predmet/seminar, accessed 2013-01-11).

[3] Project requires finding the necessary literature, analysis of similar problems and solutions, identification of project requirements, definition of technical objectives, planning and time management, creation of alternative solutions, decision making, solution implementation, writing technical documentation and presentation (http://www.fer.unizg.hr/predmet/pro, accessed 2013-01-11).

[4] In Croatia, Ministry of Science, Education and Sports, in cooperation with Croatian Academic and Research Network and Ruđer Bošković Institute Library, formed the Centre for Online Databases, which offers access to more than forty databases. There are many journals of which full text is available, and there are others which give abstracts of the articles. Because of licensing agreements, these databases can only be used by current scientific-teaching staff and students. This is a very important tool for locating the latest developments and information in technical engineering and computing. In 2012 FER arranged a licensing agreement for IEEE database access, which will, henceforward, surely contribute to the quality of the suggested materials.

[5] At first, a paper form. An online form (with data organized in Access database) was created in the 2nd year of service, and since the 3rd year of service, a segment from "Quilt CMS" system has been used.

[6] http://lib.fer.hr/

[7] www.online-baze.hr

[8] National and University Library, PMF (Faculty of Science) Library, University of Zagreb, Ruđer Bošković Institute Library, Zagreb City Libraries.

[9] "Quilt Content Management System was developed at University of Zagreb Faculty of Electrical Engineering and Computing. Its core functionality is management of all internet and intranet contents of Faculty. Today it also includes a large number of applications that support complete education process at the institution." [8].

[10] Three graduated librarians share the work on these requests separately.

A. The analysis

In the period between 2006 and 2012 (October 1st 2012) there were 1,919 requests submitted, out of which nine requests remained unsolved (2009/2010).

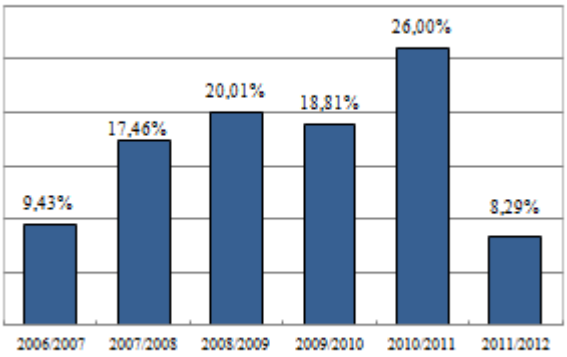


Figure 1 – Requests through the academic years, expressed in percentage

The largest number of requests was registered in the academic year 2010/2011, while the smallest number was registered in the academic year 2011/2012 (Figure 1). There was a sharp decline in the number of submitted requests in the final year of analysis, since there was a shortage of staff (due to the Library audit[11]), and the Library staff was not able to receive requests via the "Quilt CMS" system.

The higher number of requests in a four-year period (2007/2008 – 2010/2011, Figure 1) can be explained by the availability of the service via the web form, that is its accessibility through the Quilt CMS system. It is important to point out that by the latter the service received significant rise in the number of requests, but also a qualitative shift in its visibility in the formalized educational processes.

Over the years four different types of forms were developed for reference transaction, out of which two are no longer in use (printed form and the first version of the web form). During the above mentioned period requests were submitted in other ways as well (by e-mail, over the phone, or personally at the reference desk) – see Figure 2.

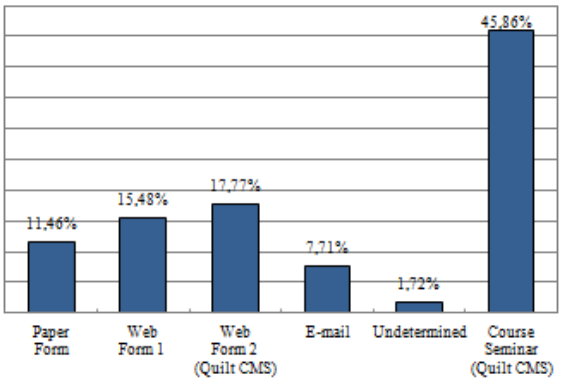


Figure 2 – Type of reference transaction in the period between 2006 and 2012, expressed in percentage

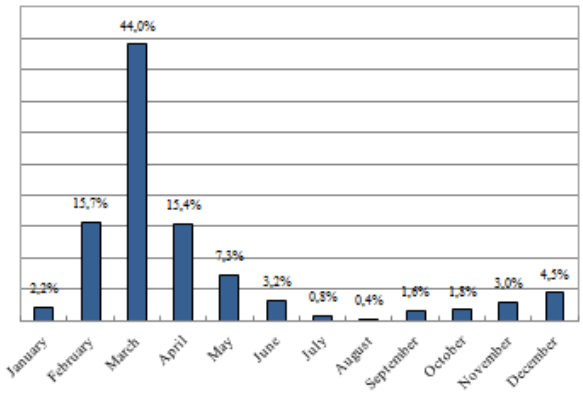


Figure 3 – Frequency of requests by months for the period between 2006 and 2012, expressed in percentage

Incorporating the service into the teaching process (Seminar course web form, "Quilt CMS" system) made the most important contribution to the rise in the number of requests - on average, more than 60% of requests per year were submitted via the "Quilt CMS" system. However, the highest increase was registered regarding requests submitted via e-mail – in the academic year 2011/2012 there was a 50% increase of such requests as compared to the previous academic year.

In the distribution of requests over the year, there is a striking imbalance (Figure 3) - third (44%), second and fourth month of make 3/4 of the total number of requests.

The users were asked to specify the type of paper they need librarian's assistance with. Figure 4 shows different types of papers noted by users. Most requests were related to seminar papers, graduate papers and bachelor thesis.

[11] Running an audit of the Central Library, as well as twelve department libraries of the Faculty

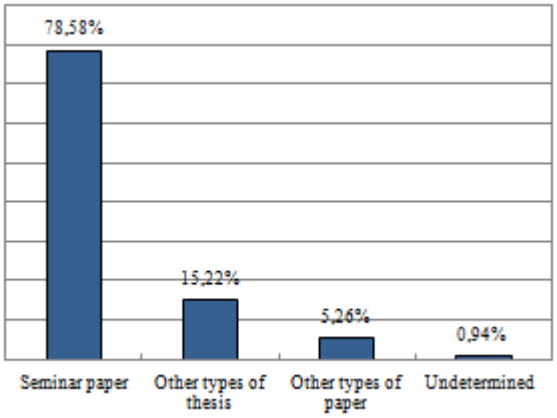


Figure 4 – Types of papers users need assistance with, expressed in percentage

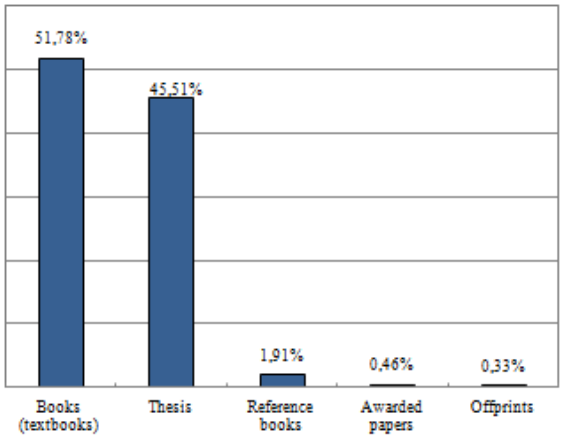


Figure 5 – Recommended FER Library holdings, according to type of materials, expressed in percentage

1,180 unique topics (formed by users and/or mentors) were attached to the total number of requests submitted (1,919). Along with a topic, key words could be attached to the request as well. Key words were also specified by librarians; however, up until January 2011 it remained unspecified who attached key words to a specific request.

There were 6,377 unique key words attached to the submitted requests. The most frequently used terms do not appear in large numbers, which proves the uniqueness of topics and requests.

Out of the total number of requests, in 63.11% librarians suggested at least one title of printed materials belonging to the Library holdings. Therefore, for 2/3 of requests, librarians suggested sources from the Library holdings; 2,170 unique titles (which makes less than 10% of the total Library holdings) were recommended, that is, the Library holdings were recommended 5,438 times altogether.

According to type of library holdings (Figure 5), most (51.78%) printed materials were textbooks, as expected, since textbooks make up the largest part of the Library holdings. Master thesis and doctoral thesis were recommended in a slightly smaller percentage (45.51%). A third of suggestions regarded external sources (links, digital materials available via online databases, or otherwise).

There were 2,181 recommended documents in digital format (pdf), out of which most are scientific papers from credible online sources (databases of peer-reviewed literature).

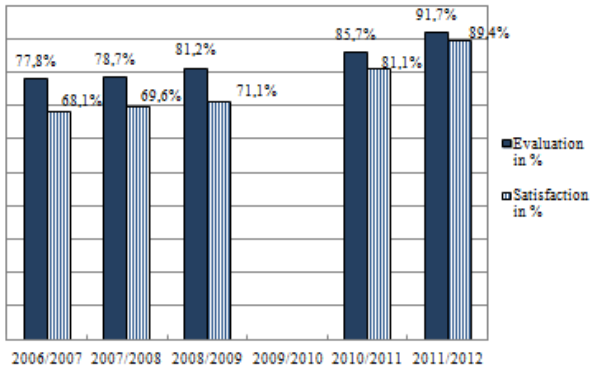


Figure 6 – Evaluation and satisfaction by the service through the academic years, expressed in percentage

B. User and mentor feedback

In 2007 the Library staff asked the Faculty professors (specified as mentors in the submitted requests) to evaluate the Library service – the sample consisted of 84 examinees. Only 34 of them (40.05%) filled in the feedback form. The questionnaire consisted of four questions, and the feedback showed that all examinees fully support the service. Furthermore, they believe that the service was helpful to students and that it serves as a support to the teaching process of the Faculty.

Over the period between 2007 and 2012, the Library also assessed the service based on user feedback (with the exception of the academic year 2009/2010). Figure 6 shows user evaluation of the service and satisfaction of the user experience. There is a positive trend in both evaluating the service and satisfaction with it. Annual average positive growth in evaluating the service is 3.5%, while annual average positive growth in satisfaction with the service is 5.3%.

It is interesting that, although that year the service „lost“ its most attractive method of submitting the requests (Seminar course web form), the users continued

to evaluate the service as extremely high, even more than before (91.7%, Figure 6).

Users have mostly (54%, Table I**)** used the service only once, that is, they submitted only one request during the analyzed period. Among those users who repeatedly used the service (46%, Table I), the most frequent ones are those with two (20.9%) and three requests (14.5%) submitted (Figure 7). Given the number of repeated users, it could be concluded that the service has met the user expectations (if it is considered that multiple use of the service indicates high satisfaction rate).

TABLE I
ONE-TIME USERS AND REPEATED SERVICE USERS EXPRESSED ACCORDING TO THE NUMBER OF REQUESTS AND IN PERCENTAGE

Users	(N)	(%)
One-time users	1036	54.0%
Repeated users	883	46.0%
Total	1919	100%

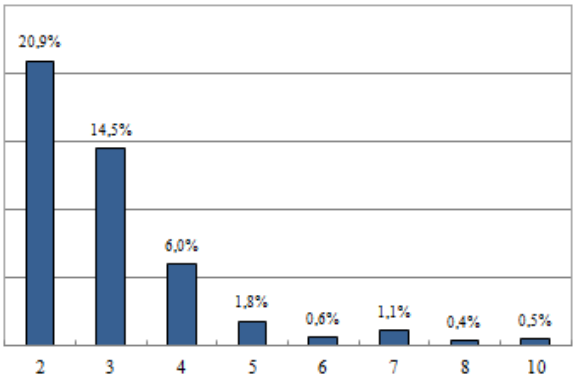


Figure 7 – Recurrence of requests, expressed in percentage

IV. CONCLUSION

FER Central Library has recognized the need to replace traditional services with the new online service and content, and it strengthened its patron-centered attitude with e-students as its focal-point. With its digital reference service the Library serves students, Faculty and staff by promoting and supporting the thoughtful use of information technology in finding, evaluating, and using information for scientific and academic work, and thus can be considered as a vital element in the process of non-formal learning.

The Library's goal for the future is to build up its support to distance education system by expanding the scope of services and introducing chat-based reference, a text-messaging service, as well as by building an institutional digital repository.

REFERENCES

[1] J. Petrak, and T. Aparac-Jelušić, „Knjižnice na hrvatskim sveučilištima: tradicija i promjene", Vjesnik bibliotekara hrvatske, 48, 1, p. 15, 2005.

[2] Ab. Kadir Wan Dollah, „Digital reference services in academic libraries". Kuala Lumpur : University of Malaya Press, 2012.

[3] T. Aparac Jelušić, „Obrazovanje na daljinu i podrška knjižnica", Edupoint, 29 (IV), 2004.

[4] D. Colardyn, and J. Bjornavold, „Validation of Formal, Non-Formal and Informal Learning: policy and practices in EU Member States", European Journal of Education, Vol. 39, No. 1, 2004.

[5] Faculty of Electrical Engineering and Computing, „Progress Report 2009/2010, 2010/2011", Zagreb: University of Zagreb Faulty of Electrical Engineering and Computing, 2011.

[6] http://www.fer.unizg.hr/cms [2013-04-18].

[7] J. Lisek, and B. Marijanović, „Halo, ovdje Knjižnica FER-a! (Pre)poznajete li me?", Proceedings of the 33rd International Convention on Information and Communication Technology, Electronics and Microelectronics, 2010, pp. 352-354.

[8] S. Tomić, „Upravljanje podacima u sustavima za podršku nastavi : magistarski rad", Zagreb: University of Zagreb Faculty of Electrical Engineering and Computing, 2011.

Gaming in the adolescent's population - pilot research

Janko Žufić, Dražen Kiralj
Juraj Dobrila University of Pula, Department of Educational Science, Pula, Croatia
jzufic@unipu.hr
kiralj87@gmail.com

Abstract: This article describes which games are played by pupils of the 7th and 8th grade of the Elementary schools and 1st and 2nd grade of the High schools in Istria. Games have been grouped in ten most popular genres: Action, adventures, arcades, horror, life simulation, driving and flying simulation games, sport, RTS, RPG and mmo ie. mmorpg games. For each genre there are listed examples of the most popular games. In this research the participants remained anonymous and they provided data for each genre - which games they play most often, how long they play daily. The research also went into direction to study behavior of the surveyed persons and whether some of them developed symptoms of "addiction". Over 180 pupils from 10 schools in Istria participated in this research. Received data indicates the following: Over 53,6% of pupils polled play computer games more than 120 minutes a day and 5.6% of them have developed symptoms of "addiction".

Keywords: RTS, RPG, MMO, MMORPG, online games

I. INTRODUCTION

Playing video games (Playing on a console, online games and computer games) is a common free time activity in the adolescent population. In addition to the numerous positive effects that gaming may bring (see [1]), there is also a possibility of developing a number of negative effects such as: neglecting friends and family [2], having lower academic achievements [3], having shorter attention spans [4], gaming over an extended period of time may lead to the pathological states of mind [5] and generate other problems. Due to the fact that gaming addiction has not been recognized as an addiction (in DSM-IV - American Psychiatric Association), adjusted criteria for determining gambling addiction are commonly used. The most important symptoms linked to the gaming addiction include uncontrollable desire to play games, unsuccessful efforts to control addictive behavior and continuous gaming over extended periods of time, abstinence crises when not playing, absorption in the addictive behavior and withdrawal from other pleasurable activities, school and home activities, inability to stop gaming, lying to conceal involvement in gaming [6]. It is necessary to make a clear distinction between a person who plays video games often and one that is addicted to playing games [7], [8]. In this research, the word *addiction* is used to describe a behavior that is damaging a person in several areas of life as a result of gaming. Motivation for this research was to determine how many adolescent gamers are in Istria, how many of them are gaming excessively and whether some of them developed addictive disorder.

A. *Game's genres*

In the world where development of games is progressing and where annual number of products is considerable, games are classified into 10 genres. Currently more and more games traverse the exact definition of a single genre.

ACTION GAMES - Action games are one of the oldest genres of games. Most of them are created in a 2D environment, and their development resulted in 3D, FPS, 3D platforms and adventures. In these games, the player is assuming the role of an action hero tasked with completion of certain missions. By solving the tasked mission, player goes to the next level until s/he completes all tasked missions. These types of games are currently played mostly online. Players connect to the game developer's servers and in a multiplayer environment show their gaming capabilities. The most common styles of playing are team death match (team playing), free for all (free style) and capture the flag (capture a flag from the opponent and take it to one's own base).

These games are among the bestsellers: they are sold in multimillion copies. Some of the games are: *Call of Duty, Quake, Doom, Unreal Tournament, Counter Strike, America's Army*. *America's Army* is a game that was developed specifically for the internal use of the US Army and was later adopted by gamers from all around the world through a multiplayer environment.

ADVENTURES - The player has a central role in the game in which s/he runs into puzzles and obstacles that s/he needs to overcome in order to go to the next level. Puzzles could be extremely difficult and sometimes it takes days of playing to solve a puzzle. The player uses various tools and objects embedded in the game, trying to resolve the puzzle and move onto the next level.

Until 2005 most of the adventure games were made as 2D graphics, in a cartoon like style and since then most of the adventure genres games are made as 3D. Games are mostly played offline. The most popular games of this genre are *Monkey Island, King's Quest, Broken Sword* and *Siberia*.

ARCADE - Arcade games can be a mixture of the adventure, action, SF, logical and sports games. In these games, the virtual space is distorted, physics of movement is distorted and the player receives special abilities as a result of collecting certain objects. For example, cars can fly, damages after car crashes are not

realistic etc. The most popular games of this genre are *Destruction Derby, Kula World* and *Contra*.

HORROR GAMES - It is a genre where a player is having a role of a character and s/he tries to survive in a hostile environment. Fear is an integral part of this genre. Games can be so terrifying to actually take the player to the edge of panic and fear. Sounds and melodies, as well as scenes of blood, fear, violence and death contribute the most to that state of mind. These conditions are creating a psychological obstacle to taking necessary actions needed to resolve puzzles. These types of games are not recommended for players under 13 years of age or players with heart problems and some games are forbidden in some states. Mexico has banned *Ghost Recon Advanced Warfighter 2* game, as it shed negative light on the Mexican army. Germany censors games located in the WWII (for example in *Company of Heroes* with its use of swastikas, although the gamers sometimes post them back into the game illegally), or simply bans games like *Wolfenstein*. In 2002 Greece banned all games relating to the illegal gambling. Chinese government decided to entirely ban expansion of the game *World of Warcraft – Wrath of the Lich King* with the explanation that presentation of skeletons, beast Undead and zombies break rules of their culture (after a while they came to understanding with the *Blizzard company* to change certain parts of the game and it was put back on the market). These games are mostly played by a single player in an offline mode. The most popular games of this genre are *Resident Evil, Alone in the Dark, Silent Hill* and *Amnesia*.

LIFE SIMULATION - In this genre games simulate types of life. In these games the player takes care of his/her character (making sure to feed the virtual character and enable him all living activities, etc) or of a whole community (prevent industrial collapses, demonstrations etc). Playing styles in this genre are developing quickly, games are played by players of all ages and it is mostly played online (although the offline style is also available). The most popular games of this genre are *The Sims, Second life* and *Sim City*. In 2006 of the first person who become a millionaire by selling her virtual money from the video game *The Second life* for real money [9].

FLYING AND DRIVING SIMULATIONS - These games allow players to simulate driving cars, motorbikes or flying airplanes. Simulation games are very close to the genre of sports games as they give players the emulative part of fun. In these games, players also have certain accessories that help player to get into the character as realistically as possible. Some of the accessories created by the game's developers are steering wheels, gas, brake or clutch pedals and joysticks for flight simulations. These games are mostly played online where players can compete with other players. Some of these simulations, for example flying an airplane, are used as an educational tool for flight simulations in pilot training. The most popular games of this genre are *Grand Turismo, V – Really* and *Flight simulation*.

SPORTS GAMES - Sports games are the ones that enable players to take virtual role of a particular sports player or take deeper insight into a virtual sport's situation. Game developers are offering accessories that can be used in the course of playing, such as tennis rackets or guns that players use in front of the monitor. The best known company of such accessories is EA - Games. The most popular games of this genre are *PES, NBA* and *NHL*

RTS GAMES - *Real Time Strategy* player has to develop his/her economy and army and use certain tactics to defeat the enemy as quickly and cleverly as possible. These types of games require high concentration and knowledge of a particular faction (nation, race) that they lead. Advancement depends on the player's skill and creativity in thinking of the tactic for his/her character. This type of game is played in the real time.

The most popular games of this genre are *Starcraft* and *Company of heroes*.

RPG GAMES - *Role Playing Games* are games where the player assumes the role of a character and go through the game by solving tasked problems (missions) in the frequent interaction with other NPC (non-playing characters), computer generated and lead characters. In this genre, the focus is on the story rather than on the action. The player has a full creative control over the story. The main accent in these games is put on the creation and perfection of the characters. During the game, player wins XP (*Experience Points*) that are used to further develop characters. As the main character gains experience, it becomes stronger and triumphs over stronger opponents. To move onto the higher levels, the character needs more and more XP, which in effect increases the playing time.

The most popular games of this genre are *Final Fantasy serial, Baldur's Gate, Mass Effect, Planescape: Torment, Ice wind Dale serial, Fallout, Fallout 2, Fallout 3, Diablo, Diablo 2* and *Morrowind*.

MMORPG GAMES - *Massive Multiplayer Online Role-Playing Games* are network games played by a large number of players in a combination with the RPG games. MMORPG is one of the most recent types of video games and it is also a fast-growing genre due to the large interest of players. Subscription is very common for this genre and in Croatia it usually costs approximately 100 HRK per month. The most popular MMO game is *The World of Warcraft* played by over 11.4 million people over the world, and growing [10]. It is also the most lucrative game in the history of the game industry ever produced. From the beginning up to date, the whole *World Warcraft* brand has been making $150 millions per month [11]. There are also some free of charge games such as *Rune of magic, Aion, Age of Connan, Battle of immortals, Drakensang, Star Wars old republic* (partially free). For some games it is not necessary to subscribe, but it is necessary to buy an authorized copy of the game, for example *Guild Wars 2* and *Diablo 3*.

MMORPG FPS "*Survivor*" is a subgenre of the MMORPG with the elements of FPS (*First Person Shooter*, assuming the role of the main character and looking the environment through the eyes of the fictive eyes of the character in the game. Such genres became very popular in the gaming population. First game of this genre was *3D Wolfenstein.)* This is a new genre and we expect to see more of them in the near future.

B. From optimal to excessive gaming

Playing video games starts as a result of having free time or as a need of an individual to keep up with the

trends of his peers. Gaming creates the feeling of comfort that gradually requires more gaming. When the amount of time spent gaming turns into preoccupation it is defined as excessive gaming [12]. Further amount of time spent gaming often results in neglect of the school requirements and lack of sleep etc. on the account of finding satisfaction in gaming. We can define this phase as the pre-pathological as the child is not fully conscious of the aberrant behavior and parental concern and influence are mostly inefficient. When the child becomes aware of the problem and decides to stop gaming, these attempts will fail. The child will have the feeling of being a victim of the process that is beyond its control. His behavior will start to show signs very similar to the effects caused by chemical substances. These children are entering into the zone of a high risk gaming [13]. Pathological status is seen as one where amount of time spent gaming is prolonged and attempted discontinuation is followed by crises (restlessness, irritability) or failed attempts to stop gaming. The game gradually takes over "entirety of his existence, and from the initial relief or entertainment it becomes the main activity" [14].

Warning signs of the excessive gaming and addiction, according to [13] are: a) Continuously spending more time gaming than intended; b) *Tolerance* – meaning tolerance to stimulus. More gaming is necessary to feel the same satisfaction in gaming; c) *Abstinence symptoms* – exhibited in gamers when they have no access to gaming due to technical or other problems; d) Inability to control the need for gaming or to stay away from gaming for longer periods of time or inability to stop gaming – manifested also in neglecting the basic physiological needs; e) Missing or suspending habitual activities or interests; f) Playing computer games as an escape from the problems; g) Lying about gaming – lying about the time spent gaming; h) Continued gaming in defiance of parental prohibition and awareness of the problem; i) Confrontation with friends and parents and problems in school.

C. Aim of the research

Aim of this research is to determine how often and which video games the surveyed population plays, do they game excessively and what is the percentage of the symptoms of excessive gaming in the surveyed population.

II. METHODOLOGY

A. Tasks

1. To determine which video games are the most popular/which genres adolescents play the most;

2. To analyze how much pupils play video games and what are their gaming habits

3. To determine how many games are played and for how long i.e. to determine percentage of the symptoms of addiction among the surveyed population

4. To determine whether boys are more prone to excessive gaming than girls

B. Surveyed population

Out of 184 pupils that participated in the survey, 166 pupils (81 of them were boys, 48.8%) completed the questionnaire correctly. Pupils attended the 7th or 8th grade of elementary school and the 1st or 2nd grade of secondary schools in Istria, most of them in Pula.

C. Tools

Questionnaire was created for the purpose of this research and it contained three sets of questions:

a) The first set of questions referred to demographic variables: sex, age/school grade, school

b) The second set referred to questions how much time do they spend playing video games daily, of which genre and do they buy video games.

c) The third set targeted habits of the surveyed persons. For this part of the questionnaire, a measuring tool – Scale of addiction was developed. Scale of addiction was created specifically for this research based on the adjusted parts of the scale used by Young to measure internet addiction Questionnaire [15] that she developed from the symptoms of pathological gambling cited in DSM-IV by American Psychiatric Association. Similar questionnaire was construed by Fisher in [16] and was also used by Loton [17] in his research. Bilic and Ljubin Golub [1] used adjusted Young's questionnaire in Croatia.

More details on the measuring tool "Scale of Addiction"

The scale contains eight groups of questions that correspond to the Young's [6] measures. Each group contains up to seven questions. Person can choose one of the five given options: "*I strongly disagree with the statement*", "*I mostly disagree with the statement*", "*Neutral*", "*I mostly agree with the statement*" and "*I fully agree with the statement*". Each answer rate as following: "*I strongly disagree with the statement*" scores 0 points, "*I mostly disagree with the statement*" scores 0.25 points, "*Neutral*" scores 0.5 points, "*I mostly agree with the statement*" scores 0.75 points and "*I fully agree with the statement*" scores 1 point. An arithmetic mean for the whole group is calculated and if it is lower than 0.5 the group receives 0 points, if it is 0.5 or higher than the group receives 1 point. Each group contributes to the final score by either 0 or 1 point. The final score is formed as a linear combination of answers from each separate group of questions and higher result on the scale reflects higher inclination to the gaming addiction. Young suggested that a person is addicted if it scores 5 points or more, and we added also another variable of gaming for more than 42 hours per week. In our *Scale of Addiction* we introduced the term "*Preoccupation, occupation or excessive gaming*" for those who scored 3 or 4 points and spent gaming more than 28 hours per week, thus keeping the possibility of comparison with previous researches and also introduced a more detailed distinction.

Questions that Young posed [15] were good if they applied to persons who were not addicted. However, addicted people very often are not sufficiently objective when it comes to self-evaluation and they are also prone to having typical defensive mechanism: going from denial to false justifications for their behavior. Additionally, many addicts feel guilty and lack the self control to stop it independently. Addicts are concealing their dependency or refusing to admit it to others. With that in mind, we have made two upgrades to this research

questionnaire: in the first upgrade, instead of one question, we have posed a group of questions and in the second upgrade, we have formulated some questions more carefully, for example instead of asking "Do you overreact on criticism or teasing", we have formulated it as "Friends who know me will say that I overreact on criticism or teasing"; instead of asking "Do you stay away from extracurricular activities, miss school and have worse grades?", we have formulated it as "Friends who know me will say that I am staying away from the extracurricular activities, miss school more often and have worse grades". These changes have been introduced to motivate the surveyed people to be more objective in their evaluations. We are aware of the fact that results cannot be fully compared to other researche projects, but they are more objective.

D. Method

This research was conducted in January 2013 as an online poll. Surveyed persons participated anonymously and voluntarily. In the general directions given by the attending teacher, it was explained that the goal of the research was to obtain data relevant to the time spent playing video games. In order to achieve an objective self-evaluation of the surveyed persons, the intention of determining gaming addiction was not mentioned. The Questionnaire was given to each person individually and time for completion was not limited. Surveyed persons could withdraw from the survey at any time. Data was collected on the server and if a person withdrew from the survey or turned incomplete questionnaire, data for that participant was not recorded. Surveyed persons had an option to leave their e-mail address and receive a feedback.

III. RESULTS AND DISCUSSION

Results on how many persons play certain genres, which video games and how many minutes per day are shown in the Table 1.

To our knowledge, this is the first research conducted in Croatia relative to the number of adolescence playing specific genres and time they spend gaming.

Results on the number of surveyed persons playing video games are in line with our expectation. All boys play video games, while 16.5% of girls do not play games. That number is relatively high as all schools that surveyed girls attend have computers and they use them regularly. In addition to that most of them have a computer at home. 10.7% of girls that do not play video games attend elementary school and 30% attend high school. These results could be interpreted as an earlier maturing of girls and need to not "waste time", but there are also other arguments presented bellow.

TABLE 1. HOW MANY PEOPLE PLAY VIDEO GAMES AND HOW MUCH TIME THEY SPEND ON EACH GENRE.

Genre	Playing games			Average time of gaming [minutes] / per day					
	(A) Total [%]	(B) boys [%]	© girls [%]	(D) Of those who play	(E) Of the total sample	(F) Boys who play	(G) Boy total	(H) Girls who play	(I) Total girls
Action	54,8	74,1	36,5	67,7	34,7	81,0	60,0	29,0	10,6
Arcade	41,0	40,7	41,2	31,0	12,5	32,5	13,2	28,7	11,8
Horror survival	31,3	45,7	17,6	50,5	15,8	58,2	26,6	31,3	5,5
Life simulations	67,5	66,7	68,2	44,8	24,3	36,6	24,4	35,4	24,2
Driving and Flying simulations	49,4	65,4	34,1	39,3	19,4	47,3	31,0	24,8	8,4
Adventures	39,8	23,5	55,3	23,0	9,0	25,8	6,1	21,4	11,8
Sports games	52,4	69,1	36,5	51,6	27,0	63,1	43,6	30,7	11,2
RTS	28,9	40,7	17,6	62,8	18,1	79,4	32,3	25,8	4,6
RPG	32,5	44,4	21,2	54,0	17,6	52,2	23,2	57,6	12,2
MMORPG	25,9	34,6	17,6	58,7	15,2	65,7	22,7	45,7	8,1
Don't play any genre of video games	8,4	0,0	16,5	-	-	-	-	-	-

Boys (in percents) are playing *Action games, Horror and survival games, Driving and Flying simulation games, Sport's games, RTS, RPG and MMORPG* games considerably more than girls, with a statistically significant difference. *Arcade and Life simulation games* are played equally by boys and girls while girls play *Adventures* more than boys.

When examining the average gaming times for certain games and time spent on the computor, it is necessary to mention that column D, F and H present data only for persons who play games while columns E, G and I show average data pertaining to all participants in the survey.

In general, higher percentage of boys play video games, spend more time gaming and play more genres for longer periods of times. Some of the reasons could be that: boys are more prone to competing than girls and winning and proving themselves is important to them [18], boys are more prone to playing video games than girls [7] or even that "boys are more inclined to pathological gaming than girls" [19][20]. All presented arguments confirm results obtained by this research, for example high involvement (over 40%) of boys in eight out of 10 genres of games.

Results of this research are in line with similar published research data [1].

Results of this research can also be presented by another criterion. Distinction by time spent gaming [in minutes per day]: a) Do not play video games: 9.0%; b) play from 0 to 120 minutes: 36.7%; c) play from 120 to 240 minutes: 24.7%,; d) play from 240 to 360 minutes: 9.6% and e) play more than 360 minutes per day (42 hours per week): 19.3%. Young revealed similar results and noticed that internet addicts (internet - not video games addicts) spend more than 38.5% hours per week on the internet.

Times under a), b) and c) are acceptable and if they stay the same, they can have a positive impact on the surveyed persons for example they can be used as positive learning – some of the simulation games, world of imagination and many more other possibilities since they enhance speed, focus, evoke thought and associative processes as well as hypothetical thinking, enhance motor skills [21], concentration, memory and control [22], hand-eye coordination, i.e. harmonization of senses and motorics, in general coordination of movement, sharpness of sight, orientation in space and even intelligence [17].

Times under d) and e) are very high and in combination with other factors may indicate excessive gaming or even gaming addiction.

Table 2 shows ratio of single symptoms that may occur during gaming. As previously mentioned in the Introduction, questions were based on [1] as a framework for the groups, in order to be able to compare our results with other studies with additional or rephrased questions aimed at getting more objective and detailed data.

Comparing obtained results [1] similar results have been observed as well as certain differences. Compared to other studies, the first fragment of this research indicates higher percentage of symptoms of addiction in boys 46.3% in our study, compared to 18.1% in Bilic, as well as in girls 33.8% in our study, compared to 4.2% in [1]. The second fragment also indicates an increase: in boys 57.4% vs. 18.1% and in girls 22.5% to 4.2%. In third and the fourth fragment, differences are 22.2% and 7,0% in boys and girls, respectively, in our study vs. 27.5% and 25.7% in boys and girls, respectively, in [1] and 22.2% and 11.3% vs. 23.8% and 7.2%. Fifth fragment is high in both studies 75.9% and 59.2% vs. 66.3% and 41.9%. Sixth and the seventh fragment have small proportion of 14.8 and 7.0 % vs. 20.0% and 6.6%; and 11.1% and 5.6% vs. 20.0 and 2.4%. The last, eight fragment differs in girls 22.5% vs. 11.4% while it is almost the same in boys – 27.8% vs. 26.3%.

TABLE 2. RATIO OF SINGLE SYMPTOMS THAT MAY OCCUR DURING GAMING IN BOYS (N=81) AND GIRLS (N=85)

Groups of questions	Boys [%]	Girls [%]
1. Are you preoccupied with gaming (do you think of the previous game and do you look forward to playing the next? - I loose sense of time when I play - I always try to improve myself and achieve a better result in the game - I "fear" of premature loss/end of game - Others always complain that I spend too much time gaming on the computer/console - While in school or involved in some other activities, I often think of the video games (how to move to the next level or so) - It happens sometime that I just cannot stop gaming	46,3	33,8
2. Do you feel the need to spend more time gaming in order to achieve satisfaction? - When playing video games, I feel ecstatic, irritated and relaxed. - I feel the need to spend more time gaming in order to achieve satisfaction.	**57,4**	**22,5**
3. Can you control and reduce the time spent gaming or stop gaming? - I have noticed (or my parents or friends told me) that I have neglected basic need for food, sleep, hygiene or I have let myself go etc because of gaming - I feel that world outside of the games is boring - I cannot wait to start gaming again - Because of gaming, I sleep less than usual. - My life without gaming would not be equally happy, interesting and fulfilling. - Sometime I neglect the school because of the time I spend gaming. - Sometime I feel physically and emotionally exhausted after gaming for several hours.	**22,2**	**7,0**
4. Do you feel restless, nervous or irritable when you stop gaming? - I have noticed feeling of restlessness, nervousness, irritability and aggressiveness when I did not have access to the games (no electricity, computer is out of order, when I went for a longer time somewhere where I didn't have access to the game, when I was grounded or when I tried to reduce or stop gaming).	**22,2**	**11,3**
5. Do you spend more time gaming then you initially intended? - I always spend more time gaming than I intended	75,9	59,2
6. Have you risked loosing important relationships or a possibility of getting an education because of the video games? - Sometime in school, I have problems being focused and attentive and feel restless, insecure and scared. - Friends who know me will say that I am overreacting to criticism or teasing. - Friends who know me will say that I am constantly opposing the authorities and come in conflict with my peers. - Friends who know me will say that I am staying away from the extracurricular activities, miss school more often and have worse grades. - I have noticed (or my parents or friends told me) that I have neglected my hobbies, school obligations and relationship with people that matter to me (friends, peers, even my girlfriend/boyfriend)	**14,8**	**7,0**
7. Have you ever lied to your friends, parents or someone else in order to hide how involved you are in gaming? - I have noticed (or other people told me) that I forget previous arrangements or given promises to friends, parents or other people. - I have lied to my parents occasionally (or stole money) to get the money and buy games. -I have borrowed money from friends or family, without the knowledge of my parents - If someone asks me how much time I spend gaming, I try to conceal the truth. - I lie to friends, parents or someone else in order to conceal how involved I am in gaming	**11,1**	**5,6**
8. Do you use playing games as an escape from problems, feeling of guilt, powerlessness, concern or depression? - There are problems in my family that I don't want to talk or think about. - I fear that I will not succeed in school or life. - I fear that I will not live to the expectations of others (people closest to me or people who are important to me) - I think that gaming relaxes me and brings calm from daily problems - Habitual gaming serves me to escape problems or feelings of guilt, powerlessness, concern, depression.	27,8	22,5

Boys play more often, spend more time gaming and have higher ratio of “gaming addiction” symptoms. Obtained results in this research in most parts confirm the results obtained by [1]. In total 0.8% of all surveyed persons fall into the cathegory *Occupied, preoccupied or excessive* gamers (3 or 4 symptoms in combination with gaming for 28 hours or longer per week), while 4.8% of all surveyed persons fall into the cathegory of “addicts” by Young’s definition (5 or more symptoms out of 8).

These results are consistent with the results of other studies, to mention just a few: 11.9% of surveyed persons are addicted to playing video games [23], 8% of people in the study of [5], 4.2% of people are addicted to playing video games on the console and 5% of people are addicted to the computer games [24]. Pathological patterns of gaming addiction were found in 8% of the participants in the research aged from 8 to 18 years [7], and 6,3% of German adolescents with lower academic achievements are addicted to gaming [25].

A. Limited factors

The authors are aware of some limitations of this research. One of the most important limitation is based on the fact that medical professionals have not defined “Gaming addiction”. Further limitations include a relatively small number of participants in this research, lack of information in regard to the level of self-satisfaction, lack of results relative to the lack of self-esteem, academic achievements of the participants and uniformity of the sample. Further limitations are linked to the lack of social data for the surveyed population, the questionnaire did not embed test of the “socially desirable answers” and maybe the most limited factor was the selectiveness of the questionnaire itself – the questionnaire was completed only by those who wanted to participate in the research and it was done online, thus prevented the possibility of cross-verification of the entered data. In spite of all limitations, we believe it is a good pilot-research and it can be the bases for similar researches in Croatia.

IV. CONCLUSION

Aim of this research was to determine how often and which video games the surveyed population plays, whether they game excessively and what is the percentage of the symptoms of excessive gaming in the surveyed population in the sample of pupils attending the 7^{th} and 8^{th} grade of elementary schools and pupils attending 1^{st} and 2^{nd} grade of secondary schools. The results confirmed the hypothesis relevant to the higher percentage of gamers in boy’s population than in the girl’s and to the fact that boys on average spend more time playing single genres. Percentage of excessive gaming or individuals with signs of gaming addictions among the participants in this research are relatively low, but it is not irrelevant.

REFERENCES

[1] Bilić, V., Ljubin Golub, T. (2011.) "Patološko igranje videoiraga: uloga spola, samopoštovanja i edukacijske sredine", Hrvatska revija za rehabilitacijska istraživanja, 47 (2) pp. 1-13

[2] Padilla-Walker, L. M., Nelson, L.J., Carroll, J.S., Jensen, A.C. (2009): "More than a just a game: video game and internet use during emerging adulthood" Journal of Youth and Adolescence, 39, 2, pp. 103-113.

[3] Anand, V. (2007.) "A study of time management: The correlation between video game usage and academic performance markers", Cyberpsychology & Behavior, 10, pp. 552–559.

[4] Tahiroglu, A. Y., Celik, G. G., Avci, A., Seydaoglu, G., Uzel, M., Altunbas, H. (2010.) "Short-term effects of playing computer games on attention", Journal of Attention Disorders, 13 (6), pp. 668-676.

[5] Porter, G., Starcevice, V., Berle, D., Fenech, P. (2010) "Recognizing problem video game use", Austrailan & New Zealand Journal of Psychiatry, 44(2) pp. 120-128.

[6] Young, K. (2009): Understanding onling gaming addiction and treatment issues for adolescents, American Journal of Family Therapy, 37(5), pp. 355-372.

[7] Gentile, D. A. (2009) "Pathological Video-Game Use Among Youth Ages 8 to 18", Psychological Science, 20(5), pp. 594-602.

[8] Rehbein, F., Kleimann, M., Mössle, T. (2010) "Prevalence and risk factors of video game dependency in adolescence:results of a German nationwide survey", Cyberpsyhology, Behavior and Social Networking, 13(3), pp.269-77.

[9] http://www.index.hr/xmag/clanak/virtualni-svijet-second-life-dobio-svoju-prvu-stvarnu-milijunasicu/333328.aspx viewed 13.2.2013.

[10] http://www.mmo-champion.com/content/2293-WoW-Loses-600k-Subscribers-Diablo-3-Beta-Q3-2011-Arena-Passe-Blue-Posts

[11] http://wow.joystiq.com/2008/01/24/how-much-money-blizzard-is-really-making-from-10-million-subscri/ viewed 13.2.2013

[12] Davis, R. A. (2001.) "Cognitive behavior model of pathological Internet use". Computers in Human Behavior, 17, pp. 187-195.

[13] Bilić, V., Gjukić, D., Kirinić, G. (2010.) "Mogući učinci igranja računalnih igrica i videoigara na djecu i adolescente", Napredak 151 (2), pp. 195 – 213

[14] Valleur, M., Matysiak, J. C. (2008.) "Patologije ekscesa: droga, alkohol, igre, seks – zastranjivanje naših strasti". Provincija franjevaca trećoredaca: Biblioteka Oko tri ujutro. Zagreb

[15] Young, K. S. (2004) "Internet Addiction: A New Clinical Phenomenon and Its Consequences", American Behavioral Scientist, 48, pp. 402-415.

[16] Fisher, S. (1994) "Identifying video game addiction in children and adolescents", Addictive Behaviors, 19, pp. 545−553.

[17] Loton, D. (2007) "Problem video game playing, self esteem and social skills: an online study". BA Thesis. Victoria University

[18] Van Vugt, M., De Cremer, D., Janssen, D. P. (2007) "Gender differences in cooperation and competition: The malewarrior hypothesis", Psychological Science, 18(1), pp. 19-23.

[19] Wölfling, K., Thalemann, R., Grusser-Sinopoli, S.M. (2008) "Computer game addiction: A psychopathological symptom complex in adolescence". Psychiatrische Praxis, 35(5), pp. 226-232.

[20] Chiu, S. I., Lee, J. Z., Huang, D. H. (2004) "Video game addiction in children and teenagers in Taiwan", Cyberpsychology & Behavior, 7, 5, pp. 571-581.

[21] Laniado, N., Pietra, G. (2005.) "Naše dijete, videoigre, Internet i televizija (Što učiniti ako ga hipnotiziraju)". Studio TiM, Rijeka

[22] Boot, W. R., Kramer, A. F. Simons, D. J., Fabiani, M., Gratton, G. (2008.). "The efects of video game playing on attention, memory and executive control", Acta psychologica, 129(3), pp. 387-398.

[23] Grüsser, S. M., Thalemann, R., Griffiths, M. D. (2007): Excessive computer game playing: Evidence for addiction and aggression? Cyberpsychology & Behavior, 10(2), pp. 290-292.

[24] Thomas, N. J., Martin, F. H. (2010) "Video-arcade game, computer game and Internet activities of Australian students: Participation habits and prevalence of addiction", Australian Journal of Psychology, 62(2), pp. 59-66.

[25] Wölfling, K., Thalemann, R., Grusser-Sinopoli, S.M. (2008) "Computer game addiction: A psychopathological symptom complex in adolescence", Psychiatrische Praxis, 35(5), pp. 226-232.

Compact flexible router based remote experimenting system

T. Szepe*, Sz. Toth* and Z. Gingl*
* Department of Technical Informatics, University of Szeged, Szeged, Hungary

Abstract - The rapidly changing needs of modern education can be effectively supported by today's small, intelligent devices including tablet PCs, smart phones programmable routers and credit-card sized computers like the Raspberry PI. Since there are many available sensors on the market, it is rather easy to connect real world signal to these devices like temperature, light intensity, acceleration and many more. This way an exciting, cheap, transparent and very efficient software-based real experimentation can be carried out at different levels of education.

Our solution is based on a flexible, easily programmable popular router and a universal sensor-to-USB interface with open source software and hardware plus well documented educational applications. The router provides a computer platform for data acquisition and control while also serves as a remotely accessible device through its wired and wireless ports. In our presentation we shall briefly introduce the system setup and the connection of sensors and actuators. We have developed and tested a simple but efficient temperature control example educational application that can be remotely monitored and controlled via the internet. The transparent hardware and software can be used to attract students' attention, also supports the development of creativity by allowing modifications of the software and hardware configuration.

I. INTRODUCTION

One of the main problems of natural science and engineering education is to find methods that can keep the attention while most of the required and sometimes quite a lot of information is shared. Today information exchange and processing is fast, the students want to quickly understand and learn which can be a problem with many disciplines. Computers and graphical user interface software can be very efficiently used to simulate phenomena, and although they can provide user interaction also, they still appear rather artificial and do not cover real-world effects completely. As in the case of learning computer programming, mechanical and electronic design, many other fields need a lot of practicing, investigation of real systems and real signals. Comparing lecturing and laboratory practicing is often made in the frame of economical efficiency and of course less students can be taught in a larger room, the equipment's and experimental setups are costly in most cases and they are rapidly changing due to today's short time-to-market periods. On the other hand industry, science and teaching need creative participants who not only know but even understand the processes, who can find the required solutions quickly and reliably.

This work was supported by grant TAMOP-4.2.2/A

These skills can only be developed efficiently in smaller groups, by a close interaction between the lecturer and students who can even work in a project fashion.

Since there are more and more sensors, sensor interfaces and low cost user-configurable or programmable electronic devices are available, therefore widely usable efficient experimentation can be supported at different levels of education and in different disciplines including informatics, physics, chemistry, biology as well [1 - 6].

Just to name some examples, real world signals can be monitored and by the Arduino microcontroller kit, by sensors connected to the Beagleboard or Raspberry PI single board computers – the latter is especially developed to support creative mass education – or even by simple interfacing of sensors to the audio port of a notebook computer or to a smart mobile phone. Comparing these solutions to the professional educational kits shows much lower cost and availability, much more transparent experimentation and measurement solutions. Many students can easily have their own device and set of sensors, they can change the experiments, try their own ideas therefore their creativity and practical knowledge can be significantly improved. On the other hand, it is important to teach the proper usage of sensors and devices, to estimate the reliability, accuracy and to show the limitations in the same time.

Note that most of the above mentioned devices are network-ready, so students can even share raw data or processed information, they can learn about sharing data over the wired or wireless internet and other networks. They can see that a complete device can have real (hardware) and virtual (software) parts that can provide the different instrumentation and experimentation functions and they make the whole system together.

Besides low cost single board computers, microcontroller boards, Lego robots other small, very low cost and widely available devices can be efficiently used – like routers. Routers also have open source development platforms, have wired and wireless network interfaces and USB ports. Sensor and actuator to USB interfaces can turn them into a very powerful local computerized instrument or a remote laboratory server.

In a usual educational environment there is some kind of shared internet available, which is usually provided by using a local router. For the proposed system a reprogrammable router with USB connection is required. One of possible choices for an easily reconfigurable router

is the TP-Link TL-WR1043ND, which supports the popular open source software package called OpenWrt [7]. In this way an existing hardware is used preserving the original functionality of the router, only minimal additional hardware components are needed.

The router can handle simultaneous connections from many different sources, including smart phones, tablets, or other pc-s through wifi, or connected to internet it allows remote laboratory experiments. The flexible sensor-to-USB interface is easy to duplicate, hence it is completely open source including schematics and microcode, this means it can be reused to fully customize on purpose. The device uses cheap and widely available components to make, that means even students can reproduce or redesign it for themselves. It does not contain complicated BGA packaged chip like in the famous Rapsberry PI or BeagleBone which requires expensive tool to replicate and provides cheaper alternative to the commercially available counterparts [8, 9].

In the following we demonstrate the use of the popular TP-Link TL-WR1043ND router and the open source hardware and software sensor-to-USB interface in a remotely controlled experiment.

II. IMPLEMENTATION

A remote experimentation system has to be connected to the Internet for a long period of time and it must be capable of providing a static homepage with parameter handling capabilities. Usually it means a desktop PC will run a webcontainer and the rest part of the experimenting setup will be connected to it. Such a configuration is usually not enough cost and power efficient. Fortunately other solution is possible using existing infrastructure. In many high schools we can find wifi capable routers in or near the class rooms and some of them already equipped with a USB port. Running the proper firmware and software configuration a router can be used as a server device without loosing its original capability to operate as an access point.

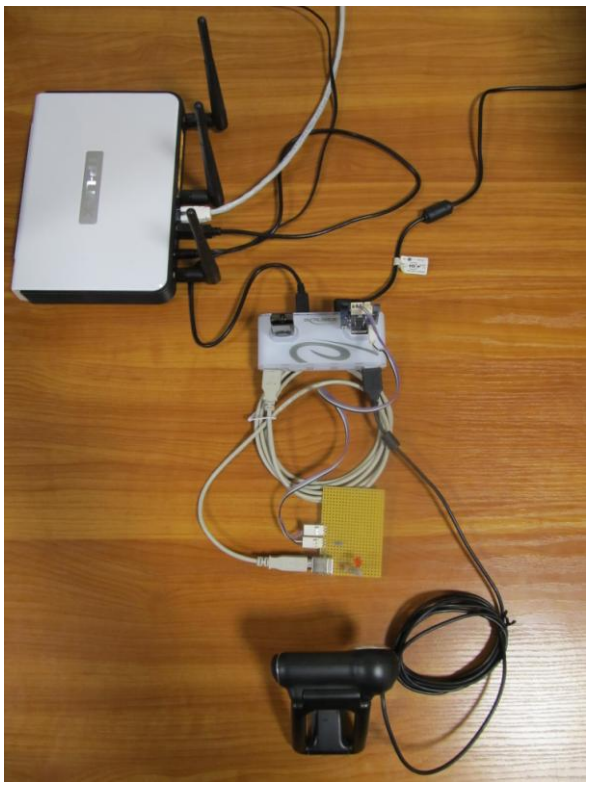


Figure 1. shows the hardware components

A. Hardware components

Using a commonly available router with USB port and good price/performance value was a top priority in hardware selection. It is advisable to choose a model supported by OpenWrt [7] or similar actively maintained project like Gargoyle [10]. Our choice for implementation was the TP-Link TL-WR1043ND router with three 802.11n capable wifi links, four Gigabit ethernet ports and a USB port. We used an active USB Hub in order to connect multiple USB devices with more than 3W of cumulated power usage. Built in flash memory was extended with a one gigabyte pendrive and a Logitech Quick Cam Pro 9100 webcam served as a video capture device. The outline of the connected hardware components can be seen in Fig. (1).

The Edaq-24 open source sensor-to-USB interface is based on the C8051F350 microcontroller from Silicon Laboratories [11, 12]. It incorporates a high precision 24-bit analog-to-digital converter, voltage reference, oscillator and rich set of digital peripherals in a single chip. Using a special arrangement of only a few passive analog components and ultralow on-resistance logic level MOSFETs most sensors can be connected without any additional circuitry. For low level signals the on-chip programmable gain differential amplifier provides the required signal scaling. Sensors that can be directly connected include thermistors, photoresistors, photodiodes, accelerometers, Hall-effect magnetic field sensors, Wheatstone bridge pressure sensors, thermocouples and many more. The very high resolution and low noise makes possible to measure extremely small signals, for example a simple Hall-sensor can be used to measure the Earth magnetic field quite accurately.

The device has digital output pins also, with the option of pulse width modulation (PWM) outputs to drive LEDs, small loads or even motors and large current actuators if external transistors are used.

The device connects to its host computer via a popular USB-UART chip (FT232RL). The chip was picked for its reliable operation and widely available drivers. Note the most Linux systems even do not require installation of drivers.

For users who do not have the possibility to build their own Edaq-24 sensor-to-USB interface some lower performance options are available like the Arduino Uno and Nano [13] or alternative setups can be developed with limited measurement capabilities using a sound card [14 - 16]. These solutions may require additional circuitry and even external amplifiers to allow the use of various sensors.

B. Software components

All software needed for the experimenting system can be categorized into three groups. Starting from the bottom

to the up first category is the C code running on the micro controller, which handles the measuring procedure wit the ADC, the PWM output and all the basic communication on the USB channel. Second category is the server-client architecture with both part running on the router. Their task is to keep the USB resource available for communication and serve as many queries as possible to the Edaq-24 board. The third category contains the operating system of the router which is a linux kernel based firmware called OpenWrt.

On the Edaq-24 a simple communication protocol was implemented. Basically all valid command have to start with a "@" character and the following character will specify the command type. If necessary a third byte has to be sent as a parameter. To set the output to binary mode "@t" has to be sent. If we want to measure the second channel "@C2" is the command and a simple measurement will be done after "@s". In the last case the measurement can be read out as a byte sequence. The source code for the Edaq-24 and other custom written source codes and configuration descriptions are available in [17].

A daemon server application was developed to run as a background process on the router to preserve the hardware resource for USB communication. Keeping in mind that many queries are possible in a short period of time, it is not a good idea to open a writer and a reader stream every time when communication needed, therefore the server process will open them only once and keep them alive. Communication is possible with the server process through a small portion of shared memory, where the implemented protocol is similar as the one on the Edaq-24 board with some extensions. Additional command characters are used to make direct messages possible with the router like terminate the server process with "@X". Thread safety for the shared memory is guaranteed with a semaphore that ensures the given command was processed by the server thread. Passive waiting for a new command is achieved by signaling using OS call and whenever possible the server thread waits in sleep mode. The shared memory is initialized by the server with its own process id which is later used by the client thread in the signaling process.

OpenWrt firmware was configured using the detailed description found in [7]. To reproduce a working configuration some additional install steps are required. Most importantly a pendrive should be mounted automatically at the booting time of the router where most of the installed software should be deployed. In order to access them as the router own packages, all deployed files should be linked to the bin folder of the root file-system. In this way the router file-system will be extended. In order to run the web interface lighttpd and perl script interpreter should be installed. The webcam requires mjpg-streamer and the USB communication may require libusb and libftdi libraries. All important configuration steps can be found in [17].

III. DEMO APPLICATION

To demonstrate one possible use of the proposed setup, a demo application was designed with which basic control algorithms can be examined. The aims for the controllers are to reach and maintain a specific temperature by applying a certain voltage across a heater resistor.

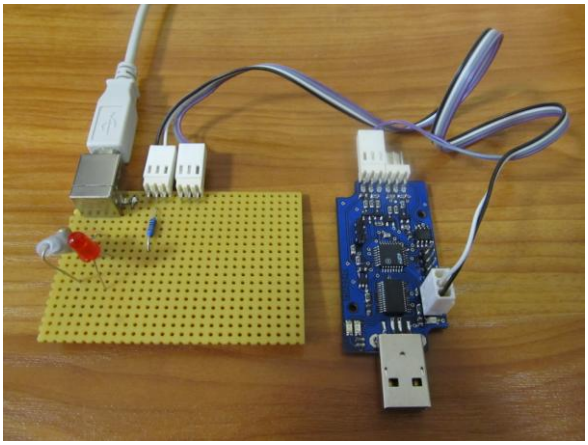


Figure 2. shows the demo board with the Edaq-24

A. Demo hardware

The demo heater board role is to give a realistic example of a delayed controlled heating system similar to a building with a thermostatically controlled heater. Thus the main component of the board is a thermistor coupled to a heater resistor as shown in Fig. (2) connected to the Edaq-24 board.

The resistance of the cheap thermistor sensor will correlate with the temperature of the environment. The accuracy of the temperature readings will be under 1°C with the used 10Kohm model in room temperature, which makes it an adequate sensor for the task. Note that, the 24bit ADC built into the micro-controller makes it possible to use high precision sensors without preamplification, thus the thermistor is directly connected to the input of the data acquisition board through an inner pull-up resistor.

The heater resistor requires much higher current than the PWM output could provide, so a transistor will switch the 5V presented on a USB port. The separation of the switched voltage through the heater resistor is a practical concern, because the micro-controller on the data acquisition board already less then a hundred milliamps. Making it more attractive an LED is switched on simultaneously with the heater to make it visible for the experimenter. Remote users can watch the blinking using a regular USB webcam.

Note that using a second PWM output to control an electric fan could be used to alter the heat loss similar to the dropping of the outside temperature raise the heat loss of buildings in winter. With basic electrical knowledge the proposed heater board can be built with a few cheap components even on a breadboard with the possibility to further upgrade with new features, like air ventilation.

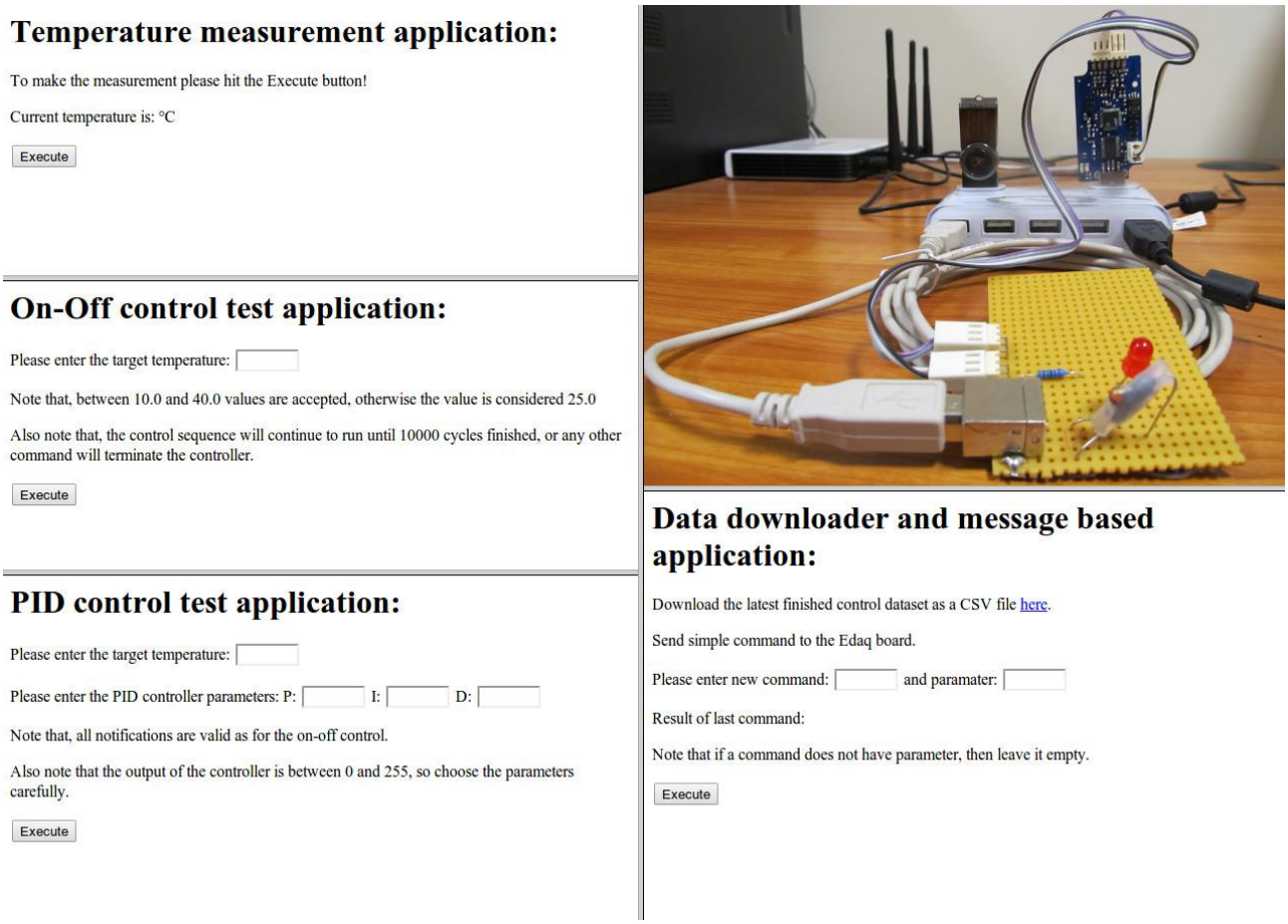


Figure 3. Shows the screenshot of the remote application

B. Demo software

A web based user interface was developed to illustrate the difference between the on-off or also known as bang-bang controller and the variations of the PID controller.

Main components of the web UI are depicted in Fig. (3), which is a simple HTML that uses frameset for the subtasks. Each frame source comes from a separate content provider, which makes the system modular and more robust. All user inputs are HTML forms with different number of text inputs submitted to the server. A PERL script handle the posted data and works as an interface between the HTML code running in the browser and the client application running on the router.

The top-left section contains a simple application with which a remote temperate reading is possible. By clicking on the button the measured temperature will be shown in the text.

On-Off controlling can be started using the middle-left part of the web UI, where a target temperature is required as an input field. By clicking on the button the controller will be initialized to keep the resistor temperature at the given value. Started controllers will keep going until a certain iteration number has been reached or any new message was received by the server thread. When the controller stops, it will write a CSV file containing all temperature measurements target temperature and control signals with timestamps in milliseconds. The CSV file can be downloaded using the bottom-right section of the UI.

PID controlling application can be found in the bottom-left section. It can be used to test simple proportional, PD, PI or PID controllers by giving the proper parameters. Setting the target temperature is also available in this subframe.

The top-right part contains a video-stream of the setup allowing instantaneous visual feedback of the experiment. Note that in order to see the quick blinking of the LED higher frame-rate may be required than the default when using mjpg-streamer. Also note that higher frame-rate will increase the required bandwidth needed for smooth playback.

Main function of the application in the bottom-right corner is to enable simple communication with the experimenting setup. All implemented commands with or without parameter can be called here. Second role of this part is to offer a link with which the last logged controller data can be downloaded.

To illustrate the difference between the on-off and a proportional controller a temperature measurement experiment was performed. Identical parameters such as the target temperature available time frame and similar starting temperatures were set highlight the characteristic difference between the two controller types. The results with the measured and target temperature and the control

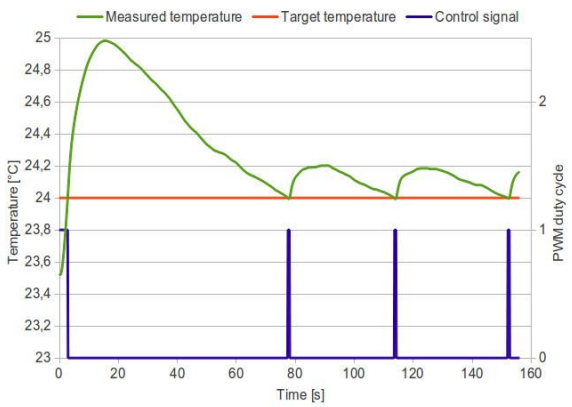


Figure 4. shows the characteristics of the on-off controller

signal are depicted as a function of time in Fig. (4) for the on-off controller and in Fig. (5) for the PID controller. The results clearly showed that the on-off controller had a significantly higher overshoot compared to the PID controller.

IV. Results and Conclusion

Cheap, broadly available hardware and some custom made open source extension boards with open source software aid highly efficient education. They can be used in various real experiments for example teaching many disciplines and showing that the experiments are transparent and scalable. Students can have their own setup, they can realize different instruments using the same hardware just by replacing the software that performs most of the instrumentation functions. The low cost, multi user experimenting equipment significantly reduces the need for alternative real laboratory resources.

In our paper we have shown a simple way to use a popular router and an open source sensor-to-USB interface for real and exciting remotely controlled experiments. The system can be used in various educational fields. Undergraduate or graduate students of engineering can use the remotely controlled laboratory experiment to examine the behavior of a real system even from home at any time, no further laboratory apparatus is needed. The lecturer can even ask for doing homework as well. The transparent system provides scalable education:

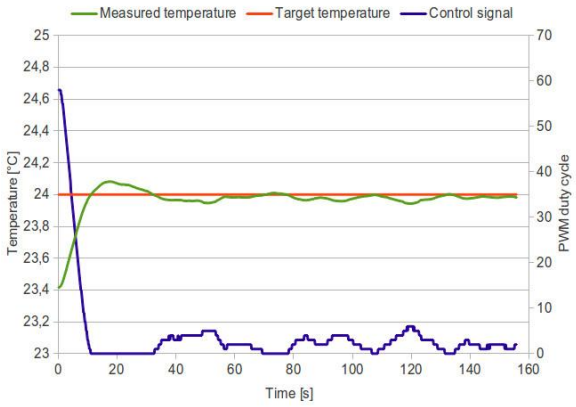


Figure 5. shows the characteristics of the PID controller

students can use already developed experiments; they can be asked to add further sensors; they can port the method to many other real mechanical and electronic systems; they can be involved even in software and hardware modifications and developments. Note also that larger projects can serve as a basis of the students' thesis both at the bachelor or master of science level, their creativity can be significantly improved. Note that the easily available, cheap, transparent computer controlled system can support teacher demonstration and home experimentation in the education of students at age of 12 to 18 years also.

The whole solution is open source therefore it can be used as a basis for many additional measurements and experiments. Note that it can be easily ported to other popular very low cost platforms including the Beaglebone and Raspberry PI single-board computers.

Acknowledgment

The publication is supported by the European Union and co-funded by the European Social Fund. Project number: TAMOP-4.2.2.A-11/1/KONV-2012-0073.

References

[1] Kopasz K, Makra P, Gingl Z. Edaq530: a transparent, open-end and open-source measurement solution in natural science education Eur. J. Phys. 2011, 32:491-504.

[2] ComLab: Computerised laboratory in science and technology teaching. http://e-prolab.com/en/index.html

[3] Stanislav Holec, Martin Hruška, and Jana Raganová. Integrated science through computer-aided experiments. Informatics in Education 2004, 3:219–228.

[4] F Schauer, F Lustig, and M Ožvoldová. ISES—Internet School Experimental System for computer-based laboratories in physics. INEER special volume: Innovations 2009 - World Innovations in Engineering Education and Research, pp:109–118.

[5] B Murovec and S Kocijancic. Educational data acquisition system with USB interface. EUROCON Computer as a Tool. The IEEE Region 8, 2003, 1:490–494.

[6] R Sallier, R Missler, and A Schütze. VenDASys — A versatile experimentation platform for educational purposes. Education Engineering (EDUCON), 2010, IEEE pp:205–210.

[7] OpenWrt Linux distribution for embedded devices https://openwrt.org

[8] Vernier LabQuest. http://www.vernier.com/labquest/

[9] Pico Technology: Oscilloscopes and data loggers for education. http://www.picotech.com/education.

[10] Gargoyle Router Management Utility http://www.gargoyle-router.com

[11] Gingl Z, A photoplethysmograph experiment for microcontroller labs. INTERNATIONAL JOURNAL OF ELECTRICAL ENGINEERING EDUCATION 2012, 49(1):42-60.

[12] Gingl Z, Fabricate a high-resolution sensor-to-USB interface. 2011, EDN 56 (22): 54-57.

[13] LIU, Jianhua; GRANT, Christopher; GALLAGHER, Donald. Adding new functions to the remote airfield lighting system. In: IEEE SoutheastCon 2010 (SoutheastCon), Proceedings of the. IEEE, 2010. p. 106-110.

[14] Chris Butlin. Electronic media: DrDAQ - an educational datalogger and virtual oscilloscope from Pico Technology. Physics Education. 2001, 36(1):85–86.

[15] Hunt, M. B. & Dingley, K. Use of the sound card for datalogging Phys. Educ. 2002, 37:251-253.

[16] Gingl Z., Mingesz R., Makra P. & Mellár J. Review of sound card photogates. Eur. J. Phys. 2011, 32:905-924.

[17] Remote laboratory applications http://www.inf.u-szeged.hu/robotics/remote

Web 2.0 tools in Croatian higher education: an overview

M. Čarapina*, D. Bjelobrk* and S. Duk**
* The Polytechnic of Zagreb/Department of Informatics and Computer science, Zagreb, Croatia
** The Polytechnic of Zagreb/Department of Electrical Engineering, Zagreb, Croatia
mia.carapina@tvz.hr, dunja.bjelobrk@tvz.hr, sanja.duk@tvz.hr

Abstract - This paper provides an overview of Web 2.0 tools usage in higher education among students and teachers in Croatia. The survey is based upon the research papers in the period from year 2008 to 2013 and represents alterations in usage quantity through referred time period, demographic correspondences, students' and teachers' preferences in usage of Web 2.0 tools, impact of their use in education and other statistical data important for general understanding of past and present situation of Web 2.0 education in Croatia.

I. INTRODUCTION

The overview presented in this paper is based upon research in the area of e-learning in Croatia for the period of five years, between years 2008 and 2013. This paper presents usage of Web 2.0 tools among students and teachers in higher education institutions in Croatia. The goal of the paper was to determine if some significant alterations can be perceived when observing the elapsed time and demographic parameters, and to make an insight in past and present state in this educational area.

For the purpose of this study, a literature search was undertaken in January 2013 in the bibliographic database Croatian Scientific Bibliography (CROSBI) available at http://bib.irb.hr. The database was searched by the keywords *web* and *education*, contained in the paper title, abstract and keywords section. The query returned approximately 300 research papers from which only few were relevant for this study. Consequently, other relevant literature was found in accordance with references stated in selected papers.

The structure of the rest of the paper is as follows: In Section 2 general concept of Web 2.0 is described and Web 2.0 tool classification is presented. Section 3 is divided into two parts, first of them presenting research and data relevant from students' perspective on tools usage, and second presenting teachers' perspective on using Web 2.0 tools in educational purposes. Finally, in the last Section conclusions are given regarding presented data.

II. WEB 2.0 IN EDUCATION

The term Web 2.0 generally defines the changes in the ways developers and users use the Web in comparison to static web pages of the early era of the World Wide Web [1]. New possibilities allow interactivity and collaboration among users on the Web. The concept of Web 2.0 has significantly influenced e-learning since the Web 2.0 applications encourage active user participation in creating, sharing and structuring data.

A. Classification of Web 2.0 tools

There are many different classifications of Web 2.0 tools for educational purposes which are mostly based on their diversity in functionalities and appliance.

Orehovački, Konecki and Radošević [2] define 6 categories of the Web 2.0 services based on their purpose: photo management and sharing, social networking and communication, learning materials development, replacement for standard desktop applications, learning management systems, and collaboration and creative learning services.

Similar classification is made by Gligora Marković, Rauker Koch and Frančić [3] in which they distinguish 8 different categories: tools for media exchange, communication tools, cooperation tools, creative learning tools, creation of learning tools, replacement of the standard desktop applications, learning management systems and tools for social bookmarks.

Since the categories overlap, the two abovementioned classifications can be conjoined in one general distribution:

- Communication and social networking tools are basically networking services where students can be a part of community and share knowledge and experience. This group of tools also encompasses text messaging applications and video or audio conferencing. The most popular example is Facebook which also has incorporated chat application.
- Tools for media exchange are used for sharing videos, photos and other types of files. Some of the services for sharing are Flickr, Odeo and YouTube.
- Collaboration or cooperation, creative learning applications and tools for development of learning materials assist collaboration between multiple users on the same project and motivate them to be creative. These tools can also be used for creation and organization of the materials. Some of these tools are Thinkature, NoteMesh, Mojiti and Imagination Cubed.

- Replacement tools for standard desktop applications are web applications that are usually used for word processing, presentations and diagrams. Some of these applications are Xwiki and Writeboard.
- Social bookmarking tools are used for organization and sharing the resources on the Internet. Representative tools are Magnolia and Delicious.
- 2.0 Learning management systems incorporate all or some of the above mentioned elements of Web 2.0.

All of the mentioned tools were part of one or all surveys presented in the rest of this papers.

III. CROATIAN EXPERIENCE IN USE OF WEB 2.0 TOOLS IN HIGHER EDUCATION

The scope of the published research papers regarding Web 2.0 tools usage in Croatian higher education includes the University of Zagreb: Faculty of Economics and Business (FEB), Faculty of Organization and Informatics (FOI), Faculty of Humanities and Social Sciences (FHSS-Z), Accredited College of Business and Management Baltazar Adam Krčelić (ACBM) and University Computing Centre (SRCE); then the University of Rijeka: Faculty of Humanities and Social Sciences (FHSS-R) and the Faculty of Teacher Education (FTE); and the University of Juraj Dobrila in Pula: Department of Preschool and Primary School Teaching (DPPST). Hereafter an overview of these surveys is given.

A. Students' perspective

Vlahović and Požgaj conducted two surveys described in papers [4], [5] and [6] at the Faculty of Economics and Business at the University of Zagreb. Two papers present student's inclination towards social networking as informal learning tool. The surveys were carried out among the students of the first year of Business Economics study. The demographic data of the survey participants is presented in the TABLE I.

The first survey presented with papers [4] and [5] was conducted in the December 2009. The goal of the survey was "to investigate students' preferences and involvement with different social networking services in particular their presence on Facebook, but also other Web 2.0 services"[5]. The majority of survey students (73%) were female students. The social networking services that were questioned in the survey were Facebook, MySpace, Twitter, Hi5, LinkedIn and option of entering other service. The research showed that 96% of students are registered on one or more social networks but the majority of them (79%) preferred Facebook among other services. Two excerpted reasons for profile registering are friends' suggestion (31%) and to renew a friendship (29%). They stated chat (63%) as the most useful Facebook feature and also photo and video comments (51%) and sharing (43%). Regarding usage of other Web 2.0 tools video sharing services, such as YouTube, Google!Video and Yahoo!Video, are stated to be the most used ones (82%). The second mostly used service is wiki (10%) and blog is the third (4%). The Web 2.0 tools that were questioned in the survey were wiki, blog, video sharing services, Flickr, Delicious and podcast. The survey showed that students spent in average less than seven hours using social networking and Web 2.0 services (43%) or seven to fourteen hours (42%) a week.

Another survey [6] by Vlahović and Požgaj was conducted at the same Faculty and published in November 2010. The goal of the research was to investigate "student attitude towards social networking services and their involvement and habits in using some of these services" and "concerns about using social networks" [6]. The majority of study participants were female students (74%). The social networking services that were questioned in the survey were Facebook, MySpace, Twitter, Hi5, LinkedIn and option of entering other service. Again, Facebook came out as the most popular service since 98% of survey participants have a profile on this social network. The research showed that 61% of students find Web 2.0 services useful as an e-learning tool. The authors also defined four distinctive groups of typical student attitudes to using Web 2.0 services and concluded that groups of students with the most positive attitude towards using

TABLE I. DATA SUMMARY OF CONDUCTED SURVEYS

		Survey participants information			
Paper reference	**Research year**	***Institution / Study***	***Number***	***Male | Female***	***Age or study year***
[4] [5]	2009	FEB / Business economics	195	27% | 73%	1st
[6]	2010	FEB / Business economics	184	26% | 74%	1st
[7]	2012	FHSS-R, FTE and DPPST / Teaching studies	83	19% | 81%	3rd-5th
[8]	2011	FHSS-Z	304	N/A	1st-5th and Postgraduate
[2]	2008	FOI / Information systems	415	76% | 24%	19-35 2nd-4th
[9]	2009	FOI / Information systems	285	75% | 25%	20-24 2nd-4th
[11]	2010	FHSS-Z	107	N/A	4th and 5th
[12]	2010	FHSS-Z ACBM	397	N/A	1st-5th and Postgraduate

Web 2.0 services in learning process are the "skilled active online users" and "cautious passive online users".

Kurelović, Tomljanović and Ružić-Baf conducted a research [7] which was published in May 2012 among students at the Faculty of Humanities and Social Sciences and the Faculty of Teacher Education at the University of Rijeka and at the Department of Preschool and Primary School Teaching at the University of Juraj Dobrila in Pula. The goal of the research was to determine "whether the students of teaching studies, the future teachers, recognize the pedagogical value of Web 2.0 tools" [7]. The Web 2.0 tools that were questioned in the survey were wiki, blog, media sharing, social networking, social bookmarking, collaboration on documents, tools for creative learning and learning authoring tools. The demographic data of the survey participants is presented in the TABLE I. The majority of survey students were females (81%) at the fourth year of teaching studies (58%). The research showed that students most frequently use Web 2.0 tools for media sharing (70%), instant messaging (65%) and social networking (57%). The most often used tool for learning is wiki (41%) followed by media exchange tools and social networks (35%), and document collaboration tools (34%). The research showed that students find Web 2.0 tools as very important (55%) or moderately important (40%) for their future job as a teacher. The majority of them (48%) stated that would like to use video content as a teaching method. As negative aspects of Web 2.0 usage in educational purposes 51% of them state that Web 2.0 activities require too much time and effort and possible lack of control in the teaching process (46%). General conclusion of research was that future to be teachers have positive attitude towards Web 2.0 usage in education.

Golubić and Lasić-Lazić conducted an online survey [8] in May 2011 with the goal to determine relevance of the higher educational institutions presence on the social networks. The data they collected also presents social network preferences of survey students which are relevant for this paper. The demographic data of the survey participants is presented in the TABLE I. It must be noted that the survey participants were not just students, but also non-teaching stuff and teachers affiliated with Faculty of Humanities and Social Sciences (87%) and few other institutions (7%) in and outside Croatia. Most of the survey participants were undergraduate and graduate students (78%) and postgraduate students (6%), therefore the obtained data can be considered relevant for this study. Research showed that Facebook is the mostly (76%) used social network. The majority (62%) of survey participants is not aware of the social network presence of their institution, but those who responded positively stated Facebook as most frequent answer. Authors discuss the ease of use and quick dissemination of information as positive aspects of social networks because of which they can be used as tool in education and communication. They state that institutional user can easily learn and adapt to it.

A number of surveys were conducted at the Faculty of Organization and Informatics at the University of Zagreb. Orehovački, Konecki and Radošević [2] conducted a survey which was published in 2008 with a goal to determine cognition and usage rate of Web 2.0 technologies and their application in e-learning domain among students with high level of information and communications technology skills. The Web 2.0 applications included in the survey were divided into 6 categories and include 38 different applications. The demographic data of the survey participants is presented in the TABLE I. Most of the survey participants were male (76%) and third year students (50%). Although the majority of them (more than 80%) rate their knowledge of information technology as great or excellent, the survey showed low level of knowledge regarding Web 2.0 technologies. The most known and frequently used are wikis (89.40% passively and 21.93% actively) and blogs (77.11% passively and 18.80% actively). Regarding learning activities survey showed that most frequently used applications are those for social networking and communication (41.51%) and photo management (38.07%). The most used social networks are Facebook and MySpace. The most used video applications are YouTube and Google Video. Authors concluded that students have very poor knowledge about majority of Web 2.0 tools and are more interested in passive than active use of applications.

In 2009 another survey [9] was published authoring Orehovački, Bubas and Konecki. The goal of the research was to determine factors which contribute to the everyday use of Web 2.0 tools by university students in order to facilitate their adoption in education. The demographic data of the survey participants is presented in the TABLE I. The majority of survey participants were male students (75%) with high level of information literacy according to self-assessment query. The Web 2.0 technologies that were questioned in the survey were wiki, blog, audio podcasting, vlogging, RSS, social bookmarking, tag cloud, virtual world, Ajax, SOAP/WSDL. Blog (98%) and wiki (87%) were stated to be known by majority of students. This fact did not confirm authors' hypothesis that the students of informatics have sufficient knowledge of various Web 2.0 tools. The research confirmed the hypothesis that personal characteristics, types of online activities and students' motives for use of the Internet/Web are associated with the use of Web 2.0 tools.

In [10] Bubas, Orehovački, Balaban and Ćorić discuss activities performed as a part of EduWeb 2.0 project that was conducted during the academic year 2009/2010. The idea of the project was to test up to 50 different Web 2.0 tools and present the experience with their use. More results on usefulness and benefits of the project can be found in their paper [10].

In January 2010 Banek Zorica and Ivanjko conducted an online survey [11] at the Faculty of Humanities and Social Sciences at the University of Zagreb with goal to gain insides in students' perspective on online services and their perception of the library. The demographic data of the survey participants is presented in the TABLE I. The majority of respondents were students from the Department of Information Sciences (43%) and the Department of Croatian Language (30%). Survey showed that students spent on average 15 hours per week on Internet. The questioned online services were e-mail, browsing videos and pictures, social networks, instant messaging/chat, reading and using wikis, playing online

games, using cloud computing, social bookmarking and tagging, using RSS, editing wikis and reading blogs. Mostly used online services are e-mail (96%), browsing video and pictures (70%), social networks (67%), instant messaging/chat (62%) and reading and using wikis (60%). The researchers also wanted to investigate students' attitude towards presence of the institutional library on the Facebook. The results of the survey showed that most of the students that have Facebook account would like to receive notifications from the faculty library group about new book arrivals (81%) and activities taking place in the library (72%). They also stated that they would like to use this group to search through the library catalogue (64%), communicate with the librarians (60%) and have the possibility to suggest new library services (56%). Although students see great advantages in the faculty library social presence only 31% of the survey participants are group members. The majority of students (81%) that were not members responded that they were not aware of its existence.

Another online survey [12] was conducted in December 2010 by Banek Zorica, Ogrizek Biškupić and Ivanjko with the participants from the Faculty of Organization and Informatics (33%) at the University of Zagreb with 79% of the participants from the Department of Information Sciences and Accredited College of Business and Management Baltazar Adam Krčelić (67%). The research investigated students' access to use of technology and use of social software in free time and educational purposes. Among questioned online services e-mail (79%), social networks (78%), browsing video and pictures (66%), instant messaging/chat (64%) and the use of wikis (46%) were stated to be most often used. In compare to previous research [11] the authors perceived an increase in use of social networks among students. The research also showed that most common activities on social networks are chatting (52%) and browsing photos from other users (39%). The research also showed that 98% of students use institutional educational e-learning platform, since they are forced to enroll. The most preferred way of communication with teachers as part of the learning process is e-mail (92%) while other online services had usage rate less than 20%.

B. Teachers' perspective

The only research paper found in the database that investigates the use of Web 2.0 tools by teachers in Croatia was written by Gligora Marković, Rauker Koch and Frančić [3]. They investigate the Web 2.0 use frequency among teachers in their educational process. The 96 survey participants were elementary school teachers (51.50%), secondary education teachers (47.92%) and higher education teachers (22.92%). The majority of participants were female (71.88%) aged 30-34 years. The research showed that 23.96% of them use Web 2.0 tools in education several times a week and 18.75% of them use it every day. Most of them use tools for audio and video sharing (55.21%), wiki (43.75%) and interactive presentation (40.62%). The excerpted reasons for not using Web 2.0 tools are insufficient professional education in the area (23.96%), poorly equipped classroom (17.71%) and insufficient awareness (14.58%). The lack of motivation was stated by only 2.08% of survey participants. Since the subject of this overview is past and current state in Croatian higher education, due to small number of teacher survey participants that work in higher education, the results of this research are taken with consideration.

IV. CONLUSION

The overview showed that the most known and used Web 2.0 applications are those for social networking and communication, then tools for video and photo management, wiki and blog. Services for social bookmarking and tagging, collaboration and creative learning are the least used ones. In general, the presence of Web 2.0 tools in higher education is low. Even the most popular service, Facebook, is primarily used for leisure. Students are generally not aware of the social presence of their institution or its sections on social network. Significant growth in number of Facebook users among students can be perceived when observing the time period from 2008 to 2013. The most common social network activities are chatting, and sharing, browsing and commenting photos and videos.

The research on teacher practice showed that minority of teachers use Web 2.0 tools in their educational process, thus this could be one of the reasons for the low knowledge level about the diversity of applications of this kind among students that could help them to collaborate or share knowledge and materials. Research showed that students are in majority more passive than active users, therefore teachers have to encourage them to use these tools more actively. Although the students are generally not aware of the variety and possibilities of Web 2.0 tools and their possibilities in education, it can be concluded that they are open for Web 2.0 technologies in education.

When observing the demographics, it must be taken in concern that the great number of survey participants was students with moderate or high level of information technology literacy accordingly to the subject of their study.

REFERENCES

[1] T. O'Reilly, "Design Patterns and Business Models for the Next Generation of Software," Web Squared: Web 2.0 Five Years On, October 2005.

[2] T. Orehovački, M. Konecki and D. Radošević, "Web 2.0 technologies in university education," Proceedings of the 31st MIPRO International Convention on Computers in Education, Opatija, Croatia, pp. 269-273, May 2008.

[3] M. Gligora Marković, M. Rauker Koch and M. Frančić, "Use of Web 2.0 tools in teaching," Proceedings of the 35th MIPRO International Convention on Computers in Education, Opatija, Croatia, pp. 1279-1283 , May 2012.

[4] Ž. Požgaj and N. Vlahović, "The Impact of Web 2.0 services on Informal Education," Proceedings of the 33rd MIPRO International Convention on Computers in Education, Opatija, Croatia, pp. 944-948 , May 2010.

[5] Ž. Požgaj and N. Vlahović, "Students' readiness for informal learning usingWeb 2.0 services," Int. J. of Knowledge and Learning, vol.7, no.1/2, pp.113-129, 2011.

[6] N. Vlahović and Ž. Požgaj, "Constraints in Course Design Using Web 2.0 Tools: A Croatian Case," Proceedings of The Second International Conference on Creative Content Technologies (CONTENT 2010), Lisabon, Portugal, November 2010.

[7] E. Krelja Kurelović, J. Tomljanović and M. Ružić-Baf, "The use of Web 2.0 tools in education from the point of view of future teachers," Proceedings of the 35th MIPRO International Convention on Computers in Education, Opatija, Croatia, pp. 1335-1340 , May 2012.

[8] K. Golubić and J. Lasić-Lazić, "Analysis of On-line Survey about Need for Presence of Higher Education Institutions on Social Networks: a Step towards Creation of Communication Strategy," Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces, pp. 195-200, June 2012.

[9] T. Orehovački, G. Bubas and M. Konecki, "Web 2.0 in Education and Potential Factors of Web 2.0 Use by Students of Information Systems," Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces, Cavtat, Croatia, pp. 443-448, June 2009.

[10] G. Bubas, T. Orehovački, I. Balaban and A. Ćorić, "Evaluation of Web 2.0 Tools in the e-Learning Context: Case Studies Related to Pedagogy and Usability," University Information Systems - Selected Problems, EUNIS International Congress 2010, Warsaw, Poland, pp. 259-277, June 2010.

[11] M. Banek Zorica and T. Ivanjko, "Social Networking at the University - Student Research Patterns and Online Activities," University Information Systems - Selected Problems, EUNIS International Congress 2010, Warsaw, Poland, pp. 473-482, June 2010.

[12] M. Banek Zorica, I. Ogrizek Biškupić and T. Ivanjko, "Student interaction in electronic environment," IADIS Multi Conference on Computer Science and Information Systems 2011, Rome, Italy, pp. 386-392, July 2011.

Using Microsoft Project for Project Management in Non-Governmental Organisations

Danijela Kažović*, Davorin Valenčić**
* University of Rijeka, Faculty of Economics Rijeka, Ivana Filipovića 4, 51000 Rijeka, Croatia
E-mail: dkazovic@gmail.com
** University of Applied Sciences Velika Gorica, Zagrebačka cesta 5, 10410 Velika Gorica, Croatia
E-mail: davorin_valencic@gmail.com

Abstracts - With the implementation and realization of the projects financed from European Union funds, project management in non governmental organisations (NGOs) becomes a more complex and demanding process. Managing projects according to European Union standards is a challenge and such projects require new control concepts of project management. Project management is the continuous process of planning, organizing, managing and monitoring all aspects of the project in order to achieve the project objectives in accordance with the planned time, cost and quality. The successful realization of the project depends on the project management, which requires a complex use of different knowledge areas, skills and methods. This age of constant development of science and technology opens up new possibilities and therefore there are many software packages that are used for project management support. This thesis will describe the realization of project management using Microsoft Project (MS Project) program, a specialized program for project management. Based on the results of the implementation it can be concluded that MS Project enables high-quality and successful project management.

I. INTRODUCTION

Nowadays, NGOs have become an important factor in the development of society. There is an ever increasing number of projects and the increasing involvement of NGOs when introducing changes relevant to the society. EU funds are particularly open to the development of civil society and NGOs projects in all areas of development.

Projects submitted for funding through EU funds are demanding, therefore the execution of such projects through NGOs, requires development of project management as an important way of organizing work within the organization during the project implementation phase. The scope of project management tasks has become ever broader with time. So for the successful completion of an entire project, besides the narrow area of project management, many other key fields should also be taken into account, such as: the organizational structure and the environment of the project, knowledge in the field of project applications, standards and legal framework, and general knowledge of business management and interpersonal relationships. Each of these factors can have a major impact on the success of the project.

As a support in monitoring of all factors of project management, many software programs have been developed. Among all these, MS Project is the most common choice.

II. PROJECT IN NGOS

Project is an organized undertaking towards a pre-defined goal or objectives that requires resources and effort, a unique venture having a budget and schedule. A project's success can be measured in terms of how closely it comes to meeting the goal or objective within the parameters of its budget and schedule [1][2].

Definition of the project can be applied in the NGOs as well. It is however important to point out the following definitions that characterize projects in NGO [3]:

- Project is a temporary process, which has a clearly defined beginning and end, a set of activities and tasks, and a budget, which has been developed to achieve a defined goal.
- Project is a set of interrelated activities, restricted by time, cost and range, and designed to achieve a single purpose
- Project is a temporary venture that creates a single result.
- Project includes a set of tasks or activities that have a well-defined starting point and well defined goals. Usually every task has a planned completion (deadline) and given funds.
- Project represents a series of activities related to the limited resources with the aim of achieving goals that are linked to the anticipated effects and impacts on the target group.

Depending on the scope, size and complexity, project requires less or more usage of Project Management.

III. NGOS PROJECTS WITHIN A FRAMEWORK OF EU FUNDS

Projects of NGOs that are applying for funding under the EU funds cover a wide range of areas and topics that have been identified as particularly important in the process of Croatian accession to the European Union, and relate to the issues of environmental protection and

sustainable development, provision of social services, good governance and the rule of law, democratization and protection of human rights and youth participation. Through these types of projects NGOs have an important role in the implementation of reforms in the pre-accession period [4].

In Croatia NGOs funding for development is more often provided through projects. It has become especially important as Croatia is approaching European Union, which started to open many funds and financial opportunities. Biddings of the European Commission require as a prerequisite good quality projects and implementers. Writing projects by the rules of the European Commission is a process that involves the determination of NGOs management to deal with it, fulfilling a commitment to the task, coordinated teamwork, and about a month to prepare the project so that it is successfully written and submitted in accordance with the requirements of the tender [5].

If we consider the current results and activities of a large number of NGOs that generate financial support from EU funds, it can be said that NGOs are now very reliable and desirable partners in usage of available resources from EU funds with the aim of contributing to the development of the local community.

Such a strong potential of NGOs seems to be particularly important in the context of preparations for usage of structural funds in forthcoming period when Croatia will have access to substantially greater resources from the EU budget that are intended for NGOs. Therefore, it is important to continue to strengthen the capacities for proposing quality projects aimed at improving the environment for the development of civil society and social cohesion [4].

IV. PROJECT MANAGEMENT ACCORDING TO EU STANDARDS

Project management can be defined as the implementation of knowledge, skills, tools and techniques to project activities in order to meet project requirements[6] [1]. Project management includes determining requirements, setting clear and achievable goals, establishing a balance between the competing demands for quality, range, time and cost and customization specifications, plans and approaches to the interests and expectations of the various parties involved [7] [1].

Working according to the project management methodology, from idea generation to finding ways of securing funds for the realization of all project activities, is essential in the implementation of projects financed from EU funds. If the NGO wants to apply for European Commission bidding to obtain funds, a well prepared project is needed. In the process of project management we go through different phases and difficulties. In this process, the possibilities of applying the software solution can facilitate the whole procedure of the project implementation.

The procedure of the projects implementation is always the same. For projects financed from EU funds it is very important to have a detailed project implementation report that should also include a part describing all the executed activities, together with the complete financial report of the fund usage. Reporting on the use of EU funds and securing that all prescribed EU procedures were respected is the most important part of project management. MS Project can be used to ensure good quality of activities implementation, financial management and reporting.

V. IMPLEMENTATION OF MS PROJECT IN PROJECT MANAGEMENT

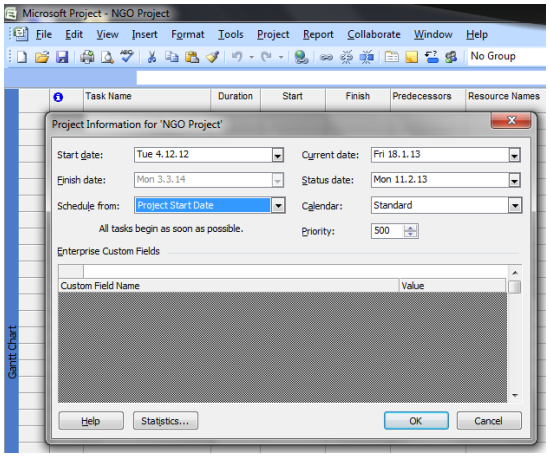
Microsoft Project is a software package for project management, which is intended for planning, monitoring and control of project realization, and it is one of the easiest and most widely used software packages for project management. It is the ideal solution for managing NGOs project according to EU standards.

For a project to be successfully implemented it is essential to carry out all planned activities within a deadline, to have financial plan at the level of the estimated budget and that project beneficiaries are satisfied with the implementation of the project and benefits derived from it. It is therefore important for project management to balance out the time, finances, and human and material resources [2] [8] [9].

With MS Project the following can be achieved [8]:

- Monitoring of all gathered information about the job, duration and resources;
- Visualization and presentation of the project plan in standard formats;
- Efficient allocation of tasks and resources;
- Exchange of information on the project within the project team;
- Communication with resources and other carriers of responsibilities.

When you begin using the program it is necessary to make the preparation of program that includes entering basic information about the project, calendar and activities. When you start entering information about the project, the framework for Project Information dialog is used (Picture 1.). That is where you enter basic information about the project such as the Start Date of project and Finish date of project. Defining a calendar is important because the project uses this calendar to estimate the time needed to execute the project. Setting up basic information about the activity provides us with the information that helps determine the characteristics of duration of activities. After entering the necessary information about the project, the program can then be used to manage the project.



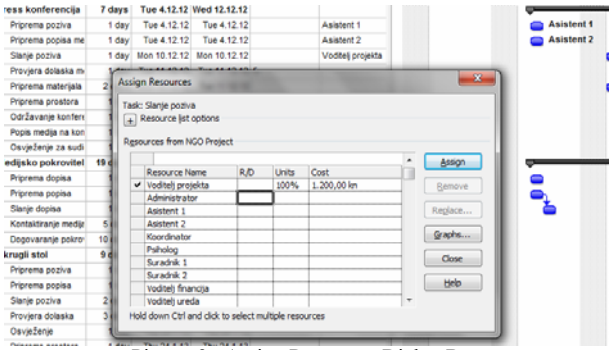
Picture 1.: Project Information Dialog Box
Source: author's work

Since MS project has a large number of features, the usage in managing the projects funded by the EU will be shown in the example of one cluster of activity that is common in many NGOs projects, and that is a media campaign. Such a cluster of activities consists of many sub-activities: collecting media sponsors for the campaign, organizing press conferences, round-table organization, organization of public hearings, etc. As a result of these activities many outputs are occurring, among which the most prominent is a press clipping that represents selective and continuous monitoring, classification and storage of information from the media.

Applying MS Project in planning, organizing and realization of media campaign would facilitate the execution of tasks within this cluster of activities for the entire project team, and primarily would facilitate for the main project manager the implementation and reporting upon completion of activities. Through MS Project project manager can achieve insight into the level of implementation of tasks at any time, track status of executed tasks, and take measures where necessary to eliminate insufficiently completed tasks.

Since cluster activity consists of a series of activities, all the activities in the cluster will be grouped into one. The name of cluster activities is written in bold letters, and a click on it opens the individual activities. When viewing activities different views can be used, such as: View - Calendar, View - Network Diagram, View - Tracking Gantt and others.

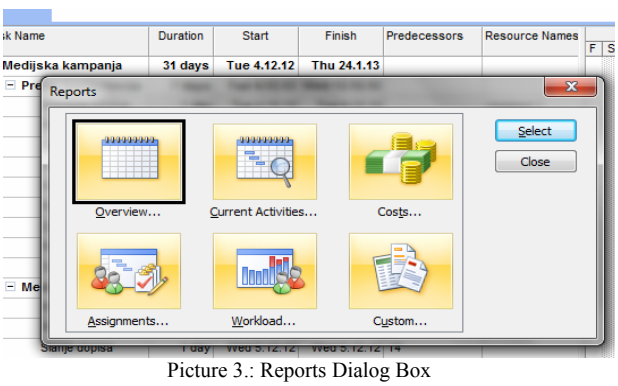
After entering the cluster activities for Media campaign, it is required to enter the necessary resources for the successful implementation of planned activities. Entering of resources implies the necessary human and material resources. MS Project puts focus on two aspects of resources, their availability and their costs. Availability determines when and how the resources can work on the planned activities within the cluster, while the costs show a cash value that is associated with the use of these resources, which should be in line with the project budget. Resource Sheet is used to enter the required resources, such as all employees on the project and all the necessary planned material resources. For the implementation of cluster activities Media campaigns human resources or project team play the most important role. Allocating the necessary resources to certain activity is done using Assign Resources (Picture 2.).



Picture 2.: Assign Resources Dialog Box
Source: author's work

If we enter all the necessary data for all the activities, it is simple for a Project manager to follow the execution of everything planned and realized through multiple Views and Reports. Gantt Chart is a common way to show a project plan, especially when the plan is presented to others. It is useful for entering and adjusting detailed information about the task, as well as to analyze the project. Gantt Chart for cluster activity media campaign will show each activity within the cluster, which human resources are in charge for the implementation of activities, which material resources are used in the implementation, how long does the enforcement action period lasts (days, months), date of commencement of activity, the date of completion of activity, etc. . It will also allow hyperlinking on achieved project outputs.

Alongside Views, which is the essential function in project management, especially in the phases of monitoring and finalizing the project and preparation of final reports, we also have a Reports function that is as equally important. Reports are pre-defined formats for printing data from MS Project. Through reports we can monitor the execution of tasks through different options: Overview Reports, Current Activity Reports, Assignment Reports, Workload Reports, Cost Reports, Custom Reports (Picture 3.).



Picture 3.: Reports Dialog Box
Source: author's work

Further usage of MS Project offers a number of other features that provide support to project management primarily focused on improving of the productivity and quality of work on projects.

VI. CONCLUSION

Project management is important to all organizations (companies, NGOs, government institutions, etc.) that conduct activities periodically or continuously through projects. It is a system that enables project managers to handle projects, allowing the realization of project efficiency.

Working without project management brings to higher risks and less integration of all the components of the project. Project management is particularly important for the design, implementation, monitoring and scheduling of changes at a particular time, which directly reduces the risk of delays and increased costs.

The realization of the goals of the project is not easy because we live in the age of rapid development of science and technology. One of the basic goals of NGO's projects is to reduce risk and uncertainty and increase the likelihood of achieving the intended results at set intervals and the estimated budget of the project.

MS Project as an instrument of support shows the path that we must choose to achieve the desired results. MS Project allows the project manager to consistently follow the objectives and planned activities within a set range, time, budget and quality. It can be concluded that MS Project is a program that supports all projects and a source of multiple benefits in their realization.

The intention of application of MS Project in the implementation of projects is to divide the tasks and to control the implementation of distributed tasks. With the implementation of program we can track how much time is spent on particular activities, what are the results of activities, what are the costs of each activity given the time and resources spent. With the monitoring of activities we can find better solutions in the division of tasks and in the realization, we can also achieve savings, and increase the efficiency of the project team. And it is especially crucial for planning future projects also, because with further analyzes and reports we can reach to a large number of useful information.

With Croatia joining the European Union a new structural funds to finance projects are opened. These funds are larger and more complex than previous projects from pre-accession funds. Taking into account that in future, projects will become more complicated and that for project management higher skills will be required, the application of MS Project on such projects is very important and it presents an appropriate choice in planning, organizing and implementing projects.

REFERENCES

[1] "A Guide to the Project Management Body of Knowledge", Project Management Institute, Pennsylvania, 2008.

[2] P. E. Harris: "Planning and control using microsoft Office project and PMBOK Guide Fourth Edition", Eastwood Harris Pty, Victoria, Australia, 2010.

[3] M. Field, L. Keller: "Project management", Thomson Learning, London, 2007.

[4] "Europski fondovi za organizacije civilnog društva u Hrvatskoj: zbornik uspješno provedenih projekata financiranih sredstvima Europske unije", Vlada Republike Hrvatske - Ured za udruge, Zagreb, 2010.

[5] "A Guide to European Union Funding for NGOs - Accessing Europe's Largest Donor", ECAS - European Citizens Action Service, Brussels, 2005.

[6] J. Westland: "Project management Guidebook", Method123, Kennesaw, USA, 2003.

[7] Z. Zekić: "Projektni menadžment - upravljanje razvojnim promjenama", Ekonomski fakultet Rijeka, Rijeka, 2010.

[8] E. Marmel: "Microsoft Project 2007 Bible", Wiley Publishing, Indianapolis, 2007

[9] "Microsoft Project 2007: The missing manual", O'Reilly Media, Sebastopol, USA, 2007.

Heuristic Evaluation in the Human Computer Interaction Course

Dušanka Bošković, Nihad Borovina
Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina
dusanka.boskovic@etf.unsa.ba, nihad.borovina@etf.unsa.ba

Abstract - This paper presents a model for teaching Heuristic Evaluation in a Human Computer Interaction (HCI) course. The course objective is to teach students to design HCI as an approach to develop usable software. The students are involved in the process of the Heuristic Evaluation to learn to address usability problems as a part of an iterative design process. The evaluation supports other course objectives: combine theoretical knowledge as usability principles, interaction design patterns with practical experience obtained through project work. Student achievements and assessment are presented indicating the assignment results.

I. INTRODUCTION

This paper describes how the use of the heuristic evaluation can help to maximize the learning experience in the HCI education by aligning theoretical knowledge as usability principles with the practical experience in implementing an user interface (UI). The evaluation encourages the students to identify many problems in the UI design and to propose reasonable solutions.

The model for teaching heuristic evaluation described in this paper have been applied in the HCI course, offered as elective course within the Master Programme at the Automatic Control and Electronics Department at the Faculty of Electrical Engineering, University of Sarajevo. For the first time it was offered in the Fall 2009 semester and has been offered every Fall semester since then.

The rest of this paper is organized as follows. The Section II provides the background information for the course and the evaluation method. The Section III describes the context in which the Heuristic Evaluation (HE) is taught. Section IV presents the tool designed for the evaluation and used in a homework assignment, the assignment results and grading. Finally, conclusions are presented in the Section V.

II. BACKGROUND AND MOTIVATION

A. Educational objectives

The HCI course is focused on designing software/system interface as an approach to develop usable software. Students gain knowledge on designing and implementation of human-machine interface for complex systems such as: SCADA systems, systems for supervision, safety and signalization, chemical and power plants control.

Syllabus of the course includes the following topics: usability principles, interface types and characteristics, methods and tools for designing user interface, prototyping techniques, user interaction design patterns, evaluation of user interface usability and trends of user interface development. Traditional lectures covering theoretical principles are followed with lab exercises, homework assignments and a final project.

There is no paper or oral exam. Students are engaged in a project to design and implement user interface by using modern software tools and visualization components. Instead of an oral exam, overall knowledge and ability to discuss HCI issues are assessed using a written report, where students answer a set of questions related to the lectures, but using their project as a case study.

B. Theoretical context

Heuristic Evaluation is introduced by Jakob Nielsen in 1990 [1], as a systematic usability inspection method. In the HCI textbooks it is defined as a method that typically involves an expert playing a role of the user, analyzing aspects of an interface, and identifying any potential usability problem by using a set of guidelines [2].

Evaluators observe and use the UI and assess if it is in compliance with established usability principles, called the "heuristics". Nielsen identified a basic list of ten rules to be used when conducting such evaluations: visibility of system status; match between system and the real world; consistency and standards, user control and freedom; recognition rather than recall; flexibility and efficiency of use; aesthetic and minimalist design; error prevention; recognize and recover from errors and help and documentation [3].

In the HE Usability problems are to be resolved as a part of an iterative design process, and very important part of the evaluation are recommendations to fix a problem.

III. TEACHING HEURISTIC EVALUATION

Teaching Heuristic Evaluation starts with a traditional lecture introducing definition, method, practice, benefits and disadvantages. The evaluation methods are concluding part of the course and reflects many important topics discussed earlier as UI design principles, UI design methods and interaction design patterns [4]. It is a “closing a circle” and all pieces of the course are fit into the right place.

The lecture is followed with a lab where students are getting familiar with the HE procedure. They are equipped with the Heuristic Evaluation Report designed as a tool to assist them in the process. The report is tabular, addressing all ten usability principles, enforcing systematic presentation of problem that needs to be balanced with a proper solution proposal.

During the lab, students perform HE for assigned web pages according the procedure defined in [1]. Usually, three to five students analyze the same page. They work at the beginning individually, and complete their own evaluation. After that, they work together in order to make an aggregated HE report as foreseen in [1].

After the lab, to demonstrate the knowledge they acquired, students are given two homework assignments related to the HE. The first assignment is to perform evaluation of a web site of their choice, and the second is evaluation of a project developed by other students. The first assignment is done individually, and the other is done by a group of students.

It is necessary to emphasize that the objective of assignments is to improve usability as utility and operational ease of use [5] and not to assess how attractive user interface is. The later objective is also very important, especially for designing e-commerce web pages [6], but this is not aligned with the objectives of the HCI course described in this paper.

IV. HEURISTIC EVALUATION ASSIGNMENT

A. *Heuristic Evaluation Report*

The Heuristic Evaluation Report comprises ten tables, each of them related to an UI design principle. The table has a title row quoting the design principle, followed by a description, objective of the principle and a few typical questions that can help analysis. The example table is presented in the Fig. 1.

For each identified problem student need to fill in the following data in the HE Report: Usability problem, Problem location, Level and Recommendations for improvement. Students are required to document at least 10 problems and to put them in the most appropriate category.

Students have a freedom to choose a web page for the evaluation for the first assignment. Initially, there were only two recommendations for the selection of a web page:

- it is an official website of company, organization or institution which present the information and / or provide services,
- it involves some kind of user interaction (browsing, search, registration, purchase, etc).

1. Visibility of the system (web page) status

Analyze if a system keep user informed about the system status using adequate feedback. Hint questions:

- Is the purpose of the web page clear?
- Is it clear how to navigate further on?
- Is the feedback for every user action meaningful and expected?

Ref.	Usability problem description	Problem location	Level of the problem	Recommendations for improvement
1.1.				

Figure 1. Heuristic Evaluation Report Excerpt

However, after some experience with submitted assignments, the list of recommendations was extended with advices to avoid:

- sites with extremely poor, chaotic or non-standard design,
- sites listed as "worst web pages in the world", "top 10 worst web pages" or similar,
- multimedia and art project sites,
- "wiki" sites,
- sites for registered users, without the possibility of free registration,
- sites with outdated design, tailored for the dial-up communication technology,
- sites not in local or English language.

Additional recommendation was that site should have a solid design, but with some insufficiency in the user interaction. After the improvements in the list of requirements for the web page, adequacy of the web sites also improved, reflecting the assignment grading as well.

The HE Report is slightly changed for the second assignment and adjusted for the application evaluations.

B. *Student Results*

In this section content of the student HE Reports is analyzed and both quantitative and qualitative results are presented. The analysis is done for the following assignments: Web Page HE and Student Project (Application) HE, received in the academic years 2010/2011 and 2011/2012. The total number of students were 75, 40 from the Automatic Control and Electronics Department and 35 from the Computing and Informatics Department.

The students have done a lot more than a required minimum. The number of documented problems were 60% higher than mandatory, for both assignments, web pages and student projects. Table I shows how detailed HE Reports were regarding documented pairs: problem and recommendation.

TABLE I. COMPLEXITY OF HE REPORTS

Number of remarks	Web pages evaluations		Student project evaluations		Comment
<13	26	34.7%	13	35.1%	Essential analysis
13-20	38	50.7%	13	35.1%	Extended analysis
>20	11	14.7%	11	29.7%	Deep analysis

TABLE II. BREAKDOWN OF THE REMARKS

Categories	Web pages		Student Projects	
	Num. of remarks	%	Num. of remarks	%
1. Visibility of system status	190	15.8%	81	11.5%
2. Match between system and the real world	146	12.2%	101	14.3%
3. User control and freedom	114	9.5%	59	8.4%
4. Consistency and standards	164	13.7%	79	11.2%
5. Error prevention	97	8.1%	78	11.1%
6. Recognition rather than recall	80	6.7%	76	10.8%
7. Flexibility and efficiency of use	101	8.4%	50	7.1%
8. Aesthetic and minimalist design	186	15.5%	98	13.9%
9. Help users recognize, diagnose, and recover from errors	65	5.4%	39	5.5%
10. Help and documentation	56	4.7%	44	6.2%
Total number of remarks:	1199		705	
Number of evaluations:	75		37	
Average number of comments per evaluation:	16.0		19.1	

Reports with less then 13 remarks are considered as essential. The reports that have between 13 and 20 remarks are marked as extended, and with more than 20 is considered as deep analysis. Nearly 30% of student projects evaluations may be viewed as deep analysis.

Table II shows breakdown of the remarks according to the design principle they address. Remark categories in the table are presented in the same order as in the questionnaire.

The number of identified problems slightly differ for web page and project evaluations. The web page evaluations have the highest number of problems related to the principle “Visibility of the system status”, while in the project evaluations, the focus was on the principle “Match between system and the real world”. The category “Aesthetics and minimalist design” was important in both types of the evaluation reports. Comparing the results for the “Consistency and standards” and “Error prevention”, the former is found more important for web pages, and the latter for the projects.

The results are expected and related with the purpose of web pages and project case studies. The primary goal of web searching is to find an information, to search and explore, so visibility, minimalistic design and consistency which facilitate those tasks are required. On the other hand, in student projects, especially those simulating industrial applications, real-world concepts, error prevention and ease of use play important role alongside visibility and minimalistic design.

In addition to the quantitative analysis, the level of problem identified and quality of recommendations are very important when assessing student performance. Not all remarks were properly documented, and the most common mistakes students made in the reports are:

- Identified problem is repeated in different categories. Sometimes the problem was formulated in a different way, but still it was addressing the same issue . For example, the following two remarks are the same: “Registration form does not exist or it is not visible enough” and “During the login process, error message indicates that account does not exist, but it is not clear how to open a new account”.
- Recommendations for improvement are not specific but rather general: "simplify design", "data should be organized in a more clear way", "use a consistent outlook throughout the site".
- A very significant and obvious problem on a page is not identified. Interesting example is a site organized in three parallel columns with three independent scroll down controls.
- Failing to recognize compromises and priorities, like: "too much content on a single page" for a TV programme for a whole week.
- Focusing on broken links, errors in typing and other technical flaws. It is considered as one remark covering all defects of the similar type.

C. Grading

Assignments grading was not based on number of remarks only, the level of problem identified and quality of recommendations have a significant part of the grade. Breakdown of a grade is: number of design principles addressed – 30%, number of problems identified – 30%, and suitable solutions proposed 40%.

Grading results for the web page evaluation assignment for the academic years 2010/2011 and 2011/2012, and for 75 students, are presented in the Fig. 2. The pie chart provides distribution of points awarded, with 7 points as maximum grade.

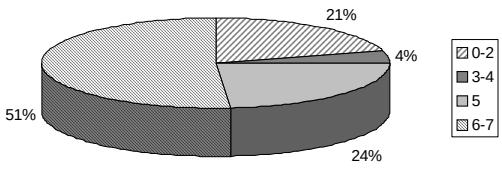


Figure 2. Grading results distribution

The advantage of this type of assignment is that cheating and superficial work is easily recognized and as a result discouraged. It is not time consuming to separate bad and moderate from good reports, but a lot of work is needed for grading good reports. Every web page evaluated by students has to be evaluated by teaching staff as well. This task is also very rewarding, and always helpful for further teaching because of many good ideas and useful examples of both good and bad UI design solutions.

V. CONCLUSION

The results presented indicate that heuristic evaluation is an appropriate method for the HCI courses. It is understandable, simple and comprehensive and it can be adopted by the student quickly. The student HE Reports provide evidence of the acquired competences. Matching of usability problems with usability principles demonstrated acquired knowledge of good design principles. Providing good recommendations for improvement demonstrated acquired knowledge of the UI design patterns.

The web page evaluation assignment posed several challenges for the students. They were supposed to find a non-trivial web-page. While finding appropriate web-page students were engaged in the inspection process, and actually performing mini-evaluations for several pages. Heuristic Evaluation Report design calls attention to solution, not only to problems.

Students gain experience in engineering approach to the HCI, some solutions are obvious after recognizing a problem, some problems ask for much more complex solution, and often there is no overall solution, but need to make trade-offs and understand prioritizing.

REFERENCES

[1] J. Nielsen, and R. Molich, "Heuristic evaluation of user interfaces," Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people, pp. 249-256, April, 1990, Seattle, Washington, United States

[2] H. Sharp, Y .Rogers, and J. Preece, Interaction Design: Beyond Human-computer Interaction. John Wiley & Sons, 2007.

[3] J. Nielsen, "Heuristic evaluation", In J. Nielsen, and R.L. Mack, (Eds.), Usability Inspection Methods. John Wiley & Sons, 1994, pp. 25-62..

[4] J. Tidwel. Designing Interfaces: Patterns for Effective Interaction Design, O'Reilly Media, 2005.

[5] ISO 9241. Ergonomic requirements for office systems with visual display terminals (VDTs). International Standards Organisation, 1997.

[6] A. Sutcliffe, "Assessing the Reliability of Heuristic Evaluation for Website Attractiveness and Usability," Proceedings of the 8th International Workshop on Interactive Systems: Design, Specification, and Verification, pp. 183-198, June , 2001 Glasgow, Scotland, UK.

Motherboard and user experience

J. Tomljanović*, T. Turina**, E. Krelja Kurelović***
Polytechnic of Rijeka
Vukovarska 58, Rijeka, Croatia
* jasminka.tomljanovic@veleri.hr
** tturina@veleri.hr
*** elena@veleri.hr

Abstract: The motherboard is the heart of the computer and also the most neglected part of the hardware. The main purpose of the motherboard is to connect all parts of a computer such as: memory, processor, disks and other devices in a better and faster way. Unlike other hardware devices development of motherboard is very unique and has a great impact on the market. While other devices developed independently of each other, it depends on the development of the motherboard whether these devices will work together. The objectives of this study were to describe the most important sections of the motherboard and its development starting from chipsets, through memory, processor, which is one of the parts that are selected first when designing a computer, to the bus and the other slots. In addition, there will be a describtion of the motherboard models and their development. An online survey was conducted and that gave the answers to the questions like: how much computer users are familiar with their motherboards, have they ever had motherboard failures and what were they.

I. INTRODUCTION – MOTHERBOARD BASICS

The motherboard is a printed board on which are located the main parts of a computer: CPU, memory, bus and other internal circuits can be soldered to the motherboard and can be added into the corresponding sockets and connectors. The motherboard is often sold separately without the processor, memory, and additional circuits so that customers could assemble a computer to suit their needs. [1]

The motherboard is located inside the case and attached with small screws through pre-drilled holes. The front of the motherboard contains ports that connect all the internal components of a computer. There is one slot for the processor, more memory module slots and several ports to connect the floppy drive, hard drive and optical drive via cable-like ribbons. Power from the power supply is provided to the motherboard with the help of a specially designed connector. The front side of the motherboard also includes slots for various cards (video, audio, etc.) that serve to expand the capabilities of the computer. On the left side of the motherboard are numerous slots which are used to connect most of the external input/output devices such as a monitor, printer, keyboard, mouse, speakers and more. Motherboard and case are made so that if you use some of the cards the connections for input/output units are located at the back of the computer to make them more accessible to use. [8]

II. CHIPSET

Chipset greatly determines the features of the motherboard, so the motherboard is sometimes classified into groups according to them. These additional integrated circuits are soldered to the motherboard and can not be changed. They determine the most important features of the computer: processor type that can be on the motherboard, memory type, bus, the method of exchanging data with the environment, etc. [1]

When the IBM (International Business Machines) constructed the first motherboard for PCs it took six different chips and later, developing that motherboard, the number increased to nine different chips that have ensured the functioning of the system. Building the copy of one of these IBM systems require more than one hundred individual chips. Therefore, the price of motherboards was extremely high and there remained little space to add other functions. In 1986 the company Chips and Technologies has introduced a revolutionary component called 82C206 - the main part of the first chipset for personal computers. The single chip had all the functions of the main chips of the motherboard for the AT (Advanced Technology) systems. This meant that, in addition to the processor, all the major components of the motherboard can be replaced with a single chip and that the functionality can be achieved with an additional 4 chip. That chipset was called CS8220. The next update of that chipset was NEAT (New Enhanced AT) CS8221 and had a set of only three chips, and that version was followed 82C836 Single Chip AT chipset that finally united all the functional chips into one chip. [2]

Other chip makers have begun to copy the idea of that chipset, but for most of them the market was short-lived and unstable, and after just a few years, most of them stopped producing and the rest were struggling to survive in the market. This was due to the new manufacturer that came into the market and in a little more than a year he dominated the chipset market. It was Intel, and since 1994. he has held the advantage.

Intel has worked for a short time a special chipset for EISA (Enhanced Industry Standard Architecture) bus, although she eventually was dormant, but Intel got the inspiration in making chipsets. Intel has become impatient waiting for chipsets which are adjusted to the new processors. It took almost two years of waiting for the chipset that supports the new processors to be developed,

and Intel could not sell their processors until the producers did not make motherboards that support them. Intel's idea was to simultaneously develop a new processor and a chipset that supports it, thereby encouraging manufacturers of motherboards in a faster response since the two main components are already on the market. Intel tested that with the presentation of the 420 chipset that is accompanying 486 processor in 1989. This has enabled motherboard manufacturers to start with the preparation and after just a few months came the first motherboard that supports it. Intel soon realized that processors and related chipsets are 90% of the components in a typical motherboard and in 1993 for their Pentium processor Intel immediately released their 430LX chipset and a fully completed motherboard. In a year Intel has dominatet the chipset and motherboard market. [2]

Most previous chipset used multistorey architecture that includes what is called the North Bridge and South Bridge and Super I / O chip [2]:

- North Bridge chip – connects to the processor, the AGP and PCI buses and memory. He is actually the control system.
- South Bridge chip - apart from the North Bridge chip is connected to the PCI bus and the ISA bus and UCB bus if the system has any.
- Super I / O chip - it is a special chip connected to the ISA bus. It includes most of the common external units into one chip.

All of the more recent South Bridge chips include the Super I / O functions and newer architecture do not need a separate Super I / O chip.

III. MEMORY

Memory, of which we speak here is called the working memory and is built into the computer itself. Working memory is the one memory that the processor uses for immediate storage and supply data. While the computer is on and processing data, the data and programs can be found in the working memory. Usually a few bits is stored in a specific location in memory. That place is called the memory location. [1]

RAM is a memory with limited access. It is a working memory in which you can write and read data from it as many times as you want. The stored data remains in this memory until they are deliberately changed or until the power supply is interrupted. The main features of RAM are capacity and speed. It is desirable that RAM has a larger capacity in order to store more data. [1]

SRAM (Static Random Access Memory) or static working memory is a type of memory in which each bit is stored in one of bistable circuits located in the memory integrated circuits. Without external shocks, bistable circuit permanently occupies one of the two stable states. The transition from one state to another encourages the appropriate signal from outside and the state can present a logical 0 and logical 1. The registered data will remain stored until a power failure or intentional changes. The benefits of SRAM memory is very fast memory access. The disadvantage is relatively large dimensions of the bistable circuits which limits the number of bistables that can be placed on a plate. [1]

DRAM (Dynamic Random Access Memory) is a type of memory where each bit is stored as a charge in miniature condenser located in the memory integrated circuits. Data exchange with DRAM memory is slower and more complicated than the exchange with SRAM memory due to the need to renew the charge in the capacitor in order to preserve data. The advantage of DRAM capacitors is the small dimensions. Therefore there can be more of them on the same plate. The disadvantage is the need for relatively complex drive frame and slow work caused by restoring memory. [1]

Because of its importance and cost computer memory is built so that it can be easily installed and modified. To facilitate the expansion of memory it is sold and installed in the so-called memory modules. Memory module is a printed circuit on which the memory integrated circuits are soldered, and on one edge there are connectors. There are several different models of memory modules that are not interchangeable. SIMM (Single Inline Memory Modules) is the oldest type of memory modules and not used more in modern computers. There are versions with 30 and 72 contact. They must be installed in pairs. DIMM (Dual Inline Memory Module) is currently the most prevalent types of memory modules. There are versions with 168 and 184 contact. They can be installed individually. RIMM (Rambus Inline Memory Module) is designed for Rambus memories. There are versions with 168 and 184 contact.

IV. SOCKET AND PROCESSOR

Port for the processor, or socket, is the connector that the processor is plugged into. In all motherboards processor is connected through the socket so it could be changed. The reason is twofold. On the one hand, making it possible to install the motherboard processor according to user preference, on the other hand it is possible to substitute in case of failure. [1]

The processor is a semiconductor device in which one plate of a semiconductor holds all the important parts of the central processing unit, such as for example the arithmetic-logic unit, internal tanks, supervisory circuits, etc. [1]

Intel was the first who invented the first processors that user could install and replace and has developed standards for the CPU (Central Processing Unit) sockets and slots that allow different models of equivalent processors to go on the same socket. The main finding was the use of ZIF (Zero Insertion Force) design for sockets, which means that the processor can be easily placed or removed without tools. ZIF sockets use the lever that engages or releases chip, and is open until the chip can be easily removed and replaced by another. ZIF sockets have a label that is usually printed on them, which indicates the type of socket. [2]

All processors were mounted on sockets, but Intel and AMD in 1990’s temporary switch to the processors that

plug into the slots on the motherboard because the L2 cache memory, which is located on a separate chip, started to be added to the processors. Then the processor is located on two different chips and the costs of such processors were higher. Then Intel began to incorporate Intel L2 cache directly on the processor die, which meant no additional chips. Shortly after that AMD has done the same. [2]

Since 2000 Intel and AMD are in constant rivalry. Since switching from one, and later 3 megahertz, introducing 64-bit processors, the introduction of multi-core processors, either on one or more chips, to the possibility of improving the performance of videos, pictures and web operations and the minimum use of energy a lot of novelty can be expected of the two leading manufacturers.

V. BUS

Bus is standardized connection to exchange data between two or more devices. By this definition bus includes all connections to exchange data. [1] Considering the place of exchange of data bus is divided into internal and external. The internal bus is used to exchange data within a computer and an external bus is used to exchange data between a computer and peripheral devices.

Bus is usually performed as a conductor on the surface of the printed circuit board, and is connected to the terminals for additional circuits in the computer. There are three main types of buses. Data bus is a set of guidelines for the transmission of electrical signals that represent data. The number of these guides is equal to the number of bits that can be processed by the CPU. Address bus is a set of guidelines for the transmission of electrical signals, which are represents the addresses, and their number depends on the computer architecture. Control bus is a set of guidelines for the transmission of electrical signals which represents the supervisory and control signals, and their number and function vary from computer to computer. Additional assemblies, which are subsequently incorporated into the computer must be connected to the bus in the computer and are built mostly as printed circuit boards with soldered electronic components and connectors on one end. In order to avoid that every computer manufacturer has its own bus several standardized internal bus was arranged. [1]

PCI (Peripheral Component Interconnect) bus is the most common in most PC's. It has general purpose and is intended for connection to different cards. It is independent of the processor, and in principle can be used with any processor which enables the processor and the bus to work at different speeds. In addition, the PCI bus can communicate directly with the memory. Benefits of PCI bus are high-speed, independance of applied processor, adaptation to new computer architectures, automatic configuration and good definition. The PCI bus supports a 'plug and play' technology which is its important advantage. [1]

Significantly improved PCI bus is a PCI Express bus that transmits data over high speed serial channels (corridor) and in essence can support multiple graphics devices. Speed per a corridor was significantly increased compared to the standard PCI bus. Within the corridors the transmission is serial with a significantly improved synchronization system. When the PCI Express bus uses more corridors for the transmission of data it performes parallel transmission data groups, which means that data has elements of parallel and serial communication. [3]

AGP bus (Accelerated / Advanced Graphic Port) is a bus connector primarily intended to display images (graphics card). Requirements for data throughput of modern graphics cards are such that virtually all the available capacity of the PCI bus would be spent on serving the graphics card significantly slowing the exchange of data with other additional circuitry. Therefore, one AGP bus is added the computer and is then used to display data on a monitor. [1]

VI. MOTHERBOARD MODELS

There are several models of motherboards, and when we talk about the model motherboards usually entails format, size and arrangement of elements on the motherboard. Differences between models of motherboards are physical size and shape, mounting locations, layouts, power connectors, and more. There are dozens of official models of motherboards, but only a few of them dominated the PC industry. [4] In the following paragraphs you can read more about these models of motherboards:

- Backplane Systems - components that are normally found on the motherboard are located on the adapter card that is plugged into the slot. Backplane systems come in two forms: passive and active. Passive design means that the main panel does not include any electrical connections other than bus connectors and maybe some compounds for powertrains. In the active form panel contains all the classic motherboard components except the processor. The advantage of a backplane system is that it can be easily upgraded with new processors and other properties just by changing the card but are often more expensive to upgrade. [3]
- Full-size AT - coincides with the original IBM AT motherboard design. First appeared in August 1984 when IBM introduced the IMB PC AT (Personal Computer Advanced Technology). Shortly after that reducing the size of components started to progress and thus decreasing the need for large motherboards. [2]
- Baby-AT - can completely replace the full-size AT motherboard and can be placed in several types of housing. Because of the flexibility of the Baby-AT that motherboard model was the most popular model until early 1996 when the more powerful ATX (Advanced Technology Extended) motherboard design, which is not directly interchangeable, began to replace it. [2]
- LPX - LPX (Low Profile eXended) and mini-LPX motherboard model is semi proprietary design which was originally developed by Western Digital 1987 for some of their motherboards. Most

noticeable feature is that the slots for additional extensions are placed on a raised card that plug into the motherboard. The slots are located on one or both sides raised card, depending on the system and design of the casing. [2]

- NLX - an enhanced version of the LPX design except that the NLX (New Low profile eXtended) is fully standardized, which means that one NLX motherboard can be replaced with another that is not necessarily from the same manufacturer, which could not be done with LPX motherboards. The main feature of the system is that NLX motherboard is inserted into a raised card unlike LPX systems where raised card plugges into the motherboard. All devices that normally connect to the motherboard connect to the raised card. Therefore, the motherboard can be extracted from the system without disrupting the raised card or any expansion card that is located on it. [2]
- BTX - Intel announces BTX (Balanced Technology eXtended) model in 2003 with the purpose to alleviate the increasing strength of components and cooling requirements and to enable better targeting of the circuit. Because of the new trend for dual-processor the need for BTX model decreases and in 2006 Intel gives up from further development. BTX motherboard model is not compatible with ATX model or some other design. The input / output connectors on the back are similar for BTX models and the ATX model but they are on the other side of the motherboard. Some major manufacturers have never even started to produce BTX systems and held to ATX-based systems. [2]
- ATX - ATX form is essentially a Baby-AT design rotated by 90 º. Expansion slots are now parallel to the shorter side of the motherboard and do not disturb the CPU, memory and input/output terminals. Intel also prescribed Mini-ATX, which is not an official standard as the two official versions microATX and FlexATX models that are intended for cheaper and more consumer-oriented systems and personal computers. They can use the specialised SFX / TFX (Small Form Factor / Thin Form Factor) power supply because they require less energy. [2]

VII. DESCRIPTION AND PURPOSE OF RESEARCH AND METHODOLOGY

The aim of this study was to determine how well the PC users know their motherboard, have they met with malfunctions and determine the incidence of occurrence of failures on the motherboard in each category of using a computer. According with this goal set following hypothesis:

H01: There is no statistically significant difference in the occurrence of malfunctions on motherboards with respect to the purpose of computer use.

This study was conducted using a survey method of testing. The survey was made with Google's service for creating documents 'Google Docs'. Questions in the survey are composed to give everyone an opportunity to reply, they do not engage in extensive and unnecessary detail, and provide a quality way to obtain the required information from the user.

The research results have been collected in an on-line database of Google's tools to create online surveys. The survey was conducted from June to August 2012. Survey was anonymous and collected 328 completed questionnaires.

Data were analyzed using descriptive statistics, and to test the hypothesis using the chi-square test. Statistical analysis of data was made in Statistica8.

VIII. DATA AND ANALYSIS

Table I. contains the frequencies of answers that are grouped by age. The numbers in the fields indicate the number of a certain group of people who responded to this question. Most respondents were younger than 25 years, and then followed by people between 26 and 35 years of age. It is reasonable that most participants in the first group are students while in the second group most of them are employed and the relatively even distribution of students and the unemployed respondents. Most participants own both a PC and a laptop, followed by those who only have a PC and last are laptop owners. In almost all groups, participants indicated that they spend more than 4 hours per day on a computer, and the average age of a computer in all of the groups is 3 years or less. Most common use of computers for the group under the age of 25 years for fun, 79 out of 171 participants, followed by communication and education, and work is in last place. In the second group, most people (54) also chose fun, 30 people chose work and 23 people chose communication. Education is in the last place. The next groups also use their computer mostly for the work while the other categories are poorly represented.

Acording to the type of the chipset Intel is leading in all groups, followed by AMD. A relatively small number of people do not know what chipset they have, and only a few people have chosen the option 'Other'. These responses are: nVidia (3 persons), Gigabyte, ASRock, and nForce 4.

Chart I. shows the response to the question of whether there were any malfunctions on the motherboard where it shows that in the first age group, more respondents said no although quite a large number of people responded affirmatively to this question. In second age group, the situation is reversed. More respondents encountered motherboard failure than those who did not. The third age group also has more affirmative answer.

TABLE I. TABLE WITH DANA

DESCRIPTION	Respondents age groups				
	Less than 25	26 – 35	36 – 45	46 – 55	More than 55
Total	**171**	**116**	**34**	**5**	**2**
Sex					
Male	127	104	30	3	2
Female	44	12	4	2	0
Level of education					
Primary school	3	3	0	0	1
High school degree	134	62	27	4	0
University degree or Bachelor	27	25	3	0	0
Specialist or Master degree	6	24	4	1	0
Mr. sc or Dr. sc	1	2	0	0	1
Current situation					
Student	140	37	7	1	0
Employed	13	50	23	3	0
Unemployed	18	29	4	1	2
Type of computer					
PC	56	44	17	1	1
Laptop	28	9	3	1	0
Both	87	63	14	3	1
Time spent on a computer per day					
Less than 1 hour	7	0	1	1	0
1 – 2	14	7	5	1	0
2 – 3	45	10	4	0	2
3 – 4	28	26	6	0	0
More than 4 hours	77	73	18	3	0
Age of computer in years					
1 or less	61	46	12	0	0
2	38	27	4	2	0
3	34	15	6	2	1
4	18	16	4	1	0
5 or more	20	12	8	0	1
Purpose of use					
For work	16	30	18	2	0
For communication	40	23	6	1	1
For education	36	19	6	2	1
For fun	79	54	4	0	0
Type of chipset					
AMD	40	46	13	2	0
Intel	109	76	19	3	2
Asus	14	6	5	1	0
Via Technologies	5	2	1	0	0
I dont know	13	5	4	0	0
Other	3	2	1	0	0
Was there any malfunction to the MB?					
Yes	70	69	21	2	0
No	101	37	13	3	2
If so, what malfunctioned?					
Power connector	18	13	10	1	0
Socket	12	3	2	0	0
CMOS battery	18	16	9	0	0
BIOS	10	15	5	0	0
Other	21	33	5	0	0
Did you change the MB because of a failure?					
Yes	49	54	13	1	0
No	122	62	21	4	2
Did you change the MB regardless of the failure?					
Yes	51	55	13	1	0
No	120	61	21	4	2

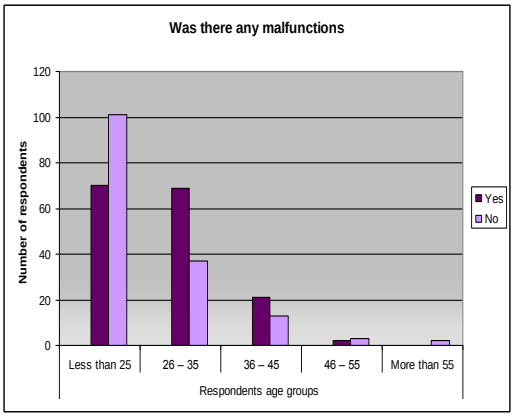


CHART I. OVERVIEW OF MP FAILURES BY AGE GROUP

Chart II. contains answers to the most common failures that the respondents had encountered. In the first 3 groups most answers were the power connector and the CMOS battery, but the largest number of respondents chose 'Other' where they had to enter themselves which component is broken. That was the aim of the questions and provoked the greatest interest. The answeres are obtained in table II.

From the answers to two questions: whether the respondents had to change the motherboard because of a failure and whether they change the motherboard, regardless of the failure it can be seen that most of the respondents in the first group did not change the motherboard because of a failure, nor have they changed it regardless of it. In the second group of answers are the similar, although a little more answered no on both questions. In the other groups, the answers are also mostly no. It is interesting that some respondents changed the motherboard regardless of any failures, thus it can be concluded that they upgrade and enhance their computer and follow the development of technology.

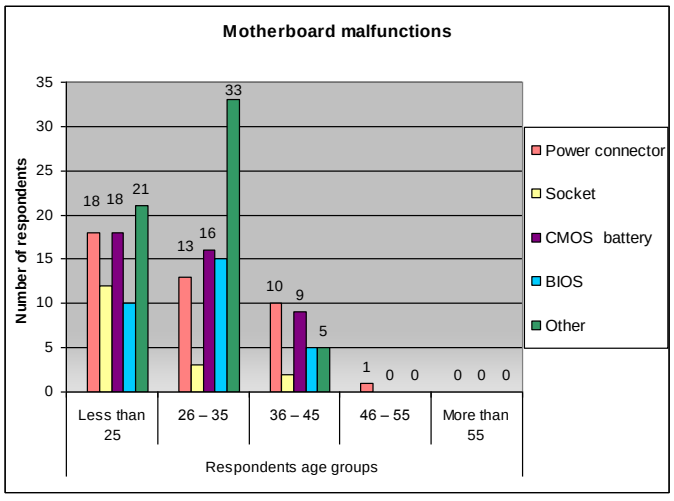


CHART II. THE MOST FREQUENT MALFUNCTION IN MP BY AGE GROUP

TABLE II. VIEW OTHER FAILURES

Other:
Capacitors (15)
Memory slots (9)
Network interface card (9)
I don't know (6)
PCI-e slot (5)
Northbridge (5)
Lightning strike (5)
Voltage regulator (4)
Chipset (6)
Sound card (3)
Grafhics card (2)
Graphics card slot
Cooler
Default error

TABLE IV. EXPECTED FREQUENCY OF FAILURES BY FUNCTION OF COMPUTER USE

Summary Table: Expected Frequencies (O1) Marked cells have counts > 10 Pearson Chi-square: 6,05817, df=3, p=,108821			
computer use	Var2 no	Var2 yes	Row Totals
For fun	69,335	67,664	137,000
For communicatio	30,872	30,128	61,000
For work	33,402	32,597	66,000
For education	32,390	31,609	64,000
All Grps	166,000	162,000	328,000

A. *Proof of the hypothesis*

The proposed hypothesis was to determine whether there is a statistically significant difference in the occurrence of malfunctions on motherboards with regard to the purpose of using the computer. Table III. i IV. shows a comparison of the observed and expected frequency of malfunctions by category of computer use. Limit value is chi-square with 3 degrees of freedom at a significance level of 5% of 7.815, while the calculated value is the chi-square (χ^2(3, N=328)=6,05817; p=0,05).

Since the calculated chi-square is less than the limit value (6.05817 <7.815), accepts the set null hypothesis and conclude that there is no statistically significant difference in the occurrence of malfunction in the computer's motherboard that are used for different purposes.

IX. CONCLUSION

The main components of the computer are located on the motherboard, so the motherboard is often sold separately to allow users to assemble a PC according to their needs. The most irreplaceable part of the motherboard is chipset that determines the processor type, the type of memory, bus and method of data exchange with the environment.

TABLE III. PLOT OF THE OBSERVED FAILURE FREQUENCY BY FUNCTION OF COMPUTER

	Summary Observed Frequencies (Marked cells have counts > 10		
computer use	Var2 no	Var2 yes	Row Totals
For fun	73	64	137
For communicatic	37	24	61
For work	27	39	66
For education	29	35	64
Totals	166	162	328

The results showed that with the increasing age of the respondents the use of a computer moves from entertainment exclusively for business purposes. Middle-aged respondents most commonly encountered with the motherboard malfunctions and their replacement due to failure. It is interesting that respondents in almost all age groups are changing their motherboard although there are not any faults which suggest that respondents upgrade their computer and follow the development of new technologies. It is confirmed that the appearance of failures on the motherboard is not a consequence of our communication via computer, or when you use it for fun, or when it enable us to carry out our important business tasks, and even when we use it to further education.

REFERENCES

[1] Grundler, M., „Kako radi računalo“, PRO-MIL, Varaždin, 2004.

[2] Mueller, S., „Upgrading and repairing PCs 20th ed“, QUE, USA Indiana, 2011.

[3] Mueller, S., „Upgrading and repairing PC's 17th ed“, 2006., http://flylib.com/books/en/4.57.1.40/1/

[4] „Motherboard Form Factors“, motherboards.org, 2004., http://www.motherboards.org/articles/tech-planations/4_1.html

[5] „Organizacija mikroračunalnog susatava“, http://www.informatika.buzdo.com/s125.htm

[6] Wilson, T. V., Johnson, R., „How Motherboards Work“, HowStuffWorks, http://computer.howstuffworks.com/motherboard.htm

[7] Comp TIAA+ Certification Study Guide, „Personal Computer Components – Motherboards and Processors“, pdf, http://www.mhprofessional.com/downloads/products/0071701451/0071701451_Chap01.pdf

[8] Fisher, T., „What's a motherboard“, About.com Guide, http://pcsupport.about.com/od/componentprofiles/p/p_mobo.htm

[9] PCTehcGuide.com, „What is a Motherboard?“, http://www.pctechguide.com/01mboards.htm

QR codes in library - does anyone use them?

Ivan Jelić* and Dina Vrkić**
*Zagreb City Libraries, Zagreb, Croatia
** Central Medical Library/University of Zagreb School of Medicine, Zagreb, Croatia
E-mail: ivan.jelic@kgz.hr, dina.vrkic@gmail.com

Abstract - Although they were initially intended exclusively for use in shipping industry, Quick Response code is gaining popularity with widespread use of smartphones and has began to be used and implemented in various types of libraries for engaging and assisting patrons. Absolutely free, allow access to more data than standard bar-codes, can be read with almost all types of smartphones and with great potential to bridge the gap between digital and physical objects. Sounds good, but is it? In this paper we will overview current usability and future applications of QR codes in libraries.

I. INTRODUCTION

Today it is clear that the most widely used device in the world is the mobile (smart)phone. Phones are mostly voice-centric devices, but a wide range of mobile devices now exist on the market offering multiple services and functions. The term smartphone is now used to characterize a mobile phone with special computer-enabled features. Despite the mobile phone evolution, one of the main disadvantages of these devices is that they still have insufficient input capabilities, providing tiny keyboards to do manual entries. Fortunately most smartphones are now equipped with several sensors that can be used to enhance and create new users interfaces. This is the case for integrated cameras that can be used to read visual codes, like Quick Response (QR) code as well as other sensing technologies such as Radio Frequency Identification (RFID) and the associated Near Field Communications (NFC). The use of these technologies does not only facilitate entering information, but it also allows using mobile phones for interactions with people, places and things, enhancing the usability and usefulness of these devices. The use of QR codes and RFID tags has significantly evolve in the last decades, they were first used to track products in the industry, but now they have contributed to develop several new concepts that integrate the physical world with the virtual one [1]. Essentially, QR codes are a low-threshold technology, convenient way to add the virtual to the physical and to provide useful content, often at the time of need [2]. But are they actually useful?

In this paper we will review possible implementation methods of QR code technology in libraries from aspect of usability and convenience. As this technology is low-cost, easy to implement and easy to use there is no doubt that in the future it will be even more in libraries, but it is difficult to predict if it will ever get wide acceptance by library patrons.

II. ABOUT QUICK RESPONSE TECHNOLOGY

Quick Response code or QR code is a two-dimensional (2D) bar code developed in 1994 by Denso Wave Corporation; QR code got this name because it was developed to improve the reading speed of complex-structured 2D barcodes. This type of code was initially used for tracking inventory in vehicle parts manufacturing. QR code is established as an ISO standard, it has been defined in the Information technology — Automatic identification and data capture techniques — QR code 2005 bar code symbology specification (ISO/IEC18004). QR code is free to use and the technology is open since its specification is disclosed and the patent right owned by Denso Wave is not exercised (Denso Wave Incorporated, n.d.).

The main characteristic of a QR code compared with a traditional bar code is that it contains information in both the vertical and horizontal directions, while a bar code contains data in one direction only. For this reason QR code holds a considerably greater volume of information. In addition it can encode several types of data including symbols, control codes, binary data, and multimedia data. The typical barcode holds a maximum of 20 digits, while the maximum data capacities of a QR code are 7,089 characters for numeric data, 4,296 characters for alphanumeric data, 2,953 bytes for binary data, and 1,817 characters for Japanese Kanji and Kana data.

FIGURE I. Comparison of QR code and Bar code [3].

QR code is faster to read than other two-dimensional code, because it contains three large square patterns in the corners that are used for position detection. Additionally, the patterns are used to detect the size, the angle and the outer shape of the symbol. When a reader scans a symbol, it first detects these patterns. Once the position patterns have been detected the scanner can rapidly read the inside-code in all directions. The inside code consists of several blocks where the information is encoded [1]. As suits something developed for demanding, fast moving environment, they included generous error correction

capabilities and QR codes can be set up to still work even if 30% of the symbol is damaged or obscured [4].

FIGURE II. Structure and components of QR code [1].

The elements contained in a QR code are the following:

- Pattern. Three big squares in the corners used for detecting the position, the size and the angle of the QR code.
- Alignment Pattern. A pattern used for correcting the distortion of the QR code. These distortions could occur for example when attaching the codes onto a curved surface.
- Timing Pattern. It consists in white and black modules arranged alternately and placed between two position patterns. It is used to determine the central coordinate of each cell in the QR code.
- Quiet Zone. A margin space that makes easier to detect the QR code. At least four cells are required for the quiet zone.
- Data Area. The area in the QR code that contains the data (for example a URL) encoded in binary numbers. The data area also includes Reed-Solomon codes to provide error correction functionalities [1].

III. COMPARISON WITH SIMILAR TECHNOLOGY

The Internet of Things (IoT) enabled users to bring physical objects into the sphere of cyber world. This was made possible by different tagging technologies like NFC, RFID and 2D barcode which allowed physical objects to be identified and referred over the Internet. The emergence of Internet of Things has provided an initiative to link our cyber lives with the physical ones by exploiting low-cost embedded tags, i.e., RFID, NFC, 2D barcode. Due to less complexity and low development and deployment cost of 2D barcodes, they have become the primary tool to create a linkage between physical objects and their cyber representations [5]. There is widespread assumption that the increase of NFC enabled mobile phones will replace the use of QR codes, but it is more likely that both technologies will coexist for some time.

The best definition of Virtual Reality (VR) is when participants are completely immersed in a fictitious world that can take some properties of the real world but that replaces it with a simulated one. The road from the real world to a virtual one may include other environments where the participants are still connected with a real world while enhancing it with other contents, merging both worlds in some way. This concept can be referred as Augmented Reality. Augmented reality (AR) refers to augmenting the real-world with content generated with a computer or another device. This content can include sound, video, graphics, and other data [1].

RFID uses radio waves to store and retrieve data from an identification chip. These chips are known as RFID tags and they provide limited information, but connected to the wealth of information that our libraries hold on our stock and our users and thus they become smart objects, acting as mobile connectors to a world of usage and behaviour information. RFID is now widely used in the industry for several applications including security, access control, transportation and tracking of the supply chain. Also in libraries they are primarily used to label stock and have tended to be used in much the same ways as barcodes previously. They have made it easier for library users to issue and return items through self service equipment and some libraries have used them to aid stock taking. There is so much more that we could do with them though, using them to display some of the data linked to our physical items that is normally inaccessible to users [6].

Usually an RFID system requires three main components: the reader/writer, RFID tag and application software for processing the information. The RFID reader comprises an antenna, a transceiver and decoder. The reader periodically transmits signals to search for tags in their vicinity. When it captures a signal from a tag, it extracts the information and passes the data to the processing subsystem. An RFID tag or transponder consists of an antenna, a radio transceiver and integrated circuit for storing and processing information. There are several types of tags. A tag contains writable memory where data is stored to be transfer later to RFID readers. The internal memory capacity of a tag depends on its model and varies from tens to thousands of bytes. RFID technology is classified into the short-range wireless communications, which are systems that cover distances of less than 100 meters [1].

Near Field Communication (NFC) is also a short-range high frequency wireless communication technology which enables the exchange of data between devices at distances fewer than 10 cm. This technology is an upgrade to RFID technology; it was designed and marketed by the NFC Forum. The International Standard Near Field Communication - Interface and Protocol, ISO/IEC 18092 (NFCIP-1), defines communication modes for NFC interface and protocol. According to this standard, NFC can operate in active or passive mode. In active mode, the devices generate their own electromagnetic field independently, while in passive mode only one of the devices is capable of generating an electromagnetic field and the other extracts energy from it to operate and transmit the required information. One of the key elements of NFC enabled devices is the ability to read different types of tags. This facility of NFC technology is a key enabler for many applications. NFC tags are passive devices with no power of their own. In order to read a tag

the users almost touches it with an NFC-enabled device. A small amount of power is taken by the NFC tag from the reader/writer to power the tag electronics. The tag is then enabled to transfer a small amount of information to the NFC reader [1]. NFC could maybe have knocked QR codes aside if Apple had decided to include NFC in its mobile devices. NFC is less maligned but more precipitous. Being kept off Apple's devices has effectively kept NFC from penetrating the mainstream technology market.

Augmented Reality has not seen the pick-up predicted by its potential, though that may change with the growth of some types of wearable computing technology [7]. AR is another means of integrating the digital and physical world, though in a much more interactive manner than with QR codes. AR is best defined as an overlay or augmentation of the physical environment with sounds, graphics and perhaps location-based information (geolocation) [8]. A person uses applications, such as Google Goggles or Layar, that use a mobile device‟s camera to identify objects and provide more information on the object, layer information over the view from the device‟s camera or locate the device and provide information based on the user`s location, and thus provide more interactive experience and enhance perceptions of reality [9].

IV. ANALYSIS OF QR CODE USE IN LIBRARIES

It is safe to say that library patrons in social driven world want to feel connected to and involved with the information, they are more social now than ever before. We need to encourage this type of interaction between information the library provides and the users. So, QR codes are recognized by many as means of encouraging this type of interaction. Of the examples we reviewed, the most useful and successful implementations are those that provide some value-added service for the users. Adding value means that using the QR code is better or easier than using the existing service or that the library is providing a service that is uniquely useful via QR codes [10].

Some reasons for implementing QR codes in libraries are:

- to dynamize relationship with patrons,
- to give an extra access point to website,
- to improve the image of the library [11],
- linking from print to electronic journal holdings, on the boxes containing back copies of print journals, together with the display stands holding current copies, QR codes took users to link resolver, searching electronic holdings and linking to the appropriate electronic journals where available,
- library exhibits that include a QR code link to songs, videos, Web sites, surveys, etc. or other information that augments the exhibits,
- linking to library audio tours for orientations,
- added to print handouts for additional information on mobile friendly sites,
- QR code with text that loads the library's text message reference service and other contact information into the patron's phone,
- art shows or permanent art in libraries with a QR code linking to the artists Web sites,
- in catalog record to offer patrons basic info about an item, including the location and call number, users can scan the code and head to the stacks rather than writing or printing,
- taped to video DVD cases, linking to mobile-friendly video trailers,
- code placed on audio book cases for author interviews or books for reviews,
- code placed on study room doors connecting to room reservation forms [2],
- providing an electronic alternative to physical books,
- promoting online audiovisual materials,
- embedding video help,
- bringing external resources into the library,
- QR codes displayed across the library that linked to help desk, prompting the users phone to ring IT help; library help; or send an SMS to our "text a librarian", whichever was appropriate for the location [6], and
- to embed QR codes in the physical space, people can scan the codes and instantly engage with the library via social media.

One of the initial uses of QR codes has been within library catalogue records, and also one of the most popular, including the University of Ryerson Library and Archives in Toronto, San Diego State University Library and the University of Bath and Huddersfield Libraries. The codes are embedded in each item record, and, when scanned, send shelf location information to the user`s mobile device. Such information can vary, but typically includes basic bibliographic information such as title and author, as well as call number, floor and library building. By using QR codes, library users do not need to record the exact location information as this data has been stored on their mobile device, and they are certain to have all of the necessary information to find the item in the library`s collection [9].

The University of Bath is the predecessor of applying QR codes in education. They have incorporated this technology in several aspects related with the learning process. For instance, they use QR codes in the library to provide information about the books. They have also developed an enhancement for Moodle which automatically includes the QR code for the page that has been printed. The QR code contains the URL of the page on that particular Moodle course. They have also added QR codes on posters that can be found around the campus,

on Websites and service blogs for bookmarking, in handbooks linking to activities, and in marketing materials from departments.

According to Susono and Shimomura [12] in Japan almost 100% of college students have mobile phones, for this reason they are commonly used for education purposes. In their work they reported the use of mobile phones and QR codes to conduct surveys during class, this with the intention of providing feedback to the teacher at the middle of a long class (i.e. 90 minutes). With this project students answer a survey using their mobile devices and QR codes to choose from different options. They send the answer to a server and the teacher can have immediate feedback in order to improve his/her class if needed.

However, there are certainly some downsides to implementing QR codes in libraries that we have to mention. QR codes are only one application among many on devices that are relatively new, they are not well known. Users need to be educated about what the codes are and their use. Scanning in poor light or off a computer screen can be troublesome, as some camera phones cannot read the code properly and it becomes blurry. Furthermore, the more information stored in a code (for example, a long URL), the more complicated (or pixelated) it will be, and thus harder to scan. Thus, the inconvenience associated with scanning a QR code (downloading an application, opening the application and then scanning the code) may be too great, depending on the situation. The difficulty associated with the installation of a third-party application to a mobile device (the QR code reader), as well as needing to open the application on the device, scan the code and then load the information contained within the code is still a barrier (real or perceived) that can inhibit their use.

Another barrier is convincing users that scanning the QR code will provide them with valuable information or an enhanced experience. In most of projects that implemented QR codes in libraries there is huge disappointment with the actual usage of new service. In the analysis of pilot project carried out in the Eda Kuhn Loeb Music Library of the Harvard College Library Wilson [13] states that "none of the three on-line resources were viewed via QR codes more than five times each over the course of the entire semester, and the actual utility of those page views was minimal, at best. Of the three sites, only the "Finding Concert Reviews in Periodicals" appears to have been accessed for use, as the other two research guides had only single page-views, and no recorded time on the sites themselves. Legacy and current usage statistics indicate that the sites are being used, with anywhere from 31 to 53 site visits over each of the past two academic semesters, but once the data is examined at the platform level, mobile usage was negligible in comparison to conventional on-line access." Among the likely reasons for the lack of acceptance of QR codes in this pilot project are a wider than previously believed misunderstanding of what QR codes are and what they can do. Despite their ubiquity in the public space, a significant portion of the population appear not to know exactly what they are, or even what the term QR code means. Abuse of the QR code technology to gather information on users is another potential negative. While QR code generators may be downloaded for free, it has been reported that some generators collect data about users before redirecting the user to the desired web site, and there is also a possibility of linking to malicious websites, and providing new ways of luring people into phishing scams [14].

V. CONCLUSION

When planning implementation of QR codes in library it is necessary to pay attention on potential short-falls that include: real world usage, their reliance on the scanner being stable and steady, size and distance sensitivity, but much of the argument in favor of QR codes in the library comes down to a simple cost/benefit analysis. And in this case, as long as a few simple rules are followed, the cost of employing QR codes is so low that any benefit derived from them outweighs the minimal effort involved.

There is a reason that QR codes have become so ubiquitous in print advertising, points-of-sale, and other venues: they are so easy to use, and cost so little in terms of resources, time and money, that despite low acceptance by the public, it is a technology simply too easy to ignore [13].

Bearing in mind that QR codes should not be used just for the sake of the technology, we need to ensure the learning design focuses very much on the appropriate use of QR codes where it adds value to the learning experience over and above other delivery methods.

REFERENCES

[1] I. Deliyannis, Interactive Multimedia. Rijeka: InTech, 2012.

[2] R. Ashford, "QR codes and academic libraries, " College Research Libraries News, vol. 71(10), pp. 526-530, 2010.

[3] QR Code.com, available at: http://www.qrcode.com/en/aboutqr.html. [2013-01-15].

[4] A. Walsh, "Quick Response codes and libraries," Library Hi Tech News, vol. 26(5/6), pp. 7-9, 2009.

[5] F. Razzak, "Spamming the Internet of things: a possibility and its probable solution," Procedia Computer Science, vol.10, pp. 658-665, 2012.

[6] A. Walsh, "Blurring the boundaries between our physical and electronic libraries: location-aware technologies, QR codes and RFID tags, " The Electronic Library, vol. 29(4), pp. 429-437, 2011.

[7] Joe Murphy – Librarian, Innovator. *The Death of QR Codes in Greatly Misunderstood*, available at: http://joemurphylibraryfuture.com/the-death-of-qr-codes-is-greatly-misunderstood/. [2013-01-09].

[8] D. Kane; J. Schneidewind, "QR codes as finding aides: linking electronic and print library resources, " Public Services Quarterly, vol. 7(3-4), pp. 111-124, 2011.

[9] T. Badham, "QR codes: do they provide the missing link between the physical and digital?, "" VALA 2012, 2012.

[10] M. J. Whitchurch, "QR codes and library engagement, " Bulletin of the American Society for Information Science Technology, vol. 38(8), pp. 14-17, 2011.

[11] D. Pons; R. Valles; M. Abarca, F. Rubio, "QR codes in use: the experience at the UPV Library, " Serials, vol. 24(11), pp. 47-56, 2011.

[12] H. Susono; T. Shimomura, "Using mobile phones and QR codes for formative class assessment, " Current Developments in Technology-Assisted Education, vol. 2, pp. 1006-1010, 2006.

[13] A. Wilson, “QR codes in the library: are they worth the effort? Analysis of a QR code pilot project accesed, ” Journal od Access Services, vol. 9(3), pp. 101-110, 2012.

[14] D. W. Jackson, “Standard bar codes beware – smartphone users may prefer QR codes, ” Law Library Journal, vol. 103(1), pp. 153-158, 2011.

Systems of support for persons with complex communication needs

Z. Stančić, J.* J Frey Škrinjar*, Ž.Car**, V.Vlahović Štetić***J.Pibernik****
* University of Zagreb Faculty of Education and Rehabilitation Sciences
*Department of Inclusive Education and rehabilitation
zstancic@erf.hr, frey@erf.hr
** University of Zagreb Faculty of Electrical Engineering and Computing
Department of Telecommunication
zeljka.car@fer.hr
*** University of Zagreb Faculty of Humanities and Social Sciences
Department of Psychology
vvlahovi@ffzg.hr
**** University of Zagreb Faculty of Graphic Arts
jpiberni@grf.hr

Abstract - The aim of this paper is to present the results of multidisciplinary cooperation of university scientists in the field of electronic engineering and computer science, education and rehabilitation science, psychology and graphic technology, focused on developing assistive technology for persons with complex communication needs. Based on the needs of participants with complex communication needs and final results of project, several services were developed: Body of Knowledge and three applications for ipad such as „Communicator" for users to establish communication, „Mathematical carousel" for learning basic mathematical operations and „e-Gallery" to improve the development of communication skills and skills for sentence ordering and short story telling.

I. INTRODUCTION

Augmentative and alternative communication (AAC) methods - based on graphic and textual symbols rather than written words to represent certain objects, actions, or concepts - are highly beneficial and in certain cases crucial for speech, literacy, learning, employment and quality of life for persons with complex communication needs. According to demographic data there are 3 to 5 potential AAC users with complex communication needs per 1000 users of general population. The explanation for that considerable number of AAC users with complex communication needs is different nature of their educational needs such as Down syndrome, Autistic Spectrum Disorder, Alzheimer's disease, severe intellectual disabilities, complex physical disabilities. Experts, caretakers, and user groups contributed to growth of rich market for development, customization, and maintenance of Information and Communications Technology **(**ICT**)** that enabled AAC services in Croatia. Such services can fulfill a wide range of needs, allowing users to achieve a higher degree of accountability, planning, organizing ideas, efficient execution of activities, increased creativity, communication skills and literacy. Fulfillment of such needs further leads towards increased employment potential and in general improved quality of life.

This paper addresses the question what works and what is essential for augmentative and alternative communication of persons with the complex communication needs.

University of Zagreb Foundation for Development and Program "Together" call for proposals was set at the end of 2010. This was a reason to initiate a common project application created by four faculties with an outstanding purpose to start new values in research [1], contributing the problem area of AAC. Project was conducted in collaboration of Faculty of Electrical Engineering and Computing Department of Telecommunication, Faculty of Education and Rehabilitation Sciences, Faculty of Graphic Arts and Faculty of Philosophy, Department of Psychology, all from University of Zagreb.

Each of the faculties is known for its specific field of scientific research [2], e.g. research and development of information services and application, supported communication and individualized support, technical and technological creative actions in creating and transforming graphic media, applied psychology. Several non-governmental associations participated and contributed to the project. Those associations were: Croatian Association for Down syndrome, Association of Parents of Visually Impaired and Multi Disabilities Children - OKO, Association for Persons with Intellectual and Communication Impairments- AmoReVera, Croatian Association for Early Childhood Intervention and Association for Wolf-Hirschhorn Syndrome.

The aims of the project were:

- establishing multidisciplinary cooperation in the area of electronic, computing, educational

rehabilitation, speech pathology, psychology and graphic technology,

- directing research resources in enhancing research capacity of the four university members
- development of the model of communication based in ICT - technology and the proposal for technological development for development of e-inclusion services.

Results of the project can be best presented through eight established indicators as following:

1. The network of scientists of different profiles has been established, the network of their contacts and cooperation has been enlarged by variety of representatives from each university faculty member.
2. The Body of knowledge on supported (AAC) communication (http://www.ict-aac.hr) has been made in a form of a browser of definitions and references with existing sources of information on AA communication. This catalogue can serve wide range of users (scientists in the field, students, parents of users, experts in different areas, etc.). In order to present the area of AA communication even better, the plan is to constantly add new content.
3. The model of technological platform for development of e-accessibility has been defined.
4. The new model of graphic sign system for persons with complex communication needs has been defined.
5. Three pilot ICT - services for iOS and Apple tablets such as „Communicator“, „Mathematical carousel" and "e-Gallery" have been developed. Technical results of evaluation are presented later in this paper.
6. Pilot application Mathematical playroom is being developed and it is designed for children before entering school aiming to alleviate their learning to count, numerical quantities and basic mathematic operation.
7. The results of the research have been published in domestic and international scientific publications.
8. Student works (seminars, thesis, doctoral thesis) with topics related to the project are defined.
9. Media promotion has been carried out in several tv-shows (Z1, RTL) - presenting interesting moments in everyday life of a boy using the applications developed within the project, weekly magazine „Glas Koncila“ and newspaper (Jutarnji list).

In this paper several services such as graphic symbol design, AAC application-user communication model and two pilot applications for ipad devices and the results will be presented and evaluated.

Project "ICT systems for persons with complex communication needs" was funded by University of Zagreb Development Fund in the period 1.1.2012-31.1.2013

II. The Design Of The Adaptive Symbol System

As we already mentioned the graphic symbol design is fundamental for the usability of a communication system, especially when users are children who need facilitation in associating graphic signs with objects, people and actions from their environment [3]. In order to create emotional connections and motivate children's interaction with the system in a pleasurable way, the design of the symbols has to be aesthetically pleasing and quickly recognizable [4].

This adaptive symbol system is designed by Faculty of Graphic Arts' experts to be implemented in future ICT services. It will be used on modern mobile devices, laptops, desktop computers and in print media; in other words, it will be optimized for various screens and printers [5] [6]. Some of the main advantages of this model, when compared to the other models, is its visual adaptability, responsiveness, consistency and possibility of animation. It is tailored to Croatian users, but can be adapted to other nationalities as well [7].

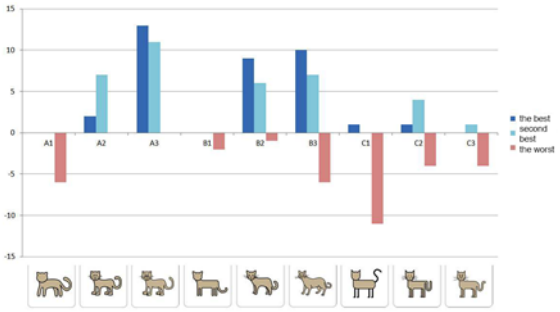


Figure 1. User's preferences in terms of symbol design

The evaluation of usability and accessibility is relatively new in the field of alternative and augmentative communication and for that reason the evaluation methods have to be adapted to persons with complex communication needs and their personalities [8]. In order to determine the principles of design, especially the complexity of the basic line construction and visual style, children with special needs ages 6 to 14 were taken as test subjects. Three different categories of symbols were created: animal, object and human face.

These results will determine the final design principles for the basic linear construction of the symbols (Figure 2). Also, the primary pallet of colors and the typography for symbol titles will be defined.

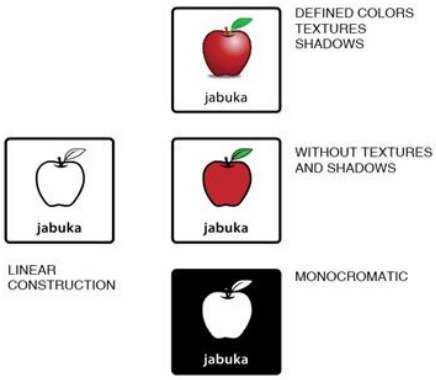


Figure 2. The Adaptive Symbol System Model

The participants were asked to rank nine prototypes for each category [9]. Prototypes differed in terms of complexity of line construction (simple, medium complex and complex) and style (technical, realistic and cartoon). Based on the research results we can conclude that users prefer complex and medium complex symbol design, while in terms of style their preferences tended towards realistic visual representation. Simple symbols ranked lowest with our test subjects (Figure 1).

III. AAC Application-User Interaction Model

This model presents possible interaction channels between AAC users and particular AAC application. It is applicable to any application that is or will be implemented within AAC software platform aimed for the component-based development and deployment of interoperable and scalable symbol based services [10].

Computer, smart phone or tablet enables interaction between AAC user and the AAC application. Input devices are mouse, keyboard, touchscreen and microphone. Output devices are screen and loudspeaker. Devices differ in the operating system, screen size, and the size of the device.

AAC applications' users could be either registered users who create their profile through registration process or non-registered users working in offline mode. Basic element of this model is an image. Augmentative and alternative communication is based on the use of graphic and textual symbols (rather than written words alone) to represent certain objects, actions, or concepts. There are available different commercial and non-commercial symbol galleries or collections of downloadable images representing different symbols. Also users can posses their own private symbol galleries with symbols having specific meaning for them (photos or specifically created images). Images are either stored locally on the device or downloaded from the Web. AAC applications should support loading of this images.

With each image, a text-label should be associated describing the meaning (word or the phrase) of a symbol. Text is entered into provided text editor with limited amount of characters. Each image has associated sound. For creating sound meaning of the symbol either text-to-speech is used or sound is downloaded from the web. Also, users should be able to record their own sound and save it as an audio description of the symbol.

Applications should support easy adding of new symbols or changing the existing by making new associated sound or text-label.

At the application output, images are displayed either sequentially or several symbols at once. Number of images displayed at one screen should be set as the parameter within the application. Text-labels turning on/off should be supported. Sound is associated with displayed symbol.

AAC application-user communication model is presented in the Figure 3.

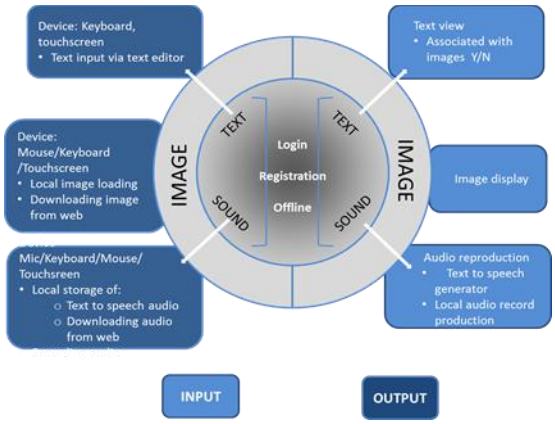


Figure 3. AAC application-user communication model

IV. Evaluation of Aplications Communicator and Mathematical Carousel

Based on the suggestions of experts from the Faculty of Education and Rehabilitation Sciences, experts from the Faculty of Graphic Arts have developed symbols and experts from the Faculty of Electrical Engineering and Computing have developed augmented communication applications for tablets. Two applications were developed: "Communicator" (application for children with complex communication needs) and application for help in learning mathematics in title "Mathematics carousel". Their evaluation was conducted by experts from the Faculty of Humanities and Social Sciences and Faculty of Education and Rehabilitation Sciences.

We have conducted study to evaluate new applications for ipad intended for work with children who have complex communication needs and learning difficulties. The purpose of evaluation was to improve the application so that it would be as adjusted as possible to users' needs.

Participants:

- four experts in the area of complex communication needs,
- 18 children with complex communication needs, age: from 2 to 6 years,
- 11 children with learning difficulties, age from 6 to 10 years.

Three phases of evaluation:

1. Focus group with experts
2.The experts used the application in one to three sessions with every child.
3. Focus group with experts

In every session they filled out the questionnaire regarding child's reactions to the application and parents' commentaries.

A. Results for Communicator

The average length of children's use of the Communicator application in three sessions was 20.1, 21.4 and 27.4 minutes, respectively.

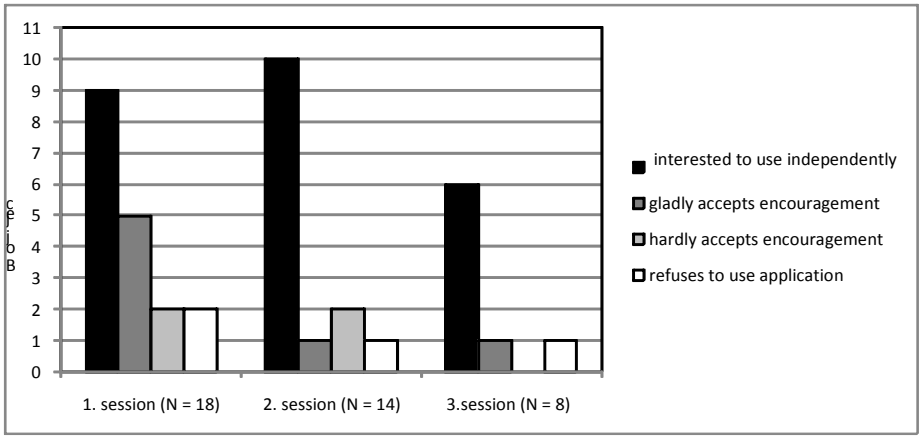


Figure 4. Interest to using Communicator application

The experts assessed children's interest in using the application. Most of the children showed very high interest to use application independently (Figure 4), and used it with positive emotional reactions.

Some of the expert's comments were:

"The option for photographing new items is great."

"The parents would like very much to get the application and use it with their child."

"The parents are very interested in buying the tablet and installing the application."

The experts' suggestions for the improvement of the application were:

- when the number of symbols is large, the application often „freezes",
- verbal reward (e.g. „That is correct") should follow the correct answer,
- symbols should be classified into different categories (e.g. clothing, activities, animals, etc.),
- font should be larger,
- female voice that reads symbols' names is clearer and easier to understand than the male voice.

B. Results for Mathematical carousel

Most children showed very high interest in application (Figure 5), especially for items with numbers and comments of children and parents were positive:

"She is looking forward to work with application. "

„He likes very much to have help and to hear praise for correct answer. "

Some of experts' suggestions for the improvement of the application were:

- symbols should be larger
- sign for equality should always be at the same place
- children do not understand symbol for zero
- different colors of squares in task and solution confuses children.

Although the applications was, due to technical problems (lack of tablets and „freezes" of the application), used by a relatively small number of children, we have obtained useful feedback information. Based on these data, the applications will be improved and available to users through the project's website.

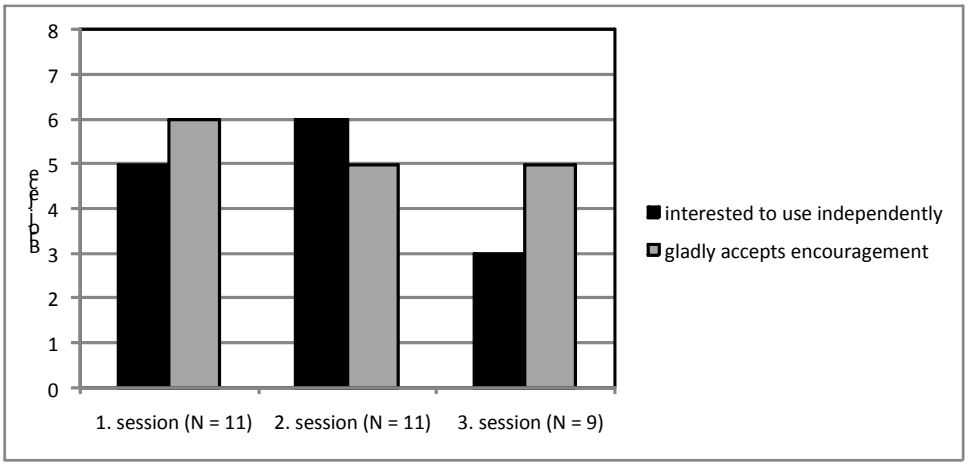


Figure 5. Interest to using Communicator application

V. COUNCLUSION

Our project and investigation brought us a network of different profiles and connect us to a network service of provider and user. For future this project created the foundation based on human right to communicate. New applications are not a magic wand because every child is unique. Applications can certainly be used in the process of teaching and learning.

REFERENCES

[1] Stančić, Z., J. Frey Škrinjar, M. Ljubešić, Ž. Car, "Multidisciplinary collaboration and ICT services for people with complex communication needs", Proceedings of the 34th International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO 2011, Opatija, 2011.

[2] Car, Ž. "Predstavljanje projekta „ICT sustavi za osobe sa složenim komunikacijskim potrebama“, Zbornik sažetaka završne konferencije projekta ICT sustavi za osobe sa složenim komunikacijskim potrebama, Zagreb, 2013.

[3] Garcia, M., A.N. Badre and J.T. Stasko, "Development and validation of icons varying in their abstractness", Interacting with Computers, I(2), pp. 191-211.

[4] Zwaga, H. and R. Easterby, "Developing effective symbols for public display", Proceedings of a NATO Conference on Visual Presentation of Information, Het Vennebes, Netherlands, 1978. Sussex: Wiley & Sons Ltd.

[5] Gittins, D., "Icon-based Human-Computer Interaction", International Journal of Man-Machine Studies 24, pp. 519-543.

[6] ISO 22727:2007. Graphical symbols – Creation and design of public information symbols – Requirements. Geneva: International Organization for Standardization

[7] Norman, D.A., "Cognitive Engineering" User Centered System Design, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 31-61.

[8] Vredenburg, K., J. Mao, P.W. Smith, T.A. Carey, "Survey of user-centered design practice". Proceedings of the CHI 2002, CHI Letters Vol. 4, No. 1, pp. 471-478.

[9] Bewley, W.L., T.L. Roberts, D. Schroit, and W.L. Verplank, "Human factors testing in the design of Xerox's 8010 'Star' office workstation". Proceedings of the ACM CHI'83 conference , Boston, MA, pp. 72–77.

[10] Car, Ž., M. Vuković, I. Vučak, J. Pibernik and J. Dolić, "A platform model for symbol based communication services", Proceedings of the 11th International Conference on Telecommunications ConTEL 2011, Graz University of Technology, Graz, 2011.

Framework for Web Application Domain Knowledge Extraction

I. Rožanc
Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, SI-1000 Ljubljana, Slovenia
igor.rozanc@fri.uni-lj.si

Abstract - A decade ago a web application e-Študent was built with aim to provide electronic support for student enrolment and examination/alumni records management at the University of Ljubljana. Due to issues emerging from the Bologna reform a new e-Študent is to be build using a modern technology in the near future. The old e-Študent encapsulates a huge amount of domain knowledge. Unfortunately, it was developed using agile approach resulting in poor technical documentation, thus an alternative approach for the domain knowledge extraction has to be defined. In the paper a framework for an effective web application domain knowledge extraction is defined. It has five elements. The main principles (1) of extraction are defined to perform effective reengineering of different application views at a defined abstract level. A proper knowledge representation using diverse models (2) has to be determined next, and the Model Driven Architecture using UML models is considered a suitable choice. The procedure (3) for extraction has to be defined using appropriate (usually custom made) tools (4) and performed by skilled staff (5), possibly members of the old development team. The use of framework is demonstrated on the web application e-Študent outlining several custom made tools, the results and the most valuable lessons learnt.

I. INTRODUCTION

In past decade the web student application e-Študent [1] proved to be one of most valuable programming assets at several faculties of the University of Ljubljana as its main functions encapsulate electronic support for student selection and enrolment, the student course selection management, detailed management of examination records and grades, the alumni records management, and several student data analysis executions.

After 2001 the three-tier web application e-Študent was built using the Oracle DB and Oracle Portal technology. It has been designed and developed at the Faculty of Computer and Information Science. It is written mainly in PL-SQL and JavaScript programming languages, and it consists of approximately 1100 different programming objects (namely dynamic pages, stored procedures, stored functions, triggers, and JavaScript code) with over 220 KLOC in total. It has 130 different reports, and its database contains nearly 120 tables. The developer team consisted of people who were themselves developers and users at the same time. By 2003 the initial release has been used by most faculties of the University of Ljubljana, consisting of more than 20000 students, lecturers and staff. The application's functionality has been systematically upgraded till 2006, and maintained later.

In 2008 the University of Ljubljana decided to replace e-Študent due to introduction of the Bologna reform. The structure of the new Bologna study programs was modified significantly, and a huge number of elective courses was introduced. The selection of those was not limited to single faculty as before, and the existing e-Študent was not able to efficiently support this issue. It was designed to be used by a single faculty and therefore different faculties were running their own instances instead of a single inter-faculty instance.

Unfortunately, the new system built after 2008 did not meet the expectations. It was too focused on implementing additional functionalities putting aside the suitable implementation of the existing i.e. essential ones. Consequently, the first e-Študent is still used at some faculties, but it is outdated due to old technology used. Additionally, in the last years it was maintained with idea of near-future replacement, thus without systematic introduction of some major improvements.

Nevertheless, it has been realized that the original e-Študent accumulated a huge amount of the domain knowledge. The e-Študent encapsulates probably the most comprehensive specification of the student examination processes at the technical faculties of the University of Ljubljana. It is precisely this key domain knowledge that should be implemented in the e-Študent successor, which has to be developed in the near future because of before mentioned reasons. The preliminary part of this initiative is the collection the e-Študent's domain knowledge and its presentation in a suitable form [2].

Technical documentation created during the systematic application development and updated during maintenance should be the primary description of the domain knowledge as it includes a software specification consisting of a number of artifacts precisely defining different aspects of the product. Providing such documentation is the easiest way to a suitable domain knowledge description, and usually no further activities are needed.

Unfortunately, in case of e-Študent (and in many other applications) software specification is mostly missing or outdated. The crucial cause of this is the agile methodology [3, 4] used during software development [5]. Although the principles of the "Manifesto for Agile Software" [6] proved successful during design and construction phase, they caused lack of documentation in later phases of e-Študent life cycle. Additionally, frequent requirement changes and an unsystematic maintenance resulted in practically unusable technical documentation.

When the application proves to be successful, it inevitably grows in both size and complexity, and the environment the application now lives in changes. Hence, the agile principle of omitting documentation to be more effective becomes an obstacle.

Thus, the domain knowledge has to be extracted from the code and other artifacts of currently active application. Reverse engineering of a software application seems to be a suitable choice as it is often performed to produce various kinds of documentation of the software application when the documentation is either outdated or even nonexistent. [7]. This process produces a number of design artifacts at a higher abstract level, usually leaving the less important details aside. The formally correct definition of reverse engineered models provides a clear insight into application description as well, and it can be used in design and development of a new system. This is in fact the essential result of this approach. However, the reverse engineering is usually a heavy duty process dependant of a number of parameters, with a strong tendency to be unique in each case. A number of questions has to be addressed before a suitable process is defined.

To effectively collect and describe domain knowledge we performed the reverse engineering of e-Študent application. Our objective was practical: to gain as much as possible results to be used in future new e-Študent development on a different platform using different technology. We had to do it in limited time, but we were not limited otherwise. Nevertheless, the process we performed forced us to think wider and to promote systematic common solutions. By performing reverse engineering of e-Študent application we were able to understand the process from different aspects, discuss different alternatives and extract the most important elements of a framework for the any web application domain knowledge extraction.

This paper consists of three sections. In second section the framework description outlined, first as the main idea and later in the subsections each of its five elements in detail: the main principles, the models, the extraction procedure, the tools, and the extraction team. Finally, the lessons learnt and a short conclusion is presented.

II. THE FRAMEWORK DESCRIPTION

The aim of the framework is to define an effective common approach for any web application domain knowledge extraction, when this has to be performed using reverse engineering activities. In this manner it is possible to systematically define why, what, how and who the extraction process will perform.

By using the framework we discover different views of the extraction process. For example, we understand the hierarchy of different activities/steps and their importance for achieving the correct final result, or alternatively we identify two distinct sets of activities. The first set of activities can be performed by less-skilled personnel using standard procedures, usually by the use of existing tools enhancing the tool use level. The second set of activities has to be executed by specialists in a particularly tailored way, usually by the use of particular approaches and custom made tools.

The framework (Figure 1) consists of five elements:

1. the main principles;
2. the models;
3. the extraction procedure;
4. the tools; and
5. the extraction team.

The main principles (1) outline the definition of extraction at the highest abstract level. The overall goal of extraction process, the type and amount of input artifacts, the expected results, the level of granularity, the level of automatisation and similar issues are addressed at this point. The main principles define WHY and WHAT is to be performed during the process of extraction.

The models (2) define a hierarchy of models or other definitions for description of domain knowledge. Different solutions might be applicable, but we promote the use of Model Driven Architecture using UML models as it is suitable for web applications description. The models define in WHAT WAY the final results will be presented.

The extraction procedure (3) systematically describes the process step by step. It follows the main principles, and it uses selected models supported by appropriate tools. The level of personnel involvement might vary depending mostly on the tools they use. By the procedure definition we answer on the HOW question.

The tools (4) perform all separate steps defined in the extraction procedure. They might be used at different levels: sometimes they perform the entire activity, while in other cases they only a part of it. Many times the existing commercial tools connected to the selected models are promoted, but usually custom made tools are needed. The set of tools define BY USING WHAT we execute the extraction process.

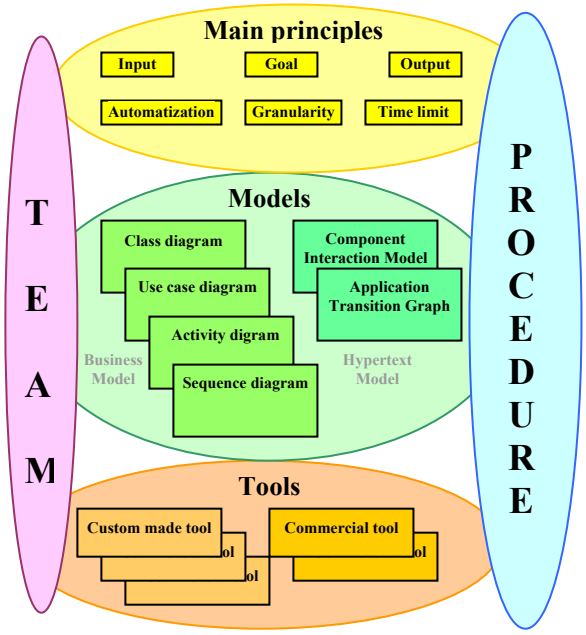


Figure 1. Framework for web application domain knowledge extraction

The extraction team (5) of professionals performs the extraction procedure. Involvement of the appropriate number of team members is key success factor, thus we often select ex-developers as team members. This element is strongly connected with the level of automatization. The definition of the team answers on the WHO question.

A. The main principles

The definition of main principles determines the nature of the extraction process. Several issues are addressed in this part. Some of them have strong dependences between them, while others can be compensated by some other. The major part of extraction approach is predefined by requirements and current state of web application, while others might be improved by careful decisions provided by the framework. For example, if a detailed extraction of domain knowledge for an application written in old technology is required using inexperienced team members, the process will be resource demanding and long. In this case framework might be used for clever selection of tools, which might upgrade the entire process.

The main principles addressed in the framework for domain knowledge extraction are the following:

The goal of extraction: the entire extraction process is driven by defined goal, usually describing the future use of extracted knowledge. The goal may outline the importance of extraction (i.e. demanding exact understanding or just vague insight), the expected aspect of extraction (i.e. demanding just description of application processing) or the amount of effort expected (i.e. demanding the extraction of data definition if it can be obtained by tools only).

The input: the common definition of process input. Extraction process might be limited to selected set of software artifacts (i.e. programming code in Java or all existing assets) or to a part of application (i.e. the artifacts describing enrollment management in e-Študent).

The output: common description of process results. The required description of domain knowledge may serve to future application development or not, it may be defined for certain (same, different, common) platform and/or technology.

The level of automatization: the extent of tool use in comparison with manual performance. This issue might require the extraction process to be fully automated using commercial or custom made tools, or just promote the use of existing tools. Usually a combination of tools and manual performance is needed.

The granularity level: this issue defines the level of detail demanded by extraction process. If the level is too low, we might spend too much time on non-important details, if it is too high we might skip some important information. Incorrect granularity level definition can cause serious consequences n extraction performance.

The time limitation: sometimes a tight time deadline for domain knowledge extraction is required. This must be logically considered in other issues definitions as well (i.e. the use of existing tools, the higher detail level).

In case of e-Študent, the main goal of extraction was to extract as much as possible domain knowledge for use during the development of the e-Študent's successor, which will use different technology. During this process all possible artifacts were used as input, and the output should be usable for automated during successor development. The use of tools was promoted, but the construction of the custom made tools was suggested to achieve the expected level of results quality. At first the deadline was tight, but later it was extended to achieve higher quality of collected results.

B. The models

By definition the models define a description of the application at a higher level of abstraction, but this can mean very different things to different people. A formal text description may meet those requirements as well, but usually a set of diverse models is used to cover different aspects of domain knowledge definition. If implemented properly the models reduce maintenance costs and prove beneficial for both managers and programmers [8].

The main principles definition limits the selection of suitable models, but some selection would be still possible. Nevertheless, our experience proved a predefined set of carefully selected models covers the requirements for all web applications, thus we define our framework strictly for this models.

For web applications it has been advised [9] that models should consist of

- a business model,
- a hypertext model, and
- a presentation model.

The business model describes the business processes, the hypertext model describes how web-pages are built and linked while the presentation model contains details of the graphic appearance of a web application. In our framework the presentation model is omitted as it is limited to user interface description only, while the other two are precisely addressed.

1) The business model

The model-driven development (MDD) promotes a definition of the software development through a hierarchy of defined models at different levels of abstraction [10, 11]. As such it is a natural choice for the formulation of the results obtained by the reverse engineering. These models are defined by the model-driven architecture (MDA) which implements the MDD. An important characteristic of MDA is promotion of the automatic generation of the lower level description models - the application code in the selected technology.

MDD is based on a notion of automatic transformations between different models describing an application on different levels. In the ideal situation, a developer would produce a platform independent model (PIM) [12], add some platform specifications to reach the platform specific model (PSM), and finally generate the application. Thus by selecting the platform independent model (PIM) as defined in MDA for the business model

we gain an important advantage for the following reconstruction of the application in the selected technology.

Only the following four kinds of UML diagrams are sufficient to produce the PIM [13]:

- class diagram,
- activity diagram,
- use case diagram, and
- sequence diagram.

Class diagrams represent the static structure of software systems and subsystems in a graphical way [14]. In the case of the non-OO business applications database objects are typically modeled in this way.

Class diagrams are the easiest to produce: the relational database tools are capable of producing entity relationship model (ERM) for the database and there exists a relatively simple transformation from the ERM to a class diagram. As the behavior information is omitted, we produce a connectional class diagram at first. To upgrade it to the proper OO class diagram, class methods have to be defined by the programming objects analysis.

A use case diagram represents system functionality by exhibiting the interactions between system users and the transactions that provide value to users. They display relationship between actors and use cases as well as dependencies that exist between users [14].

As the system is being reverse engineered it is assumed the use case diagrams produce a clear top-down picture of the problem domain. Usually use case diagram specifies a set of actions that can be performed by a certain actor. A list of actors must be retrieved from the system by inspecting the system from the user's point-of-view. Once the list of user roles is established, one use case diagram per each user role should be produced.

Activity diagrams are the most important artifacts in terms of the future reimplementation of the existing application as they denote the operational semantics (sequence, selection, iteration) of business processes [14].

It is possible to produce activity diagrams automatically from the existing code. However no tool seems to be capable generating adequate activity diagrams automatically for the applications as they face granularity problem or they include many bad system design elements. The activity diagrams are produced by the combination of the special custom made tool use and a careful manual analysis. During this all different types of code are analyzed.

Sequence diagrams express the interactions and the data flow between different objects within an application. [14]. These interactions are most often representations of procedure and function calls. While procedure and function calls can be extracted from the source code during parsing with a custom made tool, some interactions can only be obtained by dynamic analysis of the application. Sequence diagrams are needed to augment the activity diagrams.

2) The hypertext model

To adequately present the complex description of a dynamic web application the Atomic Section Model (ASM) [15] was chosen. Other formal descriptions (i.e. FSM [16] and WebML [17]) could be chosen, but the ASM has an advantage as it is a well known model capable of representing complex web applications with simple graphs. ASM is primarily developed for testing web applications [18].

Each HTML page is represented by a Component Interaction Model (CIM); all CIMs are combined together into an Application Transition Graph (ATG) representing ASM. Thus, the ATG of a web application is the formal representation of the hypertext model.

The CIM of a single page is defined by four elements:

- A set of start pages: all other HTML pages which the page is referenced from,
- A set of atomic sections: undividable basic blocks of code the HTML page is made of,
- component expression: regular expression denoting all possible sequences of diverse atomic sections when dynamically constructing HTML page, and
- a set of transitions: links pointing from and to (other) HTML pages.

The ATG is formally defined by:

- a set of defined CIMs,
- a set of all transitions of all CIMs,
- a set of variables defining all possible states of the presentation layer, and
- a set of all diverse starting pages (usually one).

Described selection of models was first used during the e-Študent extraction process. It proved to be reasonable, understandable and effective.

C. The extraction procedure

The extraction procedure precisely defines a sequence of the necessary steps to perform the domain knowledge extraction from the selected inputs to the required results.

While *the main principles* define inputs on a higher abstract level (i.e. all programming objects), a detailed list of inputs (all data and programming artifacts of existing web application) is required at this point. The database metadata (i.e. Data Catalog in case of Oracle DB) is an important input source as well. The extraction procedure defines a set of actions on input data to construct the final results describing the application on a higher abstract level. Finally, two groups of results are obtained: a large number of diagrams compliant to above mentioned models, and a set of lists of different web application objects. The models we presented in framework definition exactly define the form of expected domain knowledge description, while the lists enable an aggregated insight into the objects of existing application.

The core of extraction procedure consists of three groups of activities perform:

- the elimination of the dead-data and dead-code;
- the production of the business and hypertext models; and
- the consolidation of collected results.

The elimination of the dead-data and the dead-code is necessary as a number of redundant code and data is produced during development and maintenance. The code of alternative solutions or old test cases data may mislead the extraction process and it can result in incorrect results. The elimination is performed in two steps. First, we identify all live root objects (i.e. web application start pages). The second step is to recursively parse through all connected objects to identify all other live objects. All unreachable objects are omitted.

The production of the business and hypertext models is the core of the extraction procedure. It produces a number of diagrams for all selected models. It is performed model by model, partly by (commercial or custom made) tools and partly manually by expert performers. This part of procedure consists of the following steps:

a) The class model diagram is produced from database tables and stored procedure/functions. Usually we create first an ERD, which is later translated to proper class diagram.

b) The use case diagram requires a roles list definition first, to be able to define all functionality they can perform later. The menu structure or entry HTML page analysis may be used to define use-case diagram as well.

c) A number of activity diagrams is produced next. Each programming object's (dynamic page, stored procedure, trigger, etc.) code is first simplified by custom made tool, to be later translated to a separate activity diagram. Similar approach was used by [19].

d) The sequence diagrams are produced by static analysis for each programming object as well. The inter-objects communication is recursively examined first, and this information is later described in form of sequence diagram.

e) The CIMs are produced using all HTML pages of application (the dynamic pages). Their code is analyzed to identify all atomic sections and transitions, and those alre used to create a one diagram for each. Manual analysis is needed to define regular expressions.

f) All CIMs are combined together to create a final ATG of the web application. A large portion of this step mst be performed manually.

The consolidation of results is the last group of activities. In this step we save all collected information to a predefined repository, possibly execute additional cross-checks of consistency, and organize the future use of collected data. Presentation tools for collected data may play an important role as well.

The described procedure was used during e-Študent extraction process. As e-Študent is an Oracle Portal application written in PL-SQL, the input consisted of a number of specific programming objects (dynamic pages, stored procedures and functions, triggers, Java Script code, etc.), database objects (tables, views, sequences, etc.) and Oracle DB Data Catalog. The number of initially analyzed objects was high (over 1400 programming objects and 380 tables), but it was reduced by dead-code and dead-data elimination to final 1075 programming objects and 117 tables. Those numbers approximately correspond to the final number of separate diagrams we obtained at the end of extraction process [2].

D. The tools

A list of different tools can help to carry out separate fragments of the extraction of domain knowledge of a web application. They must be compliant with predefined main principles, and they should support selected models and execute tasks required in extraction procedure. Additionally, characteristics like effectiveness, portability or ease of use are usually desired.

The tools may be selected from a waste number of existing commercial tools or they may be custom made - thus specially created to solve a selected problem for a certain extraction process. The main advantages of commercial tools are their availability to use, standard principles and manufacturer support, but they may provide a solution just for a part of problem, they are usually expensive and require special preparation of inputs or adaptation of working environment. Custom made tools perform exactly as required. They can be adapted to selected problem and integrated into existing environment. However, building a custom made tool requires a lot of effort. Such tools are usually limited to specific problem, and they can be error prone due to limited testing.

Ideally, one heterogeneous tool to support all extracting requirements is desired. But in practice we need a set of different commercial and especially custom made tools. The ratio between the two groups depends on web technology used: for popular technologies (like Java or .NET) a large number of available commercial tools exist especially for UML modeling, while we must rely on custom made tools otherwise. Also, the hypertext model is usually not supported by commercial tools.

In case of e-Študent extraction we used just some commercial tools to perform certain parts of PIM extraction, as the PL-SQL programming language is not largely supported by those tools. The class diagram could be created automatically from existing database metadata (at least to some degree if not entirely) by tools like "UML Modeling for SQL" by Entrionics [20], "Altova UMODEL 2012" by Altova [21], and "SAP Sybase Power Designer" [22]. "UML Modeler for SQL" can be used to create a scratch of activity diagrams as well.

The major part of extraction procedure was performed by a number of custom made tools. The main principle of most of the custom made tools is parsing the metadata or programming code to select the elements we search for. The main issues addressed by the created custom made tools for e-Študent knowledge domain extraction include:

- the dead-data and the dead-code elimination;
- the class diagram methods extraction from stored procedures and functions;
- the use-case diagram creation from menu structure;
- the programming objects code simplification as a step towards activity diagram;
- the definition of all inter-object communications while defining sequence diagram; and
- the determination of all atomic blocks and transitions while defining CIM-s.

E. The extraction team

The extraction team is outlined as the last element of framework as it proves to be crucial for successful execution of the entire process. Even if preformed correctly the sole execution of formal procedure of extraction may not be enough. The expected final result usually demands a deeper understanding of web application.

The minimum requirement for each team member is to be knowledgeable with technology the web application was developed in, and the understanding of the framework models, procedure and existing tools the extraction uses. Additional operational knowledge of specific tools and technologies is often required for custom made tools development.

The extraction of domain knowledge for an average size web application is an important and especially resource demanding job. Thus, it is advised the team consists of more than one member. A suitable group of three members may include an expert analyst in UML modeling, an ex member of original development team, and an experienced programmer. Of course the optimal number of team members depends on the goal, application size, available time etc.

In e-Študent extraction process a group of ex developers was used. This decision proved to save a large amount of extraction execution time, even though the e-Študent had been developed approximately ten years ago in technology none of them uses any more for years.

III. CONLUSION

We learned several important lessons during the e-Študent domain knowledge extraction process. In fact, in paper presented systematic framework emerged from this practice. The framework's use is obligatory for correct extraction performance, but this is only the first necessary step. The process must be automated as much as possible by a suitable set of tools. The sole commercial tools do not perform well, and a number of custom made tools must be created by skilled developers with previous experience in same technology use.

Unfortunately, a large part of analysis has to be done by good old manual work. The best bet is to use a team of ex-developers as they probably already understand the application's essence to be extracted.

REFERENCES

[1] V. Mahnič, I. Rožanc and M. Poženel, "Using e-business technology in a student records information system" in Proceedings of the 7th WSEAS International Conference on E-Activities (E-Activities'08), pp. 589–594, Cairo, Egipt, 2008.

[2] I. Rožanc and B. Slivnik, "Producing the platform independent model of an existing web application" in Proceedings of the Federated Conference on Computer Science and Information Systems, pp. 1385-1392, Wroclaw, Poland, 2012.

[3] R.C. Martin, "Agile software development, principles, patterns, and practices". Prentice-Hall, Englewood Cliffs, NJ, USA , 2002.

[4] V. Mahnič, "A case study on agile on planing estimating and planing using Scrum". Electronics and Electrical Engineering, Volume 5, Number 111, pp. 123-128, 2011.

[5] V. Mahnič, S. Drnovšček, "Introducing agile methods in the development of university information systems" in Proceedings of the 12th International Conference on European University Information Systems, pp. 61-68, Tartu, Estonia, 2006.

[6] "Manifesto for Agile Software Development". http://agilemanifesto.org (last checked Februar 3rd, 2013)

[7] R. Akkiraju, T. Mitra and U. Thulasiram, "Reverse engineering platform independent models from busines software applications", in A.C.Telea (ed.) *Reverse Engineering - Recent Advances and Applications*, chap.4, pp. 83-94. In Tech Press, 2012.

[8] S. Rugaber and K. Stirewalt, "Model-driven reverse engineering". IEEE Software, Volume 21, Number 4, pp. 45-53, 2004.

[9] P.A. Muller, P. Studer, F. Fondement and J. Bezivin, "Platform Independent web application modelingand development with Netsilon". Software & System Modeling, Volume 4, Number 4, pp. 424- 442, 2005.

[10] S.J. Mellor, K. Scott, A. Uhl, D. Weise, "MDA distilled – Principles of model-driven architecture". Addison-Wesley, Boston, MA, USA, 2004.

[11] L. Favre, "Formalizing MDA-based reverse engineering process" in Proceedings of the 6th International Conference on SE Research, Management and Applications (SERA'08), pp. 153–160, Prague, Czech Republic, 2008.

[12] I. Luković, M.J. Veranda Pereira, N. Oliveira, D. da Cruz, and P.R. Henriques, "A DSL for PIM specifications: Design and grammar based implementation". Computer Science and Information Systems, Volume 8, Number 2, pp. 379-403, 2011.

[13] F. Solms and D. Lobser, "Generating MDA Platform independent model using URDAD". Journal of Knowledge.Based Systems, Volume 22, Number 3, pp. 175-185, 2009.

[14] R. Miles and K. Hamilton, "Learning UML 2.0". O'Reilly Media, Sebastopol, CA, USA, 2006.

[15] J. Offut and Y. Wu, "Modeling presentation layers of web applications for testing". Software & Systems Modeling, Volume 9, Number 2, pp. 257-280, 2010.

[16] A.A. Andrews, J. Offut and R.T. Alexander, "Testing web applications by modeling with FSMS". Software & Systems Modeling, Volume 4, Number 3, pp. 326-345, 2005.

[17] S. Ceri, F. Piero and A. Bongio, "Web Modeling Language (WebML): a modeling language for designing web sities". Computer Networks, Volume 33, Number 1-6, pp. 137-157, 2000.

[18] P. Amman, J. Offut, "Introduction to Software Testing". Cambridge University Press, New York, NY, USA, 2008.

[19] Y.Zou, J. Guo, K.C.. Foo and M. Hung, "Recovering bussines proces from business applicatons". Journal of Software Maintenance and Evolution: Research and Practice, Volume 21, Number 5, pp. 315-348, 2009.

[20] "Entrionics UML modeling for SQL". http://www.entrionics.com (retrived May 7th, 2012)

[21] "Altova UModel 2012". http://www.altova.com/umodel.html (retreived May 7th, 2012)

[22] "SAP Sysbase Power Designer". http://www.sybase.com/products/modelingdevelopment/powerdesigner (retrived May 7th, 2012)

The influence of social networks on student's evaluation and results

Aleksandar Skendzic*, Bozidar Kovacic **, Davorin Valencic ***

*University of Applied Science, Gospic, Croatia

** University of Rijeka/Department of Informatics, Radmile Matejcic 2, Rijeka, Croatia

*** University of Applied Science Velika Gorica, Velika Gorica, Croatia

**Phone: (385) 51-584-700 Fax: (385) 51- 584-749 *E-mail: askendzic@velegs-nikolatesla.hr, **bkovacic@inf.uniri.hr, ***davorin_valencic@yahoo.com

Abstract - The influence of social networks, Facebook especially, and on student population are often subject to various investigations. The analysis of the interdependence between the values that determine trends in the use of social networks in the student population is possible to determine the impact of social networks on students' evaluation and results. The starting assumption implies better performance of students who use social networks for occasional contacts with other users in relation to the students who have active status on social networks. The research deals with the validation and analysis of the interdependence of the assumptions set of relevant values. It is expected that the interdependence of time spent on the social network to publish the current status and results achieved (success) exams had a negative correlation coefficient. The results of the research will focus future research on the impact of social student population based on the calculation and analysis of the interdependence of the relevant parameters using social networks.

Keywords: social networks, impact, analysis, student's evaluation.

I. INTRODUCTION

Social networks are relatively "new" way of mass communication among their users. When we speak about this "new mode" of communication, we are obliged to consider it with some reserve. Social communication exists since ever and represents simple way of social communication with the scope of maintaining the existing circle of friends, and it's expanding. Expanding of the circle of friends results with the parallel expansion of social network of users/singles. According to Jacob Moreno, sociometric[1] procedure is one of modes (method) for exploring a social network. This procedure is used in sociometrcs that is one of the understandings in sociological logic psychology studies. The procedure is based on disclosure of the structure of relations (type of appeal, of authority, respect etc.) within a small social group (group, assembly, class, division…) and the results are represented graphically on sociograms – the graphic expression of social interactions among individuals. Nevertheless, Moreno defines the term "social communications" and based on that, it is possible to conclude that social networks can be integrated in real world and the social networks of today transfer the concepts and ideas on the widely expanded network – Internet. Internet has become the "ideal" platform for creating social networks, which in this case aren't limited by area (geographic), but have no obstacles. At the beginning, social networks were used mostly by younger individuals, while, nowadays, they are used even by elderly population. Creation of social networks on internet was allowed mostly by advanced services as MySpace, Facebook, Twitter and many others. Chapter I is going to give the introduction on the phenomena o social networks in society; Chapter II. describes social networks and communication, Chapters III. and IV. express the scope of the research, methodology and the results.

II. SOCIAL NETWORKS

Online social networks can be defined as a "service based on the web that allows individuals to (1) create a public or semi-public profile within a limited system, (2) to articulate the list of other users for sharing the connection and (3) to look and to use their own list of connections within the system."[2] Online social networks use in a way the methodology of "Open Access". They are available for all users and are not submitted to intellectual property rights and limitations. During the process of registration, the user is informed in advance with privacy rules and availability of personal data. The popularity of on line social networks is result of the fact that they are based on social interaction among the users, without being obliged to belong to any interest group like internet pages dedicated to interest communities (chat, news group, different forums and similar). Wellman[3] states that "the world is made of

[1] Moreno, Jacob Levy : The Sociometry Reader. Glencoe, Illinois: The Free Press, 1960.

[2] Boyd, D. M., Ellison, N. B. (2008.). Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication. 13 (2008.), 211.

[3] Wellman, B. (1988.). Structural analysis: From method and metaphor to theory and substance. U: Wellman,B., Berkowitz S.D. (ur.), Social Structures: A Network Approach. Cambridge, UK: Cambridge University Press. (p.37).

networks and not of interest groups". Communication among "users" of social networks is made of "introduction" of persons by own users' profiles. Users' profiles "disclose" relevant characteristics of a person (sex, age, hobby, interests, employment, activities) accompanied by multimedia material like photos, video clips or audio recordings. Identity is the key term on social networks. Identity is the mean to recognize the users of a network community. The user of a social network builds the own identity based on personal data, habits and other recognizable characteristics. It is known, that, before the popularization of social networks and the Internet as a media, people were identified by different "fazes" (clothing stiles, habits, activities) that changed in time.

Today, the most popular social network service is Facebook. Facebook has been designed as a support to other social networks at universities. The service started on 2004 as a service for on line social networks just on Harvard, and the user had to have harvard.edu e-mail address.[4] In time, Facebook started to include other universities, high schools, corporations, and in the end, all that are older than 13 years of age. Popularity of Facebook can be seen in numbers: more than 600 millions of active users per month,[5] 50% of active users log every day, average user has 130 friends, users spend on Facebook 700 millions of minutes monthly. The data that 200 million users use Facebook on mobile devices is interesting.

The statistical data for Republic Croatia are given on the next graphic (Picture 1):

Fig. 1. Croatia is at 70th place in the world on number of users of social networks in according with the size of the population, and it is notable the growth of users from +2,69% and penetration of 35,97%.[6] 365 diagram[7]

From the statistic point of view, it is possible to deduct, that Facebook has, surely, become widely expanded way of communication and from the sociological point of view it has become the new form of individual or group activity. New forms of communication, connections and the status of individuals are induced. Social networks have the phenomena of "Internet Addiction". As much this addiction relates to use of social networks, it's related to the use of internet as a media space, too. To show in more expressive way the "mass expansion" of social networks in all aspects of human society, the research conducted by Cisco Expo in 2010 in business sphere has to be noted. According to that research, 13% of companies were present in social network cannels.[8]

The term "social network" consists in a wide spectrum of users' social interactions. From acquaintances to introduction, from exchange of experiences to communication and distribution of media materials (photos, videos, speech and etc.). The results of the use show all the advantages and disadvantages of this way of communication. As disadvantages the culture of socializing, of speech, of behavior can be accented and protection of privacy is an always actual problem.

III. PREVIOUS RESEARCHES

Different researches on influence of Facebook and social networks on students have been published. The first research has been done by scientists from Lockhaven University in Pennsylvania[9]. They have published a study that examined the influence of use of Facebook on faculty grades[10]. The research was made on a sample of 1.800 students, and the conclusion was that there is a certain connection between the use of the biggest social network of the world and bad grades. Of all examined students 92 % use Facebook, and they spend averagely 106 minutes. The use of Facebook does not have any influence on students' exam grades, but students that use it for more than three hours per day have notably weaker results.

The research, also, shows that users that on social networks (e.g. Facebook) share only links or they just check out what their friends do, often have better faculty grades. On the other side, those that just publish states like: OMG, LOL and similar, in which they write what they are doing and where they are, in majority have worse results. In conclusion of this research it is stated that Facebook itself doesn't have bad influence on students, but it is important the way they use it. If it is used in the context of activities not connected or irrelevant for studying it can have bad influence on students' grades or results.

The second research has been published by the American Educational Research Association on Ohio State University. This research has showed that people using social networks have worse results on tests. In conclusion, the results of the research have shown that students that use Facebook relatively regularly (meaning daily) have much

4 Phillips, Sarah. "A brief history of Facebook". *The Guardian* (London). http://www.guardian.co.uk/technology/2007/jul/25/media.newmedia. March 7, 2008.

5 Web source: Goldman to clients: Facebook has 600 million users". *MSNBC*. January 5, 2011. http://www.msnbc.msn.com/id/40929239/ns/technology_and_science-tech_and_gadgets/. January 15, 2011.

6 Source: **Facebook statistics**, url: http://www.socialbakers.com/facebook-statistics/croatia (accesssed: 29.01.2013.)

7 Web source: http://onlinehelp.microsoft.com/en-us/office365-enterprises/ff652539.aspx (05.01.2012.).

8 Web source: http://www.cisco.com/en/US/services/ps2961/ps2664/Necessity_of_Pervasive_Collaboration.pdf (accessed: 19.12.2012.)

9 Url:http://www.academia.edu/1207833/Too_much_face_and_not_enough_books_The_relationship_between_multiple_indices_of_Facebook_use_and_academic_performance (accessed: 19.12.2012.)

10 Web source: http://thechoice.blogs.nytimes.com/2011/10/21/fbook-grades/

worse results on tests than students that don't have the habit of being on Facebook regularly. Students and other persons that use intensively social networks in the form of changing their statuses almost daily can have only 1 hour per week for intensive studying. There is also research done by The Ohio State University at Columbus which surveyed 219 undergraduate students to determine whether or not there was a link between Facebook usage and Grade Point Average (GPA)[11].

IV. SCOPES AND METHODOLOGY OF RESEARCH

Today, Facebook has more than 500 million users. The website is very popular amongst students in high school and college. Facebook allows students to socialize in multiple ways. But when this technologically enhanced social life is extended into the academic life of students, grades[12] can be severely affected.

To analyze and confront the results obtained by researches done by Reynol Junco[13] at the Pennsylvania University and the American Educational Research Association (AERA)[14], in this research similar method has been used, but on smaller sample of respondents. This research was supposed to show that even on a smaller number of respondents (100 respondents) there is a connection between students' results and the use of social networks, and it was supposed to confirm the results of previous researches. The research was done at University of Applied Sciences in Gospic at January 2013. A questionnaire has been prepared for the research sample. The scope of the research is to show correlation between results and the statement that the use of social networks has the influence on the success of students during their course of study. A questionnaire has been focused on the next topics:

- Type of social networks
- Time conducted on social networks during a day
- Time conducted on social networks during a week
- Frequency of updating social network's profile
- Number of users in social networks
- Time conducted on social networks for educational activities
- Kind of educational activities on social networks
- Average grade of study

It is expected that the correlation between the time spent on social network with the scope of publishing actual state and the results obtained (successes) on exams should have the negative coefficient of correlation.

VI. RESULTS OF RESEARCH

During evaluation of summed results, this is what was obtained (as shown in Table 1, Table 2 and Table 3):

TABLE I. TIME CONDUCTED ON SOCIAL NETWORKS DURING A DAY

Question: How much time do you spent on social network during a day?	
Answers	Number of students
Less then 1 hour	14
Answers	Number of students
1 hour	7
1-3 hours	65
More then 3 hours	14

TABLE II. TIME CONDUCTED ON SOCIAL NETWORKS DURING A WEEK

Question: How much time do you spent on social network during a week?	
Answers	Number of students
Less then 8 hour	86
8 hours	7
More then 8 hours	7

TABLE III. GRADE OF STUDY

Question: What is your study grade?	
Answers	Number of students
Satisfactory	4
Good	13
Very good	70
Excellent	13

During evaluation of summed results, this is what was also obtained (as shown in Figure 1, Figure 2, Figure 3, Figure 4, Figure 5 and Figure 6):

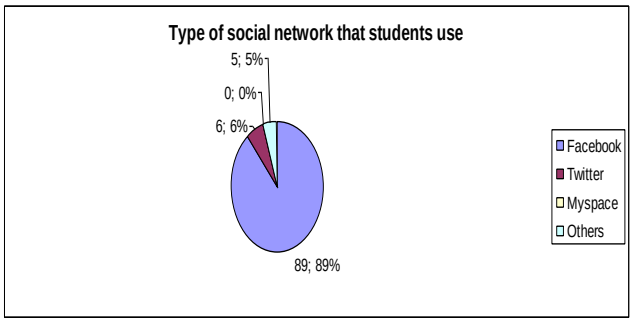


Fig. 1. Type of social network that students use

Fig 1. shows that the largest number of students use Facebook as social network (89%), 6% use Twitter and 5% use others social networks.

[11] Web source: http://yorktown-somers.patch.com/articles/student-perspective-the-academic-influence-of-the-social-network

[12] Web source: http://www.telegraph.co.uk/education/educationnews/8142721/Social-networking-teachers-blame-Facebook-and-Twitter-for-pupils-poor-grades.html

[13] Web source: http://www.academia.edu/1207833/Too_much_face_and_not_enough_books_The_relationship_between_multiple_indices_of_Facebook_use_and_academic_performance

[14] Web source: AERA, url: http://www.aera.net/EducationResearch/tabid/10065/Default.aspx

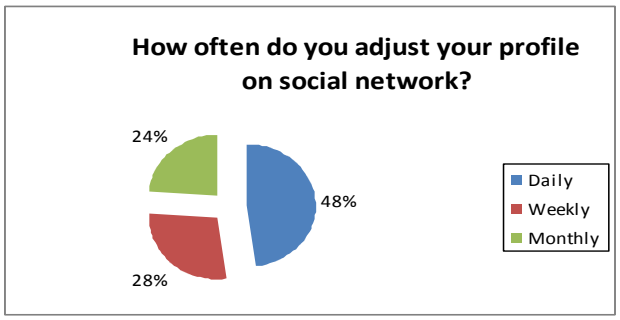


Fig. 2. Adjust social network profile

Fig 2. shows that the largest number of students adjust social network profile (48%) daily, 28% weekly and 24% monthly.

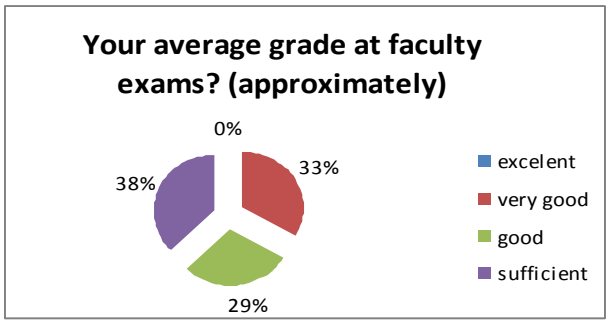


Fig. 3. Average students grades at faculty exams

Fig 3. shows that 38% of students have sufficient average grades at faculty exam, 33% very good, 29% good and nobody has excellent average grades.

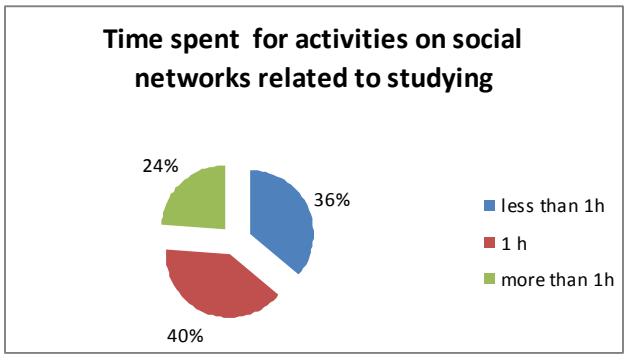


Fig. 4. Time spent for activities related to studying

Fig 4. shows that 40% of students spent about 1 hour on social network for activities related for studying, 36% less than 1 hour and 24% of students more than 1 hour.

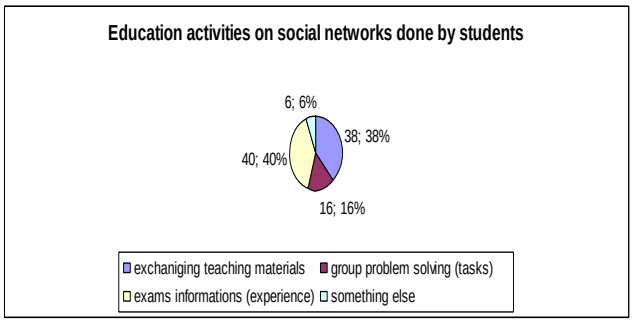


Fig. 5. Education activities on social networks done by students

Fig 5. shows that 40% of students use social networks for exams information activities, 38% use social networks for exchanging teaching materials, 16% use social networks for group problem solving and 6% do something else.

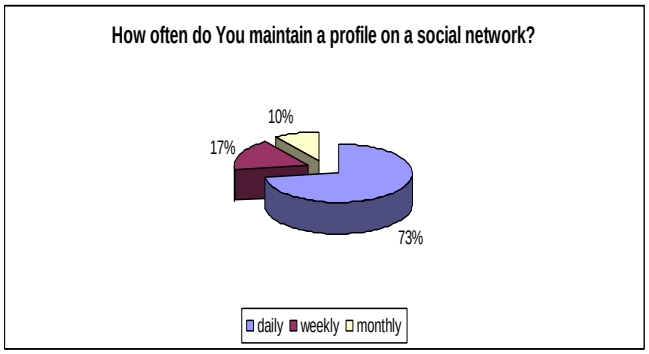


Fig. 6. Education activities on social networks done by students

Fig 6. shows that 73% of students maintain social network profile daily, 17% weekly and 10% monthly.

Correlation between daily time spent on social networks and the success of the study is -0.4328 which is a relatively weak correlation. She points to the relatively small effect of daily time spent in social networks and the success of the study, but the decrease of the trial. The correlation between the weekly time spent on social networks and the success of the study is -0.3125 which is a relatively weak correlation. Correlation defines an even smaller impact on the success of studies in relation to the daily time spent in social networks, but it should be noted the negative impact on the success of the study. We conclude that increasing daily or weekly time in social networks influence on declining performance of students. The correlation between the frequency of maintenance profiles in social networks and the success of the study is -0.4921 which is a relatively weak correlation. We conclude that more frequent maintenance profiles on social networks impact on the declining performance of students.

The correlation between the number of users with whom you regularly contacted and success in college is 0.5111 as medium correlations. She points to the positive impact of the number of contacts and performance studies, it can be concluded that some students used to exchange contact information about the study and to exchange information during the preparation of the exam. The correlation

between time spent in social networks for activities related to studying and success in college is -0.3577 which is a relatively weak correlation. She points to the negative impact of time spent in learning activities and performance studies. It can be concluded that working together to prepare for exams through social networks is common for students to study a high average, but is more common among students who have lower student success.

All calculated correlations observed variables and success studies are statistically significant at the number pattern 100 and a significance level of 5%. Influence of the number of contacts and time in social networking activities related to the study on the performance of laying courses has approximately the same correlation as for the success of the entire study and a statistically significant value. It can be concluded that social networks contribute to the success of a negative study, but their influence is relatively poorly correlated to the results of studies, that is not crucial to the success of students. Only the number of contacts in social networks had a positive impact on student achievement, which indicates the use of social networks to share information about the study, as well as developing your communication skills.

VI. Conclusion

When students are able to find a balance between socialization and education, Facebook can occasionally be a useful academic tool. Since the site allows students to stay in constant contact with each other, information about assignments, tests, homework, and due dates can be quickly spread amongst them. Through discussions on the site, students would occasionally collaborate on study guides for upcoming examinations, including midterms and finals.

Using correlation method it is expected that the correlation between the times spent on social network with the scope of publishing actual state and the results obtained (successes) on exams should have the negative coefficient of correlation. All results present relatively weak correlation. By this research done we conclude that increasing daily or weekly time in social networks influence on declining performance of students and calculation of correlation index demonstrated it.

References

[1] Wellman, B. (1988.). *Structural analysis: From method and metaphor to theory and substance. U: Wellman,B., Berkowitz S.D. (ur.), Social Structures: A Network Approach. Cambridge*, UK: Cambridge University Press. (p.p.19-61).

[2] Moreno, Jacob Levy : The Sociometry Reader. Glencoe, Illinois: The Free Press, 1960.

[3] Boyd, D. M., Ellison, N. B. (2008.). *Social Network Sites: Definition, History, and Scholarship.* Journal of Computer-Mediated Communication. 13 (2008.), 210–230.

[4] *http://thechoice.blogs.nytimes.com/2011/10/21/fbook-grades/* (accessed: 03.02.2013)

[5] *http://www.telegraph.co.uk/education/educationnews/8142721/Social-networking-teachers-blame-Facebook-and-Twitter-for-pupils-poor-grades.html* (accessed: 03.02.2013)

[6] *http://yorktown-somers.patch.com/articles/student-perspective-the-academic-influence-of-the-social-network* (accessed: 03.02.2013)

[7] Facebook statistics, *url: http://www.socialbakers.com/facebook-statistics/croatia* (accessed: 29.01.2013.)

[8] The Necessity of Pervasive Collaboration, url: *http://www.cisco.com/en/US/services/ps2961/ps2664/Necessity_of_Pervasive_Collaboration.pdf* (accessed: 19.12.2012.)

[9] American Educational Research Association (AERA), url:*http://www.aera.net/EducationResearch/tabid/10065/Default.aspx* (accessed: 18.12.2012.)

[10] Junco, Reynol, Department of Academic Development and Counseling, Lock Haven University, Lock Haven, United States,*http://www.academia.edu/1207833/Too_much_face_and_not_enough_books_The_relationship_between_multiple_indices_of_Facebook_use_and_academic_performance* (accessed: 02.02.2013.)

[11] Web source: http://onlinehelp.microsoft.com/en-us/office365-enterprises/ff652539.aspx (05.01.2012.).

[12] Phillips, Sarah. "A brief history of Facebook". *The Guardian*(London).http://www.guardian.co.uk/technology/2007/jul/25/media.newmedia. March 7, 2008. (accessed: 05.01.2013.).

[13] Web source: Goldman to clients: Facebook has 600 million users". *MSNBC*. January 5, 2011. http://www.msnbc.msn.com/id/40929239/ns/technology_and_science-tech_and_gadgets/. January 15, 2011. (accessed: 05.01.2013.).

Will universities and university teachers become extinct in our bright online future ?

J. Henno*, H. Jaakkola**, J. Mäkelä*** and B. Brumen****
* Tallinn University of Technology, Estonia
** Tampere University of Technology, Pori
*** University of Lapland, Rovaniemi, Finland
**** University of Maribor, Slovenia
jaak@cc.ttu.ee

Abstract – Members of highly developed societies do not grow their own potatoes, build their own house, marry in local neighborhood and (mostly) educate their offspring. Human society has developed a complicated networks of services to satisfy our needs in food and goods, lodging and education.

These services are based on Information. Trade is moving goods from where they can be obtained cheaper to where they can be sold profitably and the concept used for evaluation - money - is just information about ability to own/possess things; education is collecting, classifying knowledge obtained by all members of society and then arranging this knowledge to nice mouthfuls for transferring to our students.

The rapidly developing Information and Communication Technology (ITC, IT) has changed almost every aspect of our daily lives. Consumer goods are moving into on-line shops, brick and mortar shops are closed and replaced with on-line Amazons, where smart software helps you quickly find what you are looking for and gives you lot of information, what you cannot obtain in ordinary shop; salespersons are evolving into marketing and public relations professionals, networking with clients, customers, prospects, suppliers.

In education, especially in IT education, changes are even more rapid. Old-fashioned Colleges and Universities with large campus, staff and high costs are getting more and more competition from numerous educational entrepreneurs, who start massive open online courses (MOOC) with millions of users/students from all around the world and establish new on-line universities where costs of obtaining a degree are marginal compared to 'brick-and-mortal' universities and which again are open to nearly everyone independent of user geographical location.

Do we still have salespersons and teachers in 2020 - or will both professions become extinct ?

I. Introduction

Human society has created many structures to satisfy all our needs, described by Maslow's hierarchy of human needs [1].

Satisfaction of the basic, existential needs - food, clothing and lodging is measured in developed industrial societies in terms of consumption of goods. These goods are distributed by complex structure of trade and delivered to consumers by salespersons.

Satisfaction of higher level needs - social needs, the esteem of other people, the need for self-realization depends and are measured by education level of a person. In developed industrial societies education is organized by complex structure of schools and delivered by teachers.

Both these complex structures of trade and education depend on information. The achievements in Information Technology have tremendous influence and are rapidly changing both of them. Technological progress has taken over and reduced millions of man-hours in many areas - in factories, assembly lines, social organizations, fields and farms. This has rendered obsolete many established professions - typists, elevator, telegraph and switchboard operators etc etc. The wanted ads in www-sites and newspapers are looking for IT specialists - webmasters, Computer Systems Analysts, Database Administrators - these are among the best jobs listed in *U.S. News* 100 Best Jobs of 2013 [2]. But even some IT occupations are under danger with IT infrastructure of managing email and other applications is moving to the cloud.

We are producing information (and as a small part of part of it) useful knowledge in amounts which is more and more difficult to store with traditional methods and traditional media. This information is prevailingly digital and the only media large enough to store it is Internet. At first Internet made obsolete storing music on fixed-size disks - young musicians are presenting their opuses more and more directly on Internet and most of new computers already do not have CD/DVD drive - these drives are also becoming extinct [3]. Paper-based newspapers are dying out, e.g. demand for newspaper paper decreased in last year in Finland ca 10% [4]. Printed on paper scientific journals are supplemented or replaced with web-based ones. The 'Kindle Direct Publishing' system allows authors to bypass traditional publishers, publish their books digitally and put up in Amazon for sale "in hours" [5]. Many values are nowadays considered purely as (digital) information – money (more and more banks are refusing to deal with "real", physical money), many kinds of art – music, literature - Amazon is selling more e-books than paper books [6], live-streaming gameplay is aiming to kill games stored on physical media [7]. Today a new jaw bone for a patient can be created with 3D printing [8] – tomorrow 3D printers can create a whole outfit and footwear, sculptor can create a new masterpiece just typing into keyboard and all these can be delivered to customer by e-mail. NASA intends to use 3D printer to create spacecraft parts during deep space missions, the US

Army has experimented with a truck-mounted 3D printer capable of outputting spare tank and other vehicle components in the battlefield and a start-up company intends mining asteroids to 3D-print space stations [9]. Within the next decade, people will begin to print products in their own homes. They won't buy them, they won't have them shipped to their houses; they will print them. No need for shopping and most possibly we also do not have time for shopping, if our life continues accelerating as now.

And estimates obtained by comparing the current Information Revolution with the previous revolutions show, that "we are not at the end of the information revolution. We are only partway into it" [10]. Are we transforming into new species – informawore [11]? Ray Kurzweil predicted (less than ten years back) that we are approaching an era where Information Technology will be an essential part of human intelligence: "There will be no distinction, post-Singularity, between human and machine" [12] We are already there - most of us (university teachers) can not any more deliver a lecture without a computer, it has become an essential part of us.

II. Sales and Salerpersons

More and more people are making their purchasing decisions without going to a shop. They research online and make a buying decision based on received information. Even ways of obtaining very personalized items such as clothing and shoes which should fit the buyer are rapidly changing. Development of methods of 3D graphics for human body modeling and simulation together with development of 3D image capture technologies using 3D scanners has caused rapid development of digital tailoring - a complex of methods where made-to-measure clothing is produced starting with 3D scanning of a customer, extraction of essential measurements from obtained data cloud and then automatic production of a garment corresponding to exact measures of the customer [13]. If 3D scanning technologies are yet not available, people do still come into stores, but only to find the right size and fit before placing the sales order in some on-line shop. Everyone wants its goods for as cheap as possible, and people know that lowest prize could be find online.

"Buying with clicking" is also very convenient - just some clicks, eliminating the crowding and the hassle of dealing with large retail establishments and items will be delivered to your doorstep.

On-line information has created freely-flowing prizes - sellers determine the prize depending on the sales data and competitor's prizes, sometimes hourly [14]. Airlines and hotels have done this for years; they are following banks, where prizes are determined by computers (algorithmic trading) and change in a fraction of a second.

Salespersons who earlier contacted buyers face to face ("direct sales") are rapidly transforming into sales engineers [15] who sell complex scientific and technological products or services to managers of e-shop websites and researchers of online-communities trying to discover potential customers and understand their needs.

This has initiated hot debate "Is e-commerce killing brick and mortar ?" [16]. While in 2011 it was proposed that "In the battle of clicks versus bricks, retail must transform or die" [17], in the next year it was already stated that "Retail in 2021: When clicks have buried bricks" (CNET, [18]). In some fields e-shops have replaced nearly all old-fashioned physical shops. If you want in Tallinn to buy a computer, computer peripherals or home electronics, then the most convenient way is to search the site "PrizeWatcher" [19], which allows to search for an item and presents prizes in nearly all shops in Tallinn dealing with this kind of goods. But very few of them (still) have also some kind of salesroom (very reduced), in most shops you have to put an e-order. Heavy competition from online will allow only the strongest 'brick and mortar' businesses to survive - but exactly the same may hold also for universities.

III. On-line goods and on-line Knowledge

Once upon a time there was a village shop, where you could get the necessities - matches, salt, cloth etc. There was also a village school, where schoolmaster taught reading, arithmetic etc.

If you wanted something special (a new book, toys for children etc), you had to go to town. If you wanted your offspring to learn more, you sent them to school in town.

With time these targets - goods and knowledge become locally more and more available and people started to seek for 'real top' targets farther and farther away. In the beginning of the 19th century an Estonian man walked more than 2600 km from his home village to Düsseldorf art academy to study art [20]. At the beginning of 21th century Estonian entrepreneurs went to Taiwan to get cheap electronics to sell later in Tallinn.

But now-a-days it is not needed to travel anywhere - you can study or order goods with some clicks from your home computer. We are moving towards open global world, where all goods and knowledge flows freely and is accessible everywhere. Trade is based on information - buy an item where it is cheaper and sell where there (still) is need. With free flow of information gradually all goods will be available in all places - trade is based on law of connected vessels. What happens when everyone can seek the prize for every item around the whole world (and the postage expenses continue to go down) - will this be the end of trade? Will the trade business be reduced to post and the money - to e-yuans, which nobody can touch, only see on screen?

Knowledge and education behave the same way.

Every school-leaver in EU can send his/her application to every University or College in EU or USA and order whatever electronic toys from Amazon.

In USA, on-line enrollments growth rate - twenty-one percent - far exceeds the less than two percent growth of the overall higher education student population [21].

What happens when Google can answer every question for which the answer is already discovered - will this be the end of all schools, colleges and universities?

IV. Need for Education

The human resource – well-educated population is the most important recourse of every nation. Studies carried out in various countries have shown that there is a positive and statistically significant correlation between education of the population and growth rates of per capita GDP [22]. Demand for workers with higher levels of education, technical knowledge and skills is rising rapidly. It is estimated that by 2020 percentage of jobs that require a two-year or four-year degree and higher or special postsecondary occupation certificates or apprenticeships will be about 75% [23].

The economy depends more and more just on high-level fields. Despite the financial and economic crisis, high-technology manufacturing production increased in the period 2005 – 2012 in EU by 26 %, but medium-low-technology and low-technology productions shrunk during the same period by 5 % and 6 % correspondingly [24].

However, the growing economical pressures have greated very messy picture of our (higher) education.

V. Students

The current population of students, especially in IT specialties, is different and heterogeneous.

In Estonia, more than one third of students are older than 26 years. During the last five years many people disrupted their studies, but are determined to continue and often do so, but already in older age.

Most of IT students in Estonia and in Finland are already working in some field of IT - more than 75% already on bachelor level [25]. The IT field is very viable. General unemployment in Estonia in 2012 was 10% [26], but in IT field is a deficiency of specialists - the number of announcements about open jobs in IT is more than half of all open job announcements. The European Information Technology Observatory predicts 7-11% growth of IT sector in Estonia [27]. According to vision of the Estonian Council of the Information Society the number of IT specialists in Estonia should double by 2020 [28], [29].

To cope with constantly increasing need of IT professionals in Estonia, in addition to private The Estonian Information Technology College was in 2012 established IT Academy with two BCs and two international master study programs; in 2015/16 the Academy is supposed to have already 12 study programs. The Academy is funded both by the state and by private Estonian enterprises, e.g. Skype [30].

There is a clear distinction from the life template what was common for students ca 20 years ago - first you study to obtain knowledge needed during your working life, graduate and then go to work where you will use this knowledge. Currently many students in Estonia and in Finland start working even before they enter an university (if at all): they create web pages and sites, program games, tutor other students with difficult topics, e.g. mathematics, chemistry, even synthesizer programming - if you want to create your own hits. The websites with announcements about available tutoring contains tens of proposals ([31],[32] - in Estonian) from students.

In "good old times" people used to keep their jobs for all, or most, of their working life. Currently we change jobs 10-15 times during our life [33] and the frequency is increasing - young people (20..30 years old) on average change jobs every 18 months. In a poll conducted by the first author among IT bachelor students in Tallinn University of Technology nearly 70% indicated, that they already have had more than one job. These 'job-hoppers' are often considered good employees [34], who are trying out a variety of roles and workplaces while learning new skills along the way. Every new job requires learning new skills, thus life-long learning is tightly intertwined with life-long job-hopping. The divide between 'learning years' and 'working years' is vanishing. Albert Einstein: "Learning is not a product of schooling but the lifelong attempt to acquire it".

The role of university is for working students complementary and often less important than what's happening in their workplace.

The IT industry is very vibrant and innovative, exploring new things and constantly coming up with new ideas and applications. Five years ago, when the study programs for current graduates were prepared (and in many universities become fixed for several years) nobody know current IT buzzwords: Jira, Git, Toggl, NetBeans/PHPStorm, Dart, MySQL, Android, iframe, canvas, virtualization, cloud computing ....

But in (private) IT enterprises many of these words are work terms, otherwise they could not succeed in world-wide competition. Thus the professional knowledge and practical skills of working students often exceed in some areas knowledge and skills of their professors. Several years before programming of mobile devices was introduced in study programs of Estonian Universities, Estonian IT company developed mobile parking system which has been successfully introduced in several cities in Belgium, Ukraine, USA, Macedonia and is now introduced in the whole Moscow [35].

Sometimes it has been argued that working is a (very) negative phenomena - it reduces the time what students use for their studies. However, the research "Phenomenon of student working in Estonian higher education" [36] , carried out in the department of Economics of the Tartu University, Estonia, did not show very big negative influences of working. Students themselves, e.g. the foreman of the Union of Estonian Students considered working to be a positive factor: working students plan their time, understand better work environment and are much better prepared for working life when they graduate.

Working students are motivated, self-contained, know what they want and are actively seeking information/knowledge which is essential for their work or related to their interests - the two are usually closely connected. And they want it just now, just-in-time, not just-for-case [37], what is (still) the main principle for creating study programs in universities. Because of their work duties they travel a lot - below is the map of accesses by course participants of the homepage of the course "Computer Games and Virtual Environments" presented by the first author in the Tallinn University of Technology. During period when the course was

delivered, Sept 3, 2012 to Jan 26, 2013 there were altogether 1072 visits from 20 countries.

Visits of homepage of the course "Game Programming and Virtual Environments", offered by the first author in Tallinn University of Technology in the autumn term 2012

Ability to learn and communicate is very important skill in every profession; in IT this is critical - the main issue of IT is communication. Many teachers do not understand this new brave generation of 'digital natives' [38]. For digital natives “It’s only technology if it happened after you were born” [39] - and it is the technology what they know, and remarkably better than their peers.

All the technical know-how can be obtained when needed from WWW, from manuals, co-workers etc, but these sources are useless without ability to communicate and learn. Unfortunately, just those skills are degrading in 'digital natives', who are fluent with technology, but not so with ordinary human relations. Instead face-to-face talk they prefer to talking texting or blogging - even when the conversation partner is in the same room. Instead of writing a report they prefer to record a video clip, put it up to YouTube and send a link to this clip (this has happened with the first author several times).

VI.IT

Ten years ago "software engineer", "IT specialist" were a sophisticated names for a programmer. Now programming is essential for getting any kind of technical/mathematical job. And traditionally non-technological professions are also increasingly embracing some kind of programming as a way to advance their profession. Programming is becoming a much more fundamental piece of knowledge than reading or writing and IT specialists are getting the best salaries: "In Silicon Valley, annual tech salaries topped six figures" [40].

Information has very different properties compared to e.g. matter and energy. Information does not have conservation laws. Quite the contrary, information is like a fire or nuclear reaction – it feeds itself. If you have more information, you can do more; doing more you can obtain new information, which enables you doing even more and thus obtain even more information. But you have to use it quickly, value of information diminishes - yesterday's news are no news.

Thus there is an exponentially growing need for more and more talented IT professionals, who can manage, handle the information irruption.

VII. Global Growth of Information

A. *How is knowledge different from consumer goods?*

The most important difference of knowledge from consumer goods is the speed of creating new. New consumer goods are derivatives of new scientific discoveries. New consumer products, new medicine, electronic components etc are all based on scientific discoveries - i.e. new knowledge, new information. And the world’s information is doubling every two years:

"Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone." [41], and the total amount of global data is expected to grow to 2.7 zettabytes in 2012. This is 48% up from 2011 [42].

All this data - the foam together with useful information - flows to Internet. Sharing is built-in in our 'selfish genes' [43] - Internet was created to share information and this is the cheapest way to do so. It is more and more difficult to follow this flood - Google receives every minute over 2000000 searches [44].

This mass of publications, videos etc contains lot of foam - Facebook's 'See all what I'm doing now!', Twitter tweets, PhD dissertations and 'academic' publications created (in increasing rate [45]) with the 'copy-paste' method. It is difficult to estimate the global relation of bla-bla to useful knowledge, the tempo and mode of human knowledge expansion is an still rather poorly understood topic. According to International Data Corporation (IDC) report [46] only 3% of the potentially useful data in 2012 is tagged, and even less is analyzed But studies concerned with some specific field [47], [48] suggest, that the useful knowledge is also growing exponentially.

The useful knowledge is what teachers should find, systematize and teach. This makes teachers needful also in the future, but how can we cope with exponential increase of it ?

VIII. What are important IT skills?

Ten years ago, the most important IT skill was considered programming, more generally – Software Engineering (SE). But in last ten years IT has come a long way. Many-many programs have been created (software application market is currently worth around 2.2 billion dollars, and growing at a rate of 85% annually [49]).

All the time appear new software development tools which improve the productivity of human developers. In spite of rather short history of IT, especially PC IT - the 30th birthday of IBM PC was celebrated on Aug. 10, 2011 - there have been already 19 generations of computer programmers [50]. Thus today success of any project in the IT world depends on ability to work with multiple technologies, integrate everything together, communicate and learn much more than on programming skills. IT technology develops all the time rapidly. IT is constant learning, the most important is innovative approach to some (maybe old) problems - Microsoft, Apple, Google and Facebook were started by college students.

IX. New Economy

Information has one extraordinary property – it becomes worthless, if not used. As traditional jobs disappear, people will need to develop their non-automatable skills to be productive and succeed in the Hyper-Human Economy [51]. In the new based on human intelligence economy workers won't and can't go after existing jobs (they are already overcrowded), but should and would create new jobs and establish new enterprises. They identify problems to be solved and invent new solutions using their (hyper)human skills - discovery, creativity, ability to compare and generalize [52]. When Nokia announced in spring 2011 that 4000 software developers in Finland will be sacked, nobody of them complained. Many announced right away that next they will establish their own enterprise (and often mentioned game creation as their new field of operation).

The new jobs are based on new information, but this is in Europe not always used the best way. In EU are registered every year ca 65% more patent applications than in USA [53],[54], but Europeans have a great lag in creating new innovative products, services, enterprises. Americans have less patents, but use better than Europeans everything new, are more creative (and the start-up money is easier to get).

Under economical pressure, universities are more and more acting like international business enterprises, inviting foreign professors and accepting foreign students. The Tallinn University of Technology has currently biggest percentage of foreign students compared to other Estonian Universities and is preparing to accept much more students from East – from Russia, Ukraine, from former SU countries in Central Asia, from China; it has already a homepage in Chinese [55] and is advertising in Chinese Youtube (called Youku) and in Chinese Facebook - Renren.

To help students to get a grip of the USA entrepreneurship, the Tallinn University of Technology opened office in Silicon Valley and organized in summer 2012 in Tallinn the European Innovation Academy 2012, targeted to all Engineering, Science and Business students of Europe as well as professionals who are intending to pursue a career in innovative management and entrepreneurship. In this event gave talks among many distinguished lecturers Prof. Dr. Mark Harris, former head of INTEL Higher Education & Research Programs, Sir Harold Walter Kroto, Nobel Prize winner etc. The event was a great success thus in January 2013 was organized European Innovation Academy Winter Session which brought together 100 students from 30 different countries, where lectured innovation experts from world-famous and successful companies MICROSOFT and SAMSUNG [56]. These companies are supporting also the next annual European Innovation Academy, which will be in Tallinn July 8 – 26, 2013.

X. How to Cope ?

A. Curriculum

The classical idea of improving education is through improvement of curricula. With current temps of IT innovations this has become quite difficult. In order to teach something teacher himself should know the subject very well. Only few and very talented teachers can be on forefront of IT technology, if half of the content becomes obsolete in three years [57]. While computing has evolved very significantly in the past 10 years, Computer Science curricula have been much slower to adapt to these changes [58]. Curriculum is an complex emerging system [59] and it evolves first of all from bottom up, by enthusiasm of innovative teachers and students, who often introduce very different, novel and non-standard courses [60],[61] – not from top-down, as many reformers of education want it to behave. Directives "you take this topic into your course" most often result in excuses [62]"Sorry, I don't know anything about this stuff, I can't teach it".

The 'official' curriculums are rather slow in adapting new. The Computing Curricula developed jointly by the major professional societies in computing - the ACM and the IEEE Computer Society - is reworked after roughly 10 years; the current CS2013 Ironman v1.0 draft is scheduled for release in February 2013 [63]. This is the first time when this curriculum includes topics "Security" and "Parallel and Distributed Computing" - but both topics have been very important already for ten years. Feedback from practitioners to the CS2008 review indicated the importance of these two areas already in 2008, but CS2008 steering committee deferred development of those areas and they were included to the curricula only now. In the Tallinn University of Technology a Computer Security and Cyber Defense study program was introduced in 2009 and this program is popular among foreign students, since the amount of data that requires protection is growing faster than the whole digital universe and methods of protection are not keeping pace. The e-government, e-services, digital identity, digital signature etc are taught in Estonian schools already in grades 8-10 [64]. The Tallinn University of Technology is currently preparing a study program on e-Government; the program as also the study program on Cyber Defense is directed to students from China and India [65], where there are lot of students, but no possibilities to study these modern topics.

With current speed of new knowledge and technology innovations, curriculum cannot be any more fixed for years. Faculties and teachers themselves should revise their fields of teaching constantly and try to include new elements. Including new requires from teachers learning these new ideas and methods - additional work. But teachers are learning - 9% of users of the MIT Open Courseware courses are teachers [66]. And even more difficult than learning new tricks is excluding some old topics which teacher has taught for years and knows well. But this is inevitable; the first author has created for his course "Game Programming and Virtual Environments" in previous years over 100 www-tutorials and more than 1000 images, but for the last offering of this course (the autumn tem 2012) only ca fifth of them were re-used - most of the needed material were created new - the technologies, practices, possibilities have essentially improved.

B. Use of Internet

The on-line virtual world has become a great resource of teaching materials.

The Massachusetts Institute of Technology (MIT), the top of the latest Leiden Ranking of universities around the word [67] has already for over 10 years given free access to educational materials for all of its courses on the Internet. The MIT Open CourseWare (OCW) [68] is sharing materials from more than 2000 courses with an estimated 100 million individuals worldwide. Currently (winter 2013) the "Electrical Engineering and Computer Science" department offers 96 undergraduate and 143 graduate courses and these resources are extensively used, e.g. all lectures of the MIT course 'Introduction to Computer Science and Programming' together have on YouTube over 2 million hits [69]; thousands of them come from mobile devices and nearly all have rated the course 'thumbs up'.

Professors from the Stanford University Engineering School are offering on-line (for free) three of the school's most popular computer science courses [70]: *Machine Learning*, *Introduction to Artificial Intelligence* and *Introduction to Databases*. They are using technologies designed to enhance learning for Stanford students and extending them to a broad online audience. In the first four weeks after opening 300,000 students registered for these courses [71]. Lectures are delivered as short, interactive video clips which allow students to progress at their own pace through course materials. Together with courses are provided also live quizzes with instant feedback. Here are tested new technologies that allow students to rank questions that should be posed to the instructors.

The site 'Open Culture' [72], created in 2006 by group of enthusiasts leaded by Dan Colman, the Director and Associate Dean of Stanford's Continuing Studies Program brings together lot of free high-quality cultural and educational media for the worldwide lifelong learning community - audio books, online courses, movies, language lessons, eBooks. Currently (Jan 2013) they list over 700 courses (five times more than a year ago), among them nearly 100 courses on *Computer Science & Artificial Intelligence* which are both on very actual topic (*iPhone Application Development*, *Quantum Computing for the Determined*, *The Beauty and Joy of Computing* etc) and presented using modern technologies (e.g. as iTunes video clips).

On-line courses are seen as a danger to many traditional universities - they may reduce the number of new applications and thus also earnings of universities. The ratings agency Moody's report, released on Jan 16th, 2013 show that nearly half of colleges and universities expected enrollment to decline for full-time students, and a third of the schools expected tuition revenue to decline or to grow at less than the rate of inflation, this caused giving the whole sector a negative outlook [73].

Therefore many universities are not keen to give students credits for taking online courses and students typically can't use these courses as a way to satisfy requirements for a degree. But best universities (Stanford [74], US Berkeley [75], MIT [76], Yale [77] etc) have already adopted the on-line teaching model and under the pressure of high demand for these courses the attitudes are changing. In Feb 6th, 2013 the American Council on Education announced that five courses offered by Coursera [78], a Silicon Valley provider of online education stated by professors from the Stanford University, are similar enough to traditional college courses to be eligible for credit. Students who want to take the free classes for credit would have to pay a fee to take an identity-verified, proctored exam, and for a transcript to submit to the college of their choice. In USA similar transcripts are accepted by 2,000 colleges and universities for courses offered by the military or by employers. The Coursera courses are the first massive open online courses (MOOCs) which have been certified, but the number will certainly grow rapidly. Currently (Feb 2013) there are already more than 200 massive open online courses (MOOC) that also offer some kind of credits [79].

C. Use of video

Humans are expensive resource , especially lecturing professors. And the habits/abilities of 'digital natives' do not allow them to sit quietly for a whole lecture or read a long passages of text from a textbook - they prefer to absorb information in small lumps and preferably seamlessly with voice and vision.

The closest to live, face-to-face from all canned media modes are video clips. Compared to face-to-face, i.e. classroom lecture, all other modes of presenting information are far less effective and require much more mental effort, thus students still prefer the 'live' classoom with 'live' teacher, but conditions for 'face-to-face' are decreasing - students mobility, cheapness of on-line lectures etc.

Therefore he main media for delivering content in on-line courses are short video clips. Use of video clips for delivery of educational content has been strongly advocated already e.g. by R. Schank in his *Engines for Education* [80]. There are many positive examples, e.g. success of the Khan Academy [81], which explains math, science and humanities topics using over 3000 short videos (also in formats suitable for iPhone, iPad etc), which are augmented with automatically graded exercises. This simple approach was so compelling that by now more than 700 million videos have been watched by millions of viewers. In spite of some skepticism [82] there is strong evidence (e.g. [83]) for positive effect of video clips in IT courses.

Unfortunately video technology has often been misused. Long recorded lectures are much more difficult to follow than corresponding material in text. With text it is easy to scroll to the interesting passage; with video to find just a short subsection is nearly impossible without watching the whole lecture. Video should be used as short clips which are dealing with one idea only, they should be technically good (in many educational clips are shown texts which are not readable) and the presenter/lecturer should have some actor's talent – monotone reading of written notes does not convey much message.

D. Blended teaching

With now-a-days very mobile population of students the only way to reach them all is blended teaching - class-room face-to-face together with use of on-line materials. On-line materials should cover 100% of all what is

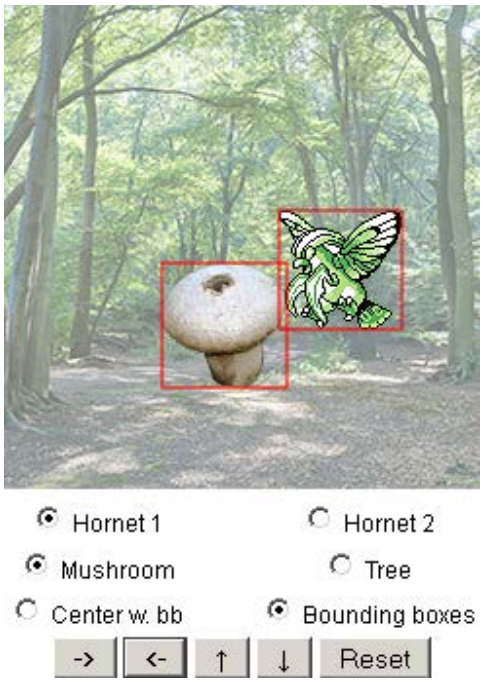
presented face-to-face in classroom. Fortunately, lot of what is needed can be already found from Internet.

There are many free resources which help to learn programming: Scratch [84] allows create simple interactive stories, games, music and video playfully experimenting, without dealing with syntax, and thus is suitable even for preschoolers; Alice [85] – innovative 3D educational programming environment allows to create an animated story, an interactive game, or a video to share on the web; Yenka [86] allows to simulate scientific experiments, create mathematical and statistical models, design electronic circuits and learn computer programming using visual control of animated characters by flowchart programs. Founded by Columbia University students Codecademy [87] offers free interactive tutorials that guide users as they write and test lines of JavaScript code directly in their browser windows. For teaching 3D modeling is available the http://www.3dvinci.net [88], which announces: "If you're a teacher, parent, or design professional looking for a steady stream of project ideas, our Project-of-the-Month subscription is ideal for you! For one year, you will receive each month three FUN, 'bite-sized' SketchUp projects, covering a variety of subjects. Projects are ready-made, step-by-step lessons; you can bring them right to your students; one project each month will be math-related, perfect for geometry lessons. The other two projects will focus on 3D design, geography, science, art, etc. ". There are 50 projects available, what should be quite enough for one-term course.

Usually it is not possible (yet) to get from Internet everything, something has to be created just for this course and because of rapid development, ca 30% has to be remade every year (some education apparatchiks think, that once an e-course has been created, the teacher is not needed any more – a total utopia!). Creating material just for your course is inevitable – it is easy to find programming tutorials or explanations of multimedia (e.g. sound, video) technology, but difficult to find suitable material which explains higher matters – connections with other topics, explains trends of development. And every year appear new technologies, e.g. (in the signal processing and multimedia) very important for applications compressed sensing, which allows hyperspectral unmixing - finding the fractions of materials in pixels of a hyperspectral images or automatic tracking of cells in phase contrast microscopic videos [89].

On-line should be also study plans (especially plans for practical exercises) – if student already knows the technology/program intended for an exercise, he/she can pass it.

On-line materials should attract students - be visually pleasant, well formatted, contain lot of images and utilize as much as possible the best feature of digital media - interactivity. New technologies, e.g. html5, are making it easier to produce interactive www-pages, which allow to demonstrate dynamics of algorithms and processes (see e.g. [90], [91]). Algorithms and processes have always many different possibilities; the following image is a part of a web-page explaining different methods of calculating collisions in games. User can select several objects with different size and shape, they can also select different methods of collision detection - bounding box with bounding box (demonstrated on picture), bounding box with center, object pixels with bounding box or center. The 'Reset' button places selected objects in random positions and now user can move the second object clicking on arrow buttons and follow, when the collision is detected. On the side of this canvas animation is displayed code, which drives this animation.



E. Pedagogy

"There's a difference between curriculum and pedagogy. Curriculum is all about what we teach. Pedagogy is about how we teach it." [92]

Even more important than IT technical skills are human skills: ability to learn, communicate and creativity – in the future we need creative, entrepreneurial IT specialists more than good programmers; many experts say that today are more relevant business skills [93]. These skills are far more difficult to develop. "We learn through experience and experiencing, and no one teaches anyone anything" (Viola Spolin), i.e. this is a message to students – if you want to learn, you have to work, just coming here does not make you wiser.

F. Flipped teaching

Often it is useful to suggest course participants to check out some on-line materials already before the time when they will be considered in classroom - the reversed/flipped teaching. This gives more time for classroom discussions and practice. We learn very little without practical exercises and even this little will soon be forgotten. With practice, the knowledge and skills needed to apply it remains with us for the whole life, just like the skill to drive bicycle - even after years without using a bicycle we do not have any problems. In the similar way work open-ended questions posed at the end of a lecture/practicum: "Now it was demonstrated, that this algorithm/process does not work as expected - how it should be modified?"

G. Diminishing role of lectures

Traditional university teaching is based on excessive use of lectures. But lectures can succeed only if audience is homogeneous and lecturer knows the audience: "The success of a lecture requires the lecturer to be aware of the capabilities of all students and for all students to have a similar background" [94]. Today this does not hold any more, thus the style of lectures should change. In classroom should not be considered technical details of programming or software systems - students already either are familiar with these topics from their work practice or should familiarize themselves with these technologies beforehand. Instead of a lecture is in classroom initiated communication with the whole group: about significance of technologies, how something could be used and/or improved, about innovations - what kind of technology is currently missing, but were good to have. Such discussions greatly promote creativity and entrepreneurship and restore the style of Socratic dialogue, where teacher is the mentor, not a squawk box.

In discussions created by the first author in his courses in Tallinn University of Technology come out some essential facts; many of them support the already established claims about 'digital natives' [95]. For instance, many students agreed, that their attention span has decreased and they do/can't deal with long texts - it does not have any sense to appoint them for home reading chapters from some textbook. But they are very good in dealing in parallel simultaneously with several topics - e.g. restricting use of laptops/tablets in lectures is totally absurd; besides, they forget their laptops as soon as the discussion in classroom becomes interesting. When the first author initiated in his course 'Multimedia' discussion about content and formatting of CV-s, nobody remained passive or looked his/her keyboard - they all understood how important this document will be in their lives - by current estimation this generation will have to prepare it more than fifteen times; several students already had haven more than one job, i.e. they are potential 'job-hoppers' [96], but some employers consider job-hoppers the best employees [97]..

For 'digital natives' some commonly used technologies seem already outdated, e.g. students do not value e-mail very high – mobile phone is more convenient. Laptop and WiFi in the whole university are the most essential conditions for successful studies (94% of students agreed).

Students considered quite professionally state of common programs and revealed problems: programming languages and systems are very brittle, error-prone and their syntax is non-uniform (this problem was discussed already in [98]); many program editors are nearly useless (e.g. the Flash IDE), error reporting is very crude (many students said, that after using the free ActionScript editor FlashDevelop [99] they will never use the Flash IDE again, although their employer has expensive Flash license); students indicated Prezi [100] as a new fresh style of presentations, but also agreed that the new html5 libraries (e.g. [101]) are already overtaking Prezi, these can be much more capable than Prezi in accessibility and extensibility, the main drawbacks (presently) are security policies and inconsistencies across browsers, which will certainly improve in the future.

Most students are using their laptops during the lecture/discussion and often find on the fly new facts/examples which were essential in the discussed topic - another example of their ability to deal simultaneously with several topics. Unfortunately our lecture hall technology allows not (yet) switching student laptops to projector so that their findings were visible to the whole group, but even now students were quite satisfied. These discussions promote just the ability to discover, analyze, compare and generalize.

H. Listen to students

As already said above, currently it is very difficult for a teacher to be aware of all new technologies and trends, students often know or discover something what is new to teacher. Therefore students are encouraged to use in their projects every suitable (free) technology or program. Leave it to students to discover what's simple and useful !.

Students often make very proper suggestions for topics of lectures and/or exercises – 'I saw tutorial of this technology, could we consider/try it in lecture/exercise' – always agree with such proposals (however many hours of sleep it takes to understand/master this new thing).

I. Gamification

Why is it so that people do not like to work, but can play videogames for hours?

Videogames have become a very essential part of modern culture and production of videogames - a very significant sector of SE and world-wide economy. In the last years the game industry grow worldwide more than twelve percent every year [102],[103].

Games give players rich and challenging experiences, support search and discovery – the main components of creativity. Many researchers and practitioners are trying to employ principles of game also to work and study processes [104].

Many people still have outdated stereotypes of gamers - 'pastime of young boys'. Currently the percentage of male gamers younger than 17 (13%) is less than the percentage of adult women gamers (37%) and nearly one-third (29%) of gamers are older than 50 years [105].

Today games are everywhere and gamers are social, tech-savvy, goal-oriented people who want to improve themselves and also the world around them [106].[107].

The principles of gamification can be applied to leaching/learning process in many ways [108]; there is also an free course "Gamification" on Coursera [109]. One of the easiest methods is presenting in class from time to time questions and small problems - who first answers correctly, will collect some bonus points which will be taken account in the whole course assessment.

XI. Conclusions and the Moral

To answer the question posed in the title of this paper: yes, (university) teachers will certainly survive - if the whole world would be educated using the same MOOC courses (however good) the whole world would be uniform, everyone similar to others - dull. There is all the time increasing need for systematization and preparing of

digestible mouthfuls of the all new information flooding us for students, and this presentation should remain individualized. We all see things a bit differently and these differences are important, they should not vanish in one world-wide MOOC. We have to explain to students the significance, importance of the considered issues in general context, not lot of details, classifications and historical development – anyone can get this kind of data from literature and Internet. If student does not understand, why this topic is important and how it is connected with issues considered earlier then there is no motivation for them to listen. Make students to think, not to memorize facts. Give them general understanding of issues in context and ability to use the right methodology in right place, formulas and details do not vanish and can be easily found.

Teachers will survive, but only those, who are constantly learning themselves, considering classroom and the whole teaching process as a two-way communication and students as active participants of this communication, who often can give significant contributions. There is no place any more to classical lecturing, which has been described as "the best way to ensure the transfer of information from the notes of the teacher to the notes of the students—without touching the students' mind". The modern educational paradigms are extensive use of virtual, interactive knowledge networks, collaboration and gamification, blending of classroom and on-line teaching in different forms of reversed (flipped) teaching, search (googling) and actively exploitation of rich depositories of on-line knowledge, just-in-time and just-when-needed "thrivable education" [110],[111]. Our's is a difficult, challenging but still very pleasurable profession which will survive.

REFERENCES

[1] Maslow, A (1954). Motivation and personality. New York, NY: Harper. pp. 1-236. ISBN 0-06-041987-3

[2] http://money.usnews.com/careers/best-jobs/rankings/the-100-best-jobs

[3] Doug Gross. Why disc drives are an endangered species. http://edition.cnn.com/2012/12/06/tech/innovation/imac-disc-drive

[4] http://www.paperonweb.com/Finland.htm

[5] Self-published e-book author: 'Most of my months are six-figure months'. http://edition.cnn.com/2012/09/07/tech/mobile/kindle-direct-publish

[6] Amazon selling more Kindle ebooks than print books. http://www.bbc.co.uk/news/technology-19148146

[7] PlayStation 4 should go all-in on cloud-streaming games. http://reviews.cnet.com/8301-9020_7-57570138-222/playstation-4-should-go-all-in-on-cloud-streaming-games/

[8] http://www.telegraph.co.uk/technology/news/9066721/3D-printer-builds-new-jaw-bone-for-transplant.html

[9] Mining asteroids to 3D-print space stations: Beyond pie in the sky? http://news.cnet.com/8301-17938_105-57565271-1/mining-asteroids-to-3d-print-space-stations-beyond-pie-in-the-sky/

[10] W. B. Arthur. Is the Information Revolution Dead? If history is a guide, it is not. Business 2.0, 3(3), March 2002, pp 65–72

[11] The Age Of The Informavore. http://www.edge.org/3rd_culture/schirrmacher09/schirrmacher09_index.html

[12] Ray Kurzweil, The Singularity is Near. Penguin Group, 2005, p.9

[13] Get bespoke tailoring in the comfort of your home with UPcload. http://www.digitaltrends.com/lifestyle/upcload-the-digital-tailor-that-sizes-you-up-with-a-webcam/

[14] Coming Soon: Toilet Paper Priced Like Airline Tickets. http://online.wsj.com/article/SB10000872396390444914904577617333130724846.html

[15] http://www.bls.gov/ooh/sales/sales-engineers.htm#tab-2

[16] http://www.zdnet.com/debate/great-debate-is-e-commerce-killing-brick-and-mortar/10086560/

[17] http://www.zdnet.com/blog/perlow/in-the-battle-of-clicks-versus-bricks-retail-must-transform-or-die/19418

[18] http://www.zdnet.com/blog/perlow/retail-in-2021-when-clicks-have-buried-bricks/19344

[19] http://www.hinnavaatlus.ee/

[20] http://www.wherethemuseum.com/2013/01/ants-laikmaa-house-museumhaapsaluestonia_nophoto/

[21] I. E. Allen, J. Seaman. Class Differences. Online Education in the United States, 2010

[22] Goldman Sachs Global Economics Group. Brics and Beyond. The Goldman Sachs Group, Inc, 2007

[23] E.E. Gordon. The future of Jobs and Careers. https://www.acteonline.org/uploadedFiles/Publications_and_E-Media/files/files-techniques-2009/Theme_4%283%29.pdf

[24] http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-SF-13-001/EN/KS-SF-13-001-EN.PDF

[25] http://tilastokeskus.fi/til/opty/index_en.html

[26] http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&language=en &pcode=teilm020&tableSelection=1&plugin=1

[27] EITO. ICT Market Report 2012/13. http://www.eito.com/epages/63182014.sf/en_GB/?ObjectPath=/Shops/63182014/Products/001-1206/SubProducts/001-1206-010

[28] Estonian Association of Information Technology and Telecommunications. Plan of Activity 2011–2013. http://www.itl.ee/?op=body&id=247

[29] http://www.epl.ee/news/eesti/it-asekantsler-kotka-it-tiiger-ei-maga-vaid-mollab-erasektoris.d?id=65588230

[30] Skype to sponsor Estonian IT Academy. http://www.baltic-course.com/eng/Technology/?doc=56557

[31] http://www.teenus.info/tag/matemaatika

[32] http://www.kuldnebors.ee/search/too-ja-koolitus/eratunnid/search.mec? pob_cat_id=11014&pob_action=show_cats

[33] Alison Doyle. How Often Do People Change Jobs? http://jobsearch.about.com/od/employmentinformation/f/change-jobs.htm

[34] Penelope Trunk. Why Job Hoppers Make the Best Employees. http://www.cbsnews.com/8301-505125_162-45040110/why-job-hoppers-make-the-best-employees/

[35] http://www.nowinnovations.com/news

[36] http://www.ut.ee/sites/default/files/www_ut/lipilastettamisefenomeneestikrghariduses2011.pdf

[37] Just-in-Time Education: Learning in the Global Information Age. http://knowledge.wharton.upenn.edu/article.cfm?articleid=236

[38] Marc Prensky. Digital Natives, Digital Immigrants. http://www.marcprensky.com/writing/prensky%20-%20digital%20natives,%20digital%20immigrants%20-%20part1.pdf

[39] T. Levin. Digital Natives and Their Customs. http://www.nytimes.com/2012/11/04/education/edlife/arthur-levine-discusses-the-new-generation-of-college-students.html

[40] IT Salaries Rising for Experienced Employees: Dice Report. http://www.eweek.com/c/a/Enterprise-Networking/IT-Salaries-Rising-For-Experienced-Employees-Dice-Report-158390/

[41] What is big data? http://www-01.ibm.com/software/data/bigdata/

[42] http://www.idc.com/getdoc.jsp?containerId=prUS23177411#.URf94mf4V8E

[43] R. Dawkins. The Selfish Gene. New York City: Oxford University Press, 1976.

[44] http://www.emc.com/about/news/press/2012/20121211-01.htm

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

[46] The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. http://www.emc.com/leadership/digital-universe/iview/index.htm

[47] Joel Mokyr. Useful Knowledge as an Evolving System: the view from Economic history. http://faculty.wcas.northwestern.edu/~jmokyr/Arrowfest.PDF

[48] Xionglei He, Jianzhi Zhang. On the Growth of Scientific Knowledge: Yeast Biology as a Case Study. http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000320

[49] http://www.idgconnect.com/blog-abstract/459/tanya-kalyan-global-the-rise-app-internet
[50] http://www.infoworld.com/d/application-development/developer-divide-19-generations-of-computer-programmers-209971?source=IFWNLE_nlt_daily_2013-01-07
[51] Samson, R.W. How to succeed in the hyper-human economy. IEEE Engineering Management Review 32:4, 2004, pp. 91-96
[52] Richard W. Samson. Hyperjobs: The New Higher-Level Work and How to Grow Into It. Futurist, November-December 2005 Vol. 39, No. 6, http://rogeliodavila.com/Creatividad/ Actividades/Future _Careers_%282%29.pdf
[53] European Commission Eurostat. Patent Statistics. . http://epp. eurostat.ec.europa.eu/statistics_explained/index.php/Patent_statistics#Main_statistical_findings
[54] The United States Patent and Trademark Office. http://www. uspto.gov/patents/process/search/issuyear.jsp
[55] http://www.ttu.ee/china
[56] The European Innovation Academy 2012 in Tallinn. http://www.ttu.ee/innovationacademy/home/
[57] Forecast #6: Professional Knowledge Increasingly Obsolete. . http://www.britannica.com/blogs/2008/12/forecast-6-professional-knowledge-increasingly-obsolete/
[58] M. Sahami, A. Aiken, J. Zelenski. Expanding the frontiers of computer science: designing a curriculum to reflect a diverse field. Proceedings of the 41st ACM technical symposium on Computer science education, ACM 2010, pp 47-51
[59] Jaakkola, H.; Henno, J.; Rudas, I.J. IT Curriculum as a Complex Emerging Process. IEEE International Conference on Computational Cybernetics, 2006. ICCC 2006, pp. 1-5.
[60] E. A. Freudenthal, M. K. Roy, A. N. Ogrey, T. Magoc, A. Siegel. MPCT: media propelled computational thinking. Proceedings of the 41st ACM technical symposium on Computer science education, ACM 2010, pp 37-41
[61] M. Guzdial, D. Ranum, B. Miller, B. Simon, B. Ericson, S. A. Rebelsky, J. Davis, K. Deepak, D. Blank. Variations on a theme: role of media in motivating computing education. Proceedings of the 41st ACM technical symposium on Computer science education, ACM 2010, pp 66-67
[62] M. Baranger — Chaos, Complexity, and Entropy. http://necsi.edu/ projects/baranger/cce.pdf
[63] Computer Science Curricula 2013. http://ai.stanford.edu/users/sahami/CS2013/
[64] http://www.roela.edu.ee/failid/dok/Roela_L-PK_oppekava_lisa-10-11.pdf
[65] http://teadus.err.ee/artikkel?cat=1&id=7846
[66] MIT OpenCourseWare. Site Statistics. http://ocw.mit.edu/ about/site-statistics/
[67] http://www.leidenranking.com/ranking.aspx
[68] http://ocw.mit.edu/index.htm
[69] http://www.youtube.com/watch?v=k6U-i4gXkLM
[70] http://news.stanford.edu/news/2011/august/online-computer-science-081611.html
[71] Daphne Koller. Death Knell for the Lecture: Technology as a Passport to Personalized Education. http://www.nytimes.com/2011/12/06/science/daphne-koller-technology-as-a-passport-to-personalized-education.html
[72] http://www.openculture.com
[73] http://www.insidehighered.com/news/2013/01/17/moodys-report-calls-question-all-traditional-university-revenue-sources
[74] https://www.coursera.org/stanford
[75] Campus extends reach to iPod generation. http://www.berkeley.edu/news/media/releases/2006/04/25_podcast.shtml
[76] MIT to make nearly all course materials available free on the World Wide Web http://web.mit.edu/newsoffice/2001/ocw.html
[77] Cautiously, Yale moves online . http://yaledailynews.com/ blog/2013/02/07/cautiously-yale-moves-online/
[78] https://www.coursera.org/
[79] OpenCulture. Free Online Certificate Courses & MOOCs from Great Universities: A Complete List. http://www.openculture.com/ free_certificate_courses
[80] R. C. Schank, Chip Cleary. Engines for Education. Routledge 1995, pp 248, http://www.routledge.com/books/details/9780805819458/
[81] http://www.khanacademy.org/
[82] R. Talbert. The trouble with Khan Academy. http://chronicle.com/ blognetwork/castingoutnines/2012/07/03/the-trouble-with-khan-academy/
[83] Martin C. Carlisle. Using You Tube to enhance student class preparation in an introductory Java course. SIGCSE '10 Proceedings of the 41st ACM technical symposium on Computer science education , ACM 2010, pp 470-474
[84] http://scratch.mit.edu/
[85] http://www.alice.org/
[86] http://www.yenka.com/
[87] http://www.codecademy.com
[88] 3DVinci. http://www.3dvinci.net/
[89] M. Möller. Multiscale Methods for (generalized) Sparse Recovery and Applications in High Dimensional Imaging. http://imaging.uni-muenster.de/teaching/Doktoranden/thesisMoeller.pdf
[90] http://www.html5canvastutorials.com/labs/html5- canvas-physics-engine-with-curve-detection/
[91] http://jsfiddle.net/zhanglei/LHWmv/embedded/result,js/
[92] D. Kuropatwa. The Difference Between Curriculum and Pedagogy. http://adifference.blogspot.com/2012/01/difference-between-curriculum-and.html
[93] Does a computer science degree matter anymore? http://www.networkworld.com/news/2009/040609-hot-tech-skills.html
[94] Laurillard, D. Rethinking university teaching: A framework for the effective use of educational technology. New York: Routledge 1993.
[95] 50 Surprising Facts About Technology Usage In College. http://edudemic.com/2011/12/tech-in-college/
[96] 0% of Gen Y Leave First Job within Two Years. http://www.experience.com/corp/press_release?id=press_release_1220541662048&tab=cn1&channel_id=about_us&page_id=media_coverage_news
[97] P.Trunk. Why Job Hoppers Make the Best Employees. http://www.cbsnews.com/8301-505125_162-45040110/why-job-hoppers-make-the-best-employees/
[98] J. Henno. User-friendly syntax: design and presentation. International Journal of Man-Machine Studies, v.28 n.5, p.551-572, May 1988
[99] http://www.flashdevelop.org/
[100] http://prezi.com/
[101] impress.js. http://bartaz.github.com/impress.js/#/bored
[102] Video Game and PC Game Market Package. http://dfcint.com
[103] Market Trends: Gaming Ecosystem, 2011. http://www.gartner.com/DisplayDocument?ref=clientFriendlyUrl&id=1724014
[104] http://gamification-research.org/
[105] The 2011 Essential Facts About the Computer and Video Game Industry http://www.theesa.com/facts/pdfs/ESA_EF_2011.pdf
[106] New Digital Citizenship Program Launched by ESA in Los Angeles. http://www.theesa.com/newsroom/release_detail.asp?releaseID=190
[107] The Future Of Gaming. A Portrait of the New Gamers. Latitude report, http://latd.com/wp-content/uploads/2011/08/ Latitude-FutureofGaming.pdf
[108] K. Werbach, D. Hunter. For the Win. How GAME THINKING Can Revolutionize Your Business. Wharton Digital Press, Philadelphia 2012
[109] https://www.coursera.org/course/gamification
[110] M.Holliday. Thrivable Education. http://humanity4point0.ning.com/group/builtspace40/forum/topics/ thrivable-education
[111] A.Laszlo, R.Rowland, T.Johnston, G.Taylor. Virtual Learning in a Socially Digitized World. World Futures: The Journal of General Evolution, Volume 68, Number 8, 1 November 2012 , pp. 575-594(20)

Cloud Computing in Education and Student's Needs

E. Krelja Kurelović*, S. Rako**, J. Tomljanović***
*Polytechnic of Rijeka, Vukovarska 58, Rijeka, Croatia
elena@veleri.hr
** University Computing Centre, J. Marohnića 5, Zagreb, Croatia
sabina.rako@srce.hr
*** Polytechnic of Rijeka, Vukovarska 58, Rijeka, Croatia
jasminka.tomljanovic@veleri.hr

Abstract: Cloud computing is not just a buzz-word, it represents a strong direction of IT industry development. Speaking of cloud computing we should distinguish three different service models: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). The scope of this work is a model of Software as a Service. This represents the lease of computing resources on a network of remote servers where applications are executed and data is stored. The application of cloud computing is very broad and growing daily because of many advantages to the users, and is driven by the increasing use of various mobile devices (laptops, tablets and smartphones) and mobile Internet access being more available. Cloud computing is applicable in education, but it implies the acceptance of these services by all involved in the educational process. Therefore, the aim of this paper is to investigate whether there is a need between our students for applications and services in the "cloud" (SaaS), the extent to which they use them and what types of applications and services are leading. The paper analyzes and interprets the results of this study which provides indications of students' willingness to "move to the cloud".

Keywords: cloud computing, Software as a Service, education, students

I. INTRODUCTION

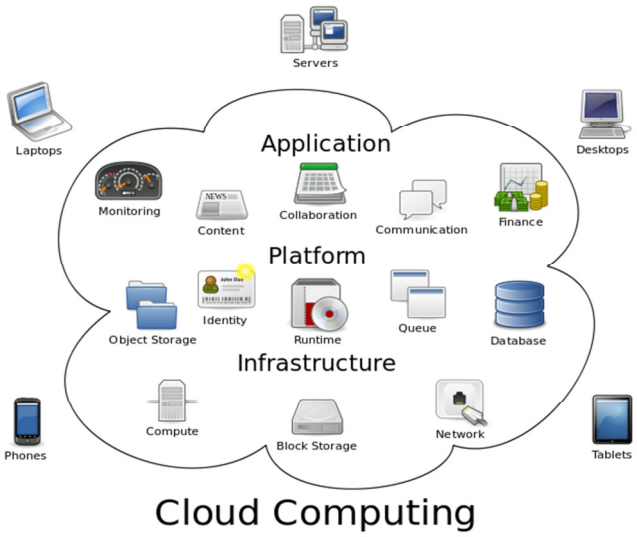
In the last couple of years "cloud computing" has increasingly been discussed. This is a relatively new trend of IT industry development, focused on users, and driven by the increasing use of various mobile devices such as laptops, tablet PCs and smartphones. Research has shown that it is one of the fastest growing sectors of the digital economy. European governments and industry plan to invest 45 billion euros in the development of cloud computing by the year 2020 [15]. In cloud computing networks of remote servers, storage systems (data centers and server farms) and their resources are being used upon user request. Term "cloud" is used as a metaphor for the Internet since it doesn't matter where the hardware and software resources that are used are located [14]. For IT professionals cloud computing is a new business model and a new technology platform for developing and deploying applications, and for end-users a new and cheaper way to use applications [3]. Cloud computing has many advantages but also some limitations, both arising from the fact that all data and applications are located somewhere on the Internet. It can be used in various activities of everyday life, including in education. In addition to providing students and teachers (usually free of charge) access to many applications and services in the cloud, which can be used in formal and informal education, cloud computing allows for greater flexibility and mobility in the use of resources for teaching and learning, greater degree of collaboration, communication and sharing of resources, and creates a personalized learning environment or virtual communities of learning and teaching.

II. THEORETICAL FRAMEWORK

National Institute of Standards and Technology of U.S. Department of Commerce defines cloud computing as a „model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction“ [10]. Gartner defines cloud computing as a „style of computing in which scalable and elastic IT-enabled capabilities are delivered as a service using Internet technologies“ [16]. „For everyday users of the Internet and computers, cloud computing is any online activity, such as accessing data or using a software program, which can be done form different devices regardless of the on-ramp to the Internet, as in Figure 1. In this vision, the data or software applications are not stored on the user's computer, but rather are accessed through the web from any device at any location a person can

get web access.“ [6]. For end-users, cloud computing means that you don't have to worry about maintaining hardware or purchase new equipment, obtaining software licenses, updating or upgrading existing software, data synchronization, etc. because all of these are included in the "cloud" service. One can say that cloud computing is the new driver of IT revolution, in which new IT services are being developed, changing the ways of access, usage, maintenance and financing services on demand. Cloud computing is characterized by scalability (extent and amount of used resources according to the needs of the application and paid on the actual use of resources), mobility and platform independency (the ability to access anytime, from any location and device). There are three types of cloud computing: Infrastructure as a Service (IaaS) is the hardware component with different forms of virtual technology rentals, platform as a service (PaaS) involves the use of the operating system and development tools in the cloud and software as a service (SaaS) which refers to the use various web-based applications that run and execute on the server [12]; [3]. These models differ in the type and extent of resources accessed and managed by users. Connecting to the cloud and using "hidden" resources enables sharing of information always and everywhere, great application scalability, service availability anytime and anywhere, data security, storage, backup copies, and more [4]. „The potential benefits of adopting cloud computing can be assessed from both the financial savings and resource management perspectives“ [8].

Figure 1. Representation of cloud computing



Source: http://en.wikipedia.org/wiki/Cloud_computing

III. CLOUD COMPUTING IN EDUCATION

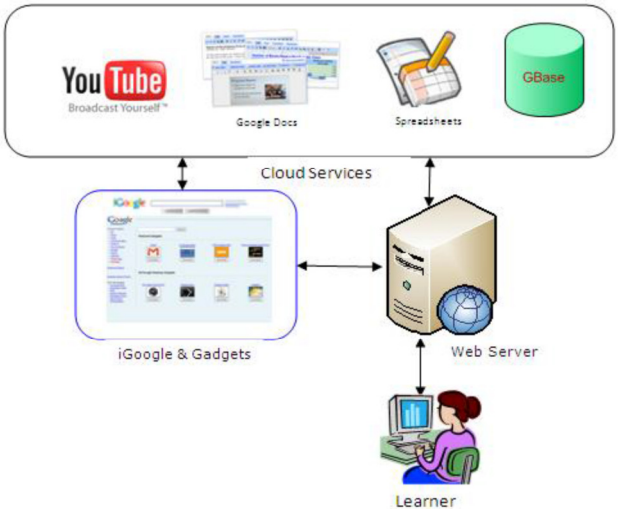
According to [11], „the potential of cloud computing for improving efficiency, cost and convenience for the educational sector is being recognized by a number of US educational establishments. For some universities, the availability of an awesome computing power through cloud computing for research purposes was welcome“. „Many educational institutions have begun their movement to cloud computing by outsourcing their student email provision … Educational institutions are also beginning to use lower level cloud services for purposes such as data storage. This may be attractive where data security is of lower concern such as where video and audio is provided as open educational resources. Another use of cloud computing which is beginning to emerge in education is for the hosting of institutional learning management systems (LMSs) in the cloud. Outsourcing the provision of LMSs such as Blackboard or Moodle to a third party makes sense for institutions who cannot justify the costs of purchasing, maintaining and supporting the hardware and software themselves“ [5]. Cloud computing is often associated with e-learning and m-learning [9]. This refers to e-learning environment mainly distributed on the cloud, in which Open Educational Resources were produced, researched and shared by participants worldwide [7].

With applications in the cloud (SaaS), students and teachers can flexibly access their data via a web browser from a computer at home, school, library, student room or some other place, and achieve rapid and efficient communication, collaboration, exchange or share documents, contacts, notes, audio / video and other data. With their use students can create „Cloud-Based Personalized Learning Environment“ (Figure 2) [1]. "The first idea that comes to mind when assessing such a cloud space for learning, would be the creative potentials that could be nurtured i.e. the endless ideas, thoughts and knowledge that could be shared, created and inspired" [7].

In addition to individual applications in the cloud, bundled applications are also available (eg. Google Apps for Education or Microsoft Live@edu with Office 365 and other applications for Education) that combine tools for communication and collaboration, office tools for working with documents, and space to store and synchronize data on demand. „Whereas a university computing service department may aim to achieve 99.5% availability for its educational services such as the LMS, Google offers 99.9% availability for its educational application suite and appears to outperform this target „ [5].

Using services and applications in the cloud, students and teachers can achieve mobility because their educational resources and necessary applications are available via portable computers and Internet-connected devices. For example, classes can be implemented outside the school / faculty or students can perform duties at various places.

Figure 2. Example of Cloud-Based Personalized Learning Environment



Source: Al-Zoube, M. [1]

According to Kop and Carroll, „cloud computing has the potential for new interaction metaphors and new ways of thinking about learning design and learning experiences“. However, capabilities of cloud computing should be recognized by all participants of educational process in order to find its uses and application in education. According to Gartner analysis the cloud is used for "school and educational services" by only 4% [2]. Another research [9] shows that only 12% of respondents are familiar with the use of cloud computing in education while 88% of them would agree that cloud computing should be implemented in the education sector.

IV. PROBLEM AND HYPOTHESES

The broader problem domain of this paper is to investigate the preconditions for the implementation of cloud services and applications in higher education. This implies the acceptance and active use of such services by students and teachers. The study is focused on the use of cloud services (SaaS) for students. The reason for selecting students as a target population of research lies in the fact that students are increasingly dependent on online services for learning and assessment [5].

The aim of this study is to determine if students show a need for cloud services and applications (SaaS) and how often they use them. It also aims to investigate the relationship between the reported need for cloud services and their active use, since it may not always be correlated. According to goals, the following null hypotheses are posted:

H1a: There is no significant difference in reported need for cloud services between male and female students.

H1b: There is no significant difference in reported use of cloud services between male and female students.

H2: There is no positive correlation between reported need for cloud services and frequency of use them.

H3: There is no positive correlation between number of devices from which students access the Internet and frequency of cloud services use.

V. RESEARCH METHODOLOGY

Research began in November 2012 and lasted until mid-January 2013. The subjects of this study were students from the University of Rijeka. For the purpose of this research a questionnaire (consisting of three sections) was created in GoogleDocs - Forms and distributed online. The first section collected general data about the respondent (gender, faculty, year of study) and data on the use of computers (number of computers and similar devices, frequency of accessing Internet). This was followed by eight statements in which participants estimated the frequency of situations indicating the need for cloud applications and services, based on a 4-point scale. The third section of the questionnaire contained 10 questions about the frequency of use of cloud applications and services. Here we listed cloud applications and in parentheses names of the most popular ones from categories, for respondents to be clear on what the question was about. In statements and questions in the second and third section of the questionnaire, the responses were given on a 4-point frequency scale (1 = none, 2 = rarely, 3 = sometimes, 4 = often / very often). The questionnaire was distributed through the website of the Student Union, Polytechnic of Rijeka. The survey was anonymous, and the questionnaire was completed by 158 students. Empirical data were analyzed using descriptive statistics, and hypothesis testing was done using nonparametric tests: Mann-Whitney U test and Spearman correlation.

VI. RESULTS AND ANALYSIS

The questionnaire was completed by 158 respondents; 58% males and 42% females. The proportion of students was 78% from the Informatics study program and 22% from other study programs; all from the Polytechnic of Rijeka. A vast amount of the students (90%) attended professional studies (first 3 years), and only a small proportion of students (10%) were from specialized studies (4th and 5th year).

On the question "Which devices do you use for accessing Internet and data?" respondents could select multiple answers, and distribution is shown on Figure 3. The respondents mostly use laptop computers (74%) and smartphones (67%) indicating their mobility. A lower number of users selected desktop computers at home (64%) and faculty (54%), and the smallest

number used tablet computers (13 %), and ultra-portable computers (5%).
Respondents reported the frequency of Internet access, and responses were grouped into three categories: weekends or a few times a week (5%), several times a day (18%), all of the time or whenever I get an opportunity (77%). Another survey by [17], on a population of 15-30 years, indicates that 80% of respondents use online services daily, and nearly 50% use the Internet via mobile phone [17].

Items with statements about frequency of situations that indicate a need for services in the cloud are grouped by category of service to which the statement relates. Respondents expressed the frequency on a 4-point scale, and the distribution of responses is shown in Table 1. These data show that the respondents expressed the greatest need for communication software to the cloud, then cloud multimedia sharing, cloud docs / office software, and the smallest need for cloud storage and file synchronization software. However, mean scores indicate that all of the needs for services in the cloud fall into the category "sometimes", with similar standard deviation.
Through the 10 questions in the final section of the questionnaire respondents expressed the frequency of the active use of various cloud services (SaaS). Question items are grouped and response distribution is shown in Table 2. Arithmetic mean of responses indicate that respondents use cloud communication software often and very often, they sometimes use cloud services for multimedia sharing and learning, and are little less likely to use cloud storage and file synchronization software and cloud docs / office software. The largest scattering in answers was regarding to using cloud services for multimedia sharing and learning.
Research [6] analyzing cloud computing activities of online Americans also reported frequency of cloud communication services being most used, and least use of cloud storage and file synchronization software, while the cloud docs / office software use was in the middle.

Figure 3. Use of computers and devices for accessing Internet and data, distribution of answers

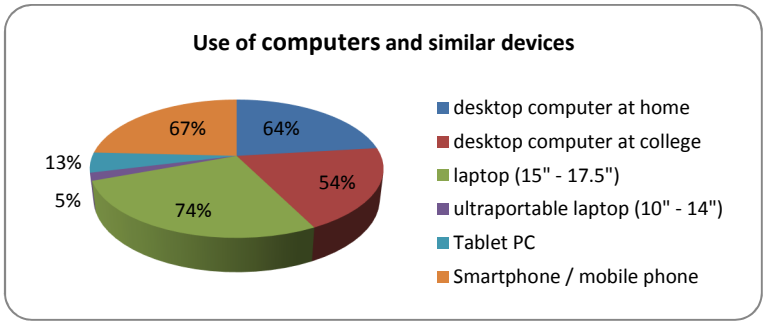


Table 1. Expresses need for cloud services, descriptive statistics

Category of service:	M	SD
Cloud storage and file synchronization SW (2 items)	2,52	0,76
Cloud docs / office software (2 items)	2,82	0,79
Cloud services for multimedia sharing (2 items)	3,05	0,85
Cloud communication software (2 items)	3,38	0,70

Table 2. Frequency of use of cloud services, descriptive statistics

Category of service:	M	SD
Cloud storage and file synchronization software	2,49	1,05
Cloud docs / office software	2,46	0,98
Cloud services for multimedia sharing	2,79	1,15
Cloud communication software	3,62	0,76
Cloud services for learning	2,55	1,00

VII. TESTING THE HYPOTHESES

A Mann-Whitney U test was conducted to evaluate the H1a hypothesis that there is no difference in need of using cloud computing services between male and female students. The results of the test were not significant, according to z = -0,109; p = 0,98 (Table 3). Thus, we can accept the stated null-hypothesis.

A Mann-Whitney U test was also conducted to evaluate the H1b hypothesis which says that there is no difference in using cloud computing services between male and female students. The results of the test were not significant, as indicated z = -1,054; p = 0,29 (Table 4). Thus, we can accept the stated null-hypothesis.

Although in the past there were significant differences in the frequency and way of using computers between men and women, so people would talk about "gender-gap", recent data shows that there is no big difference when it comes to the use of cloud services and applications [18]; [19]. Some experts believe that the road to increasing the number of female in IT sector is through cloud computing … because the characteristics of the "cloud", such a big impact on socializing, collaboration and project management, increase the interest of the women for IT [13].

For testing H2 we have used Spearman's Correlation Coefficient. The results shown in Table 5 indicates an association between variables need and use of cloud computing services and applications with rs = 0.563 and statistical significance of $p < 0.05$. The statistical analysis has shown moderate correlation. Thus, we can discard the hypothesis that there is no positive correlation between expressed need of cloud computing services and frequency of their use. Therefore we can expect that students who recognize the need for cloud services in their everyday activities on the computer will begin more actively use some of those cloud services and applications.

The third hypothesis (H3) was also tested using the Spearman Correlations. The Spearman's Correlation Coefficient indicates an association between variables "number of devices" and "use of cloud computing services" with rs = 0.164 and statistical significance of $p < 0.05$ (Table 6). But, the statistical analysis has shown weak correlation with borderline statistical significance.
The cloud services were designed for people who often use different computers and mobile devices, which is what these students do (according to Figure 2). However, in this study the correlation between these two variables has not proved strong enough as would have been expected to.

Table 3. Results of testing H1a

variable	Mann-Whitney U Test (Spreadsheet1) By variable spol Marked tests are significant at p <,05000						
	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-level	Z adjusted	p-level
potreba oblak	7308,500	5252,500	3030,500	-0,019390	0,984530	-0,020967	0,983272

Table 4. Results of testing H1b

variable	Mann-Whitney U Test (Spreadsheet1) By variable spol Marked tests are significant at p <,05000						
	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-level	Z adjusted	p-level
uporaba oblak	7015,000	5546,000	2737,000	-1,05414	0,291821	-1,12573	0,260281

Table 5. Results of testing H2

Variable	Spearman Rank Order Correlations (oblak) MD pairwise deleted Marked correlations are significant at p <,05000	
	potreba oblak	uporaba oblak
potreba oblak	1,000000	0,563467
uporaba oblak	0,563467	1,000000

Table 6. Results of testing H3

Variable	Spearman Rank Order Correlations (oblak) MD pairwise deleted Marked correlations are significant at p <,05000	
	Br uređ	Uporaba oblak
Br uređaja	1,000000	0,164522
Uporaba oblak	0,164522	1,000000

VIII. CONCLUSION

Cloud computing represents a new model of providing IT services which includes rental of resources located somewhere in the "cloud" and is considered to be the direction of future development of the IT sector. Ordinary people are increasingly online and more work is done online, from checking e-mails and using other forms of communication, writing and editing documents and collaboration, through watching movies and videos or listening to music, up to the fact that personal documents and images are being stored on web. There is no need for installation, licensing and updating of program, and there are no maintenance costs because all programs and services are available through a web browser. Cloud computing (SaaS) is based on "on-demand self-services" with "pay as use" model, although for ordinary users and for applications in education such software is usually free of charge. Cloud computing usability in education is very broad, as recognized by many educational institutions around the world. The reasons for the worldwide introduction of cloud computing for educational establishment are mainly of a financial nature, but it should be noted that „cloud“ has creative potentials because it enables that ideas, thoughts and knowledge can be created, used and shared easily. Students may create their own "Cloud-Based Personalized Learning Environment" or use m-learning and access to Open Educational Resources from the cloud. However, to achieve this, students and teachers should be willing to use services in cloud and be familiar with their advantages and limitations. The study shows that the respondents (students) are often online and at the same time using multiple computers and similar devices (tablet computers, smartphones). When using computers they are (on average) sometimes in situations that require the use of applications and services in the cloud. The largest use is noted with cloud communication software, the smallest with cloud docs / office software. Gender difference in reported need for services in the cloud, or in the frequency of use of applications in cloud is not confirmed. However, a significant correlation between the reported need for cloud applications / services and the frequency of their use is found. There is also a correlation between the number of computers and mobile devices used for Internet access and frequency of use of cloud applications / services. The students have expressed, based on the average of their answers, that they “sometimes” need and use cloud applications / services, which means that they are not yet ready for a "move to the cloud". It is necessary to educate and motivate them about cloud applications and services to make them become aware of the benefits.

REFERENCES

[1] Al-Zoube, M., E-learning on the Cloud, International Arab Journal of e-Technology, Vol.1, No.2, 2009, http://www.iajet.org/iajet/iajet_files/vol.1/no.2/E-Learning%20on%20the%20Cloud.pdf (28. 12. 2012)

[2] Bittman, T., Cloud Computing Inquiries at Gartner, 2009, http://blogs.gartner.com/thomas_bittman/2009/10/29/cloud-computing-inquiries-at-gartner/ (12. 01. 2013)

[3] Bronzin, T., „Cloud Computing“ ili programska rješenja u oblacima, Pogled kroz prozor, 2009, http://pogledkrozprozor.wordpress.com/2009/08/29/%E2%80%9Ecloud-computing%E2%80%9C-ili-programska-rjesenja-u-oblacima (24. 01. 2013)

[4] Bronzin, T., Kaldi, V., Mobilnost u oblaku, Infotrend, 186, 2011.

[5] Cloud Computing in Education, IITE Policy Brief, UNESCO, 2010, http://iite.unesco.org/pics/publications/en/files/3214674.pdf (29. 01. 2013)

[6] Horrigan, J.B., Use of Cloud Computing Applications and Services, 2008, http://www.pewinternet.org/Reports/2008/Use-of-Cloud-Computing-Applications-and-Services.aspx (26. 01. 2013)

[7] Kop, R., Carroll, F., Cloud Computing and Creativity: Learning on a Massive Open Online Course, European Journal of Open, Distance and E-Learning, 2011, http://www.eurodl.org/index.php?article=457 (21. 01. 2013)

[8] Lin, A., Chen, N., Cloud computing as an innovation: Perception, attitude, and adoption. International Journal of Information Management, 2012, http://dx.doi.org/10.1016/j.ijinfomgt.2012.04.001 (27. 01. 2013)

[9] Mallikharjuna Rao, N., Sasidhar, C., Satyendra Kumar, V., Cloud Computing Through Mobile Larning, International Journal of Advanced Computer Science and Applications, Vol.1, No. 6, December 2010, http://thesai.org/Downloads/Volume1No6/Paper_7_Cloud_Computing_Through_Mobile-Learning.pdf (28. 01. 2013)

[10] Mell, P., Grance., T., Final Version of NIST Cloud Computing Definition, NIST, 2011, http://www.nist.gov/itl/csd/cloud-102511.cfm (25. 01. 2013)

[11] Sultan, N., Cloud computing for education: A new dawn?, International Journal of Information Management, no. 30, 2010, 112. str., http://www.sciencedirect.com/science/article/pii/S0268401209001170 (28. 01. 2013)

[12] Žganec, T., na nebu ni na zemlji, Infotrend 184, 2011.

[13] *** Cloud Computing Will Increase the Number of Woman in IT. Netpremacy Global Services Investigated How, PRWEB UK, 2012, http://www.prweb.com/releases/2012/9/prweb9878941.htm (02. 02. 2013)

[14] *** Cloud Computing, Wikipedia, http://en.wikipedia.org/wiki/Cloud_computing (14. 01. 2013)

[15] *** EU želi uložiti 45 milijarde eura u razvoj računarstva u oblaku, 2012, http://dalje.com/hr-zivot/eu-zeli-uloziti-45-milijardi-eura-u-razvoj-racunarstva-u-oblaku/446783 (20. 01. 2013)

[16] *** Gartner, IT Glossary, http://www.gartner.com/it-glossary/cloud-computing (15. 01. 2013)

[17] *** GfK Croatia, Neka zapažanja o mladoj „online populaciji u Hrvatskoj, http://www.gfk.hr/public_relations/press/press_articles/008744/index.hr.html (20. 01. 2013)

[18] *** Survey: Use of cloud–based brands in the United States in 2011, by gender, Statista, http://www.statista.com/statistics/199167/use-of-cloud-based-brands-by-gender-in-the-united-states-by-gender/ (02. 02. 2013)

[19] *** Survey: Use of cloud–based services in the United States in 2011, by gender, Statista, http://www.statista.com/statistics/198928/use-of-cloud-based-services-by-gender-in-the-united-states-by-gender/ (02. 02. 2013)

Mobile Phone in the Classroom

Željka Požgaj, Vesna Bosilj Vukšić
Department of Informatics
Faculty of Economics and Business
Trg J.F.Kennedy-a 6. 10000 Zagreb, Croatia
Phone: 01-238-3333, Fax: 01-2335-633, e-mail: zpozgaj@efzg.hr; vbosiljj@efzg.hr

Abstract - **Mobile phones have changed the way we live but they have also greatly affected the education sector during the last few years. This statement is particularly true for Generation Y also known as the Millennial Generation. Nowadays, mobile learning (m-learning) creates a new paradigm for modern education, but the question about students' readiness to use mobile phones in learning process is still opened. Participants in this study consisted of undergraduate students of the Faculty of Economics (University of Zagreb). During the survey, a questionnaire was used to collect the needed data. The purpose of the survey was to investigate (1) the use of mobile phones during lectures; (2) the students' mobile phone addiction and (3) the students' perspective towards the purposes they use mobile phones nowadays and their predictions for the future. (3) the students' opinion about the purpose of the mobile phones nowadays and their possible future The results were following: (According to the research results) a majority of students use their mobile phones to access the educational materials and information, but data also show a very frequent mobile phones' adoption for a wide range of inappropriate purposes (such as entertainment and social networking) during the classes.**

I. INTRODUCTION

Intrigued by the fact that undergraduate students of the Faculty of Economics and Business (University of Zagreb) often use mobile phones more or less openly during the lectures, we wanted to examine the reasons behind such behavior. Students who use mobile phones during the lecture are obviously not concentrating on the lecture itself, they are at the same time disturbing their colleagues causing their level of concentration to drop as well which ultimately prevents the lecturer from doing his or her job properly. Based on a small sample research, we argue in this paper for the need to recognize the students' relationship with mobile phones and hence develop our understanding of their needs, routines, attitudes and perceptions of the mobile phones usage in the classroom.

According to Yu [19] the popularity of mobile phone usage indicates that the number of phone users tripled in 2012 compared to 2009 (from 165 million in 2009 to over 500 million in 2012). The research "*How Croats live*" conducted by the Mediana agency [4] shows that 96.3% of Croats have a mobile device (while the EU average is 87%). On the other hand, mobile devices such as 3G mobile phones, PDAs and pocket PCs are increasingly used for educational purposes and educational settings [17]. Mobile device applications can be used as a study aids that students can access from virtually anywhere [18]. However, the availability of mobile devices does not guarantee their use in education: to realize these benefits, students must first adopt m-learning [5]. In the last years there has been a growing interest from educationalists and researchers to explore students' practices and perceptions on the use of mobile devices in the process of learning. Although mobile devices are ubiquitous on college campuses, student readiness for mobile learning has yet to be fully explored [2].

Our argument is that mobile learning could be relevant for our undergraduate students. Therefore this study investigated the current state of undergraduate students' perception toward (1) the use of mobile phones during lectures; (2) the students' mobile phone addiction; and (3) the current role and future trends of mobile phones in their life.

The paper is organized in the following manner. First, the mobile learning in higher education is reviewed (Section 2). Next, the research methodology and the research results are described (Section 3). Finally, the results are discussed (Section 4), followed by the conclusions and directions for future research.

II. MOBILE LEARNING IN HIGHER EDUCATION

Numerous studies have explored the impact of using mobile technologies on learning process in higher education institutions [17]. According to the results of recent literature overview, mobile phones can be valuable in higher education. Mobile phones can facilitate information collection, time management, access to content, communication and collaboration [6]; [7]; [13]. The use of wireless transmission and mobile devices such as PDAs, mobile phones, laptops and tablet PCs is defined m-learning [14]. Pinkwart et al. [11] defined m-learning as e-learning that uses mobile devices. M-learning is similar to e-learning since both of these approaches provide multimedia contents and communication with other students, but it is unique in terms of flexibility of time and location [9].

However, some obstacles for the implementation of m-learning in higher education institutions are recognized by researchers during last few years. Schepman et al. [13] systemized these as followed: (1) embedding mobile learning is not always straightforward since not all students own mobile devices to support mobile learning; (2) mobile learning is not universally accepted, with some users displaying a lack of engagement or dislike; (3) not all institutions have the funds to commission tailor-made

applications for different learning settings; and (4) learners very often prefer traditional computing platforms which utilize relatively generic software which can be tailored to different learning tasks.

The benefits of instant access to learning anytime and anywhere is an enormous benefit, but will be restricted until the technology of wireless data access matures and educators learn how to apply appropriate pedagogical forms [7]. Some studies show that students are not likely to use mobile devices for learning because of the limitations of m-learning [8]; [16]. First, some technical limitations are pointed out, such as the small screens, inadequate memory, slow network speeds. Second, students are more likely to use mobile devices for entertainment (like texting with friends and checking social network services) rather than for educational purposes. Third, there are pedagogical obstacles - using mobile phones in class may hinder student concentration.

A very important issue to be examined in the future is the inappropriate use of mobile phones during the classes. A mobile phone can also be used for performing unauthorized actions during the written exam. There are three types of using the mobile phones in a prohibited way: mobile phone as a place to store exam material (formulas, definitions, questions and answers from the previous exams) in order to check on it during the exam; using the mobile phone to go online and check the exam materials stored on the web and using the mobile phone to communicate with others who are writing the same exam or somebody else outside the classroom. This third example of illicit use of mobile phones during the exam can be achieved not only through simple text messaging but through a device that consists of a headset (miniature speaker called a "bug" that is placed in the ear) and a microphone that is hidden inside the clothes. The person in possession of such a device is communicating to another person usually sitting outside the classroom. A person using the device then receives the answers to the questions through a "bug". This technique was very popular during the year 2010 so one of the University of Zagreb faculties had to install instruments that jam mobile phone signals during the exams. Because the Electronic Communications Act allows jamming mobile phone signal only in exceptional circumstances prescribed by the law, Croatian Post and Electronic Communications reacted negatively and the Faculty had to remove the jamming technology.

III. THE RESEARCH

A. Methodology

A survey was carried out at the Faculty of Economics and Business, University of Zagreb, Croatia, during the lecture in Informatics, on first-year students, in January 2013. The research approach is carried out in order to find out:

- *How students use their mobile phones during the classes?*
- *Are the students addicted to mobile phones?*
- *What is the students' perception about the way they use mobile phones during the classes and their opinion about the future of mobile phones?*

The questionnaire was prepared by the authors. Content and validity of questions were investigated by 2 experts of educational technology and information technology and were found to be satisfactory. The questionnaire consisted of 3 sections. *The first section* (4 items) asked the participants why, how and for what purposes they use their mobile phones during the classes. The aim of the *second section* (3 items) was to examine whether the students are addicted to mobile phones. *The third section* of questionnaire, consisting of 3 items, was prepared to evaluate students' perceptions of the current status of mobile phones and the future trends on smartphones adoption in learning process. Questionnaires were distributed to 71 students. Besides, the participants were asked about the gender. The majority of students were female (53) and the rest were male students (18). Research results were analyzed using MS Excel.

Our research, however, has one significant limitation. The research sample could be more extensive in order to get more generalized results. It is hard to conclude from a single-country study whether the results are valid in different countries and other research contexts. It is desirable that similar studies be conducted on a larger and international sample of higher education institutions so as to be able to make valid conclusions regarding possible differences among institutions of different size, or from different countries.

B. Results

How students use their mobile phones during the classes?

Using mobile phones during the lecture can, without a doubt, interfere with the whole process of teaching and learning. If students' mobile phones are not turned off during the lecture, it is very likely that at least one mobile phone will begin to ring which will interrupt the professor and cause a general disorder in the classroom, laughing and teasing the phone's owner will surely ensue. If students' mobile phones have only the sound turned off, it is very likely that the students will notice and respond to the notifications appearing on the screen of the phone. Not only will that lessen their attention span, but also the attention span of the student sitting next to them. Turning the phone completely off would be the most desired option, but as the survey shows (Table 1) that rarely happens. A majority of students appear to have their mobile phones turned on but with the sound turned off during the lectures, as we ourselves were able to notice during our lectures.

TABLE 1: STATUS OF STUDENTS' MOBILE PHONES IN CLASSROOM

"During the classes my mobile phone is:"	**N**	**%**
Turned on	8	11.27
Turned on, sound off	60	84.51
Turned off	3	4.22
Total	71	100.00

The next (multiple choices) question was asked in order to examine for what purposes students use their mobile phones during the lectures (Table2).

TABLE 2: FOR WHAT PURPOSES STUDENTS USE THEIR MOBILE PHONES DURING THE LECTURES?

"During the classes I use my mobile phone for:"	**N**
Checking up on Facebook	41
Reading/replying e-mails	16
Sending/replying text messages	13
Listening to music (with headphones on)	11
Playing games	24
Watching clips on YouTube	2
Checking the time	3
Using the application What's app	3
Using the mobile phone for other activities	5
Mobile phone is turned off	2
Total	120

The results in the Table 2 indicate that during the lecture, students are mostly interested in what is happening on Facebook, reading and sending e-mails or text messages. Next in popularity is playing games or listening to music. According to the results we could divide the students who use mobile phones during the lectures into two groups. The first group would be the active group, those who check their Facebook, read and send e-mails or text messages, play games or use their mobile phones for other activities. The rest who use mobile phones for checking the time or watching YouTube clips or are listening to music belong to the passive group.

Besides, the authors wanted to find out how students use their mobile phones in accessing educational materials and information published on the faculty's' Web site. The results show that 76% of students use their mobile phones to access "short" information such as test results, schedules, consultation hours etc while only 24% of students use their mobile phones for other activities on the faculty's Web site.

Finally, the students were asked how they access the learning materials (presentations, texts, case studies, etc) on the faculty's Web site (Table 3). The results show that only 4% of students use exclusively their mobile phones in order to access learning materials. 45% of students access their learning materials exclusively via personal computers while 49% of students access their learning materials either via mobile phones or via personal computers. Therefore it can be concluded that although students use mobile phones to view information, they still prefer to use personal computers to upload or download learning materials.

TABLE 3: THE USE OF MOBILE PHONES AND/OR PERSONAL COMPUTERS TO ACCESS THE LEARNING MATERIALS ON THE FACULTY'S WEB SITE

„In order to access the learning materials, I use:"	**N**	**%**
exclusively mobile phone	4	5.63
exclusively personal computer	32	45.07
mobile phone and personal computer	35	49.30
Total	71	100.00

Are the students addicted to mobile phones?

There are some very interesting results from a research announced on Cisco Connected World Technology Report (CCWTR) for the year 2012 [3]. The results are based on a survey conducted by InsightExpress, it was conducted on 1800 students and young working people aged 18 to 30 across 18 countries. The results indicated that 90% of respondents check their smartphones for e-mail updates or social network notifications before they go to sleep or in the morning, as a part of their routine. 60% of respondents admitted to subconsciously or compulsively checking their smartphones for e-mails, texts and other social network updates. One in five checks his or her smartphone for an update at least once every ten minutes. Nearly 70% of respondents said that mobile applications are important for their day to day lives.

In our research (Table 4) based on the statement "I cannot imagine my life without a mobile phone" we found out that 61% of the students involved said that they strongly or mostly agreed with the statement, 26% said that they partially agreed with the statement, and more than 12% said that are disagree. Since almost 90% of students cannot imagine their life without a mobile phone it can be concluded that a majority of population is addicted to mobile phones.

TABLE 4: A MOBILE PHONE ADDICTION

"I cannot imagine my life without a mobile phone	N	%
Strongly agree	15	21.13
Mostly agree	28	39.44
Partly agree	19	26.76
Mostly disagree	7	9.85
Strongly disagree	2	2.82
Total	71	100.00

Considering the results of studies [10]; [12] that indicate young people's addiction to social networks, especially Fabecook, we wanted to learn the students' attitudes about that phenomenon as well. To the question "Are you addicted to Facebook?" 14 students answered positively, 56 students answered that they did not consider themselves addicted to Facebook and 1 student did not answer the question.

For further analysis we focused on students going on Facebook during lectures (Table 5). From the Table 2 we can see that the answer "Checking up on Facebook" got

chosen 41 times (each student could choose more than one offered answer), meaning that 41 students follow Facebook during the lectures but 27 of them do not feel addicted to Facebook.

TABLE 5: THE USE OF FACEBOOK DURING THE CLASSES

A Facebook addiction?	**N**	**%**
During the classes I check Facebook because I am addicted to it.	14	19.72
During the classes I check Facebook, but I do not feel addicted to it.	27	38.03
During the classes I do not check Facebook and I do not feel addicted to it.	29	40.84
No answer	1	1.41
Total	71	100.00

The analysis has shown that these results correspond with other international studies considering the same problem. Therefore, we could conclude that students' might not be completely aware of their own addiction to mobile phones.

What is the students' perception about the current ways and reasons of their use of mobile phones during the classes and the adoption of smartphones in the future?

We found interesting to examine the students' perception about the way they use their mobile phones during the classes.

We were curious about how many students are aware that such behavior disturbs not only them as the users of the phone but their colleagues and professors as well (Table 6).

TABLE 6: DOES THE USE OF MOBILE PHONES DISTURB ANYONE IN THE CLASSROOM?

Using a mobile phone during the classes disturbs	**N**	**%**
The user of the phone	17	23.94
Only colleagues	0	0.00
Only teacher	6	8.45
The user and colleagues	0	0.00
The user and teacher	4	5.63
Teacher and colleagues	2	2.82
The user + colleagues + teacher	9	12.68
No one	31	43.66
No answer	2	2.82
Total	71	100.00

Research results in Table 6 indicate that more than 54% of students are aware of the fact that using mobile phones disturb themselves and/or their colleagues and/or their professors but they still carry on with such behavior. 44% of them believe that they are not disturbing anyone, but in reality only the 4% of students who actually turn off their mobile phones completely (see Table 1) do not disturb anyone.

When we asked the students why they insist on using their mobile phones during the class, we've been told that they were generally satisfied with the way the lectures were presented but that they had a very strong desire to constantly be in touch with what is happening on their social networks or with their friends. The results show that such behavior could as well be categorized as a mobile phone addiction, but obviously students do not have the right perception of this fact.

Thanks to the development of operating systems, smartphones' functions are evolving and they are becoming more and more similar to personal computers. Some authors [1]; [15] argue that smartphones will soon replace personal computers. We asked our students about their opinions regarding the possibility of smartphones replacing personal computers in learning process in the near future (Table 7).

TABLE 7: A FUTURE OF SMARTPHONES IN RELATION TO PERSONAL COMPUTERS

"Smartphones will replace personal computers in learning process in the near future"	**N**	**%**
Strongly agree	17	23.94
Mostly agree	27	38.03
Partly agree	16	22.54
Mostly disagree	7	9.86
Strongly disagree	4	5.63
Total	71	100.00

We can conclude that the respondents are not completely sure if they agree with the idea of smartphones replacing personal computers: 38% agree that it is possible for smartphones to replace personal computers, 23% strongly agree and in total 85 % agree (strongly, mostly or partially). It can be concluded that personal computers are the most commonly used devices for learning, but these may be replaced in the future by new emerging technologies such as smartphones.

IV. CONCLUSION

The goal of this paper was to examine the role of a mobile phone in the classroom. The results are based on students' opinions and attitudes which are examined through the questionnaire and are presented in the study.

The authors began with the obvious problems: students use their mobile phones as a tool for entertainment or communication during the lectures thus disturbing their colleagues and professors and hindering their own learning process. They also use their mobile phones for cheating on the exams. However, mobile phones can also be used in order to help the learning process during the lecture. Students can use them in order to solve simple tasks or find useful information. In a time of change, higher education must adopt to new conditions created by widespread adoption of mobile phones. Therefore, those involved in teaching profession are faced with the need for adaptation to advancements in technology for teaching and learning.

The most important findings of this study can be systemized as follows:

- Students are aware that using mobile phones during the lecture disturbs both other students and the lecturer.
- The main reason for using mobile phones during the lecture is entertainment.
- Students use their mobile phones to access "short" information on the faculty's Web site and they access learning materials either via mobile phones or via personal computers.
- Students believe that smartphones could replace personal computers in the near future.

Based on these results, students want to use new technologies such as mobile phones in education because they use them in everyday life. The findings of the study propose a powerful approach to facilitate the learning process of students through the use of mobile phones: m-learning needs to become an integral part of higher education. Mobile phones and personal computers are expected to be replaced by smartphones. This will positively impact the dynamic of m-learning adoption in higher education. Finally, mobile phone users' behavior must be strongly addressed: students should change their perception of the mobile phone use in the classroom. The results of this research could provide decision support for educators who wish to explore how to change students' behavior and attitude and how to focus them toward using mobile phones as learning tools

REFERENCES

[1] S. Anthony, "There can only be one: Smartphones are the PCs of the future", available at: http://www.extremetech.com/computing/134868-there-can-only-be-one-smarthphones-are-the-pcs-of-the-future, accessed: 12 th January 2013.

[2] J. Cheon, S. Lee, S.M. Crooks, J. Song: "An investigation of mobile learning in higher education based on the theory of planned behavior", Computers & Education, 59 (2012), pp.1054-1064.

[3] Cisco Connected World Technology Report (2012), available at: http://newsroom.cisco.com/feature/1114851, accessed: 23 rd December 2012

[4] Jutarnji list, "Kako žive hrvati", available at: http://www.jutarnjlist.hr, accessed: 27th December 2012.

[5] J. Keller, "The slow-motion mobile campus", The Chronicle of Higher Education, May 13, 2011, pp.84-86.

[6] S. Martin, G. Diaz, E. Sancristobal, R. Gil, M. Castro, J. Peire, " New technology trends in education: Seven years of forecasts and convergence", Computers & Education, Vol. 57, No. 3, 2011, pp.1893-1900.

[7] L.F. Motiwalla, "Mobile learning: A framework and evaluation", Computers & Education, No. 49, 2007, pp.581-596.

[8] Y. Park, "A pedagogical framework for mobile learning: categorizing educational applications of mobile technologies into four types", International Review of Research in Open and Distance Learning, 12(2), 2011, pp.78-102.

[9] K. Peters, "M-learning: positioning educators for a mobile, connected future", International Review of Research in Open and Distance Learning, 8(2), 2007, pp.1-17.

[10] A. Patterson, " Social-networkers of the world, unite and take over: A meta-introspective perspective on the Facebook brand", Journal of Business Research 65 (2012), pp. 527–534

[11] N. Pinkwart, H.U. Hope, M. Milrad, J. Perez, "Educational scenarios for the cooperative use of Personal Digital Assistants", Journal of Computer Assisted Learning, 19(3), 2003, pp.383-391.

[12] L.D. Rosen, K. Whaling, S. Rab, L.M. Carrier, N.A. Cheever. „Is Facebook creating ''iDisorders''? The link between clinical symptoms of psychiatric disorders and technology use, attitudes and anxiety", Computers in Human Behavior, 29 (2013), pp.1243-1254

[13] A. Schepman, P. Rodway, C. Beattie, J. Lambert, "An observational study of undergraduate students' adoption of (mobile) non-taking software", Computers in Human Behavior, 28(2012), pp.308-317.

[14] P. Seppala, H. Alamaki, "Mobile learning in teacher training", Journal of Computer Assited Learning, 19(3), 2003, pp.330-335.

[15] J. Spurbeck, "Three Ways Future Smartphones May Repleace PCs", available at: http://news.yahoo.com/three-ways-future-smarthphones-may-replease-pcs-174400632.html , accessed: 14th January 2013.

[16] Y.S. Wang, M.C. Wu, H.Y. Wang, " Investigating the determination and age and gender differences in the acceptance of mobile learning", British Journal of Educational Technology, Vol. 40, No. 1, 2009, pp.92-118.

[17] R. Wang, R. Wiesemes, C. Gibbons, "Developing digital in fluency through ubiquitous mobile devices: Findings form a small-scale study", Computers & Education, 58(2012), pp.570-578.

[18] J.R. Young "Smartphones on campus: the search for "killer" apps", The Chronicle of Higher Education, May 13, 2011, p.92-118.

[19] F. Yu, "Mobile/Smarth Phone Use in Higher Education", available at: www.swdsi.org/swdsi2012/proceedings_2012/paper/PA144.pdf , accessed: 14th January 2013

Future-oriented education with a modern flexible production unit

C. Ungermanns, W. Werth
Carinthia University of Applied Sciences (CUAS)
Europastraße 4, 9524 Villach, Austria
Phone: +43 4242 90500
c.ungermanns@fh-kaernten.at
w.werth@fh-kaernten.at

Abstract - This paper describes the realization of a modern flexible production unit – as output of the international SIAT project "Koop Flexible Automation", which combines various disciplines of process automation, e.g. robotics, process control for CNC-based manufacturing, image processing and analysis [1]. It will be also shown how this production unit is used in education for trainees with different levels of previous knowledge. For this purpose a modular program was developed and a startup with pupils, students and employees of SMEs (Small and medium enterprises) has been performed. The concept of the educational training and first experiences will be discussed.

I. INTRODUCTION

Today's production processes need both flexibility and a high level of automation [2]. Beside these two manufacturing aspects an important customer requirement is the quality of the final product. Therefore within the international SIAT project "Koop Flexible Automation" various flexible cells have been designed. These production units are used in teaching modern automation topics for different target groups. Figure 1 shows the flexible unit of the Carinthia University of Applied Sciences in Villach. It covers all proposed aspects: robotics, flexibility, automation and quality management. The device consists of a transportable part, which can be also used separately in a lecture room and easily decoupled from the production CNC-machine and the CNC-tool. The objective of this unit is on the one hand the implementation of a flexible, fully automated production line with special emphasis on robotics and quality aspects, on the other hand the possibility of sharing this knowledge with pupils, students and SMC (small- and medium- sized companies). It is important to mention that the system was constructed which especially fulfills the didactic demands on teaching and learning.

Due to these objectives the central part of our systems is an industrial robot. This can be seen in the most left part of Fig. 1 and in Fig. 4. Beside that an optical inspection system was designed to produce relevant data for the quality control process.

As an overview the entire process is shortly described in the next section. Then the specific training modules at CUAS are mentioned and two training units are described more detailled.

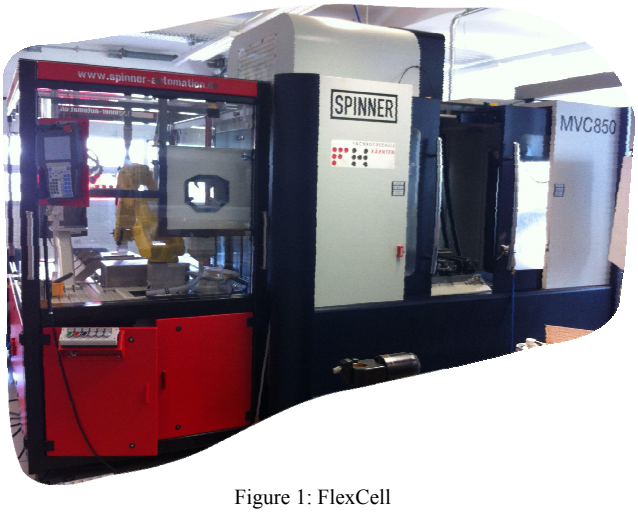


Figure 1: FlexCell

II. PROCESS AUTOMATION

The coaction of all units is described in the following section.

The process, shown in Fig. 2, starts with the manual loading of the unprocessed part into the system. In *S1*, the first optical inspection system determines the position, the size, and the orientation of the part and communicates this information to the robot. The robot handles the part into the milling machine (*S2*) where they get labeled. After the processing is done, the robot handles the part to *S4* (drying station).

Next, the robot handles the part into *S1* where a second optical inspection system reads the labeling. After this step is complete, the robot handles the part again to *S2* where is it milled. After an ultrasonic cleaning (S3) and a drying step (S4) the finished work piece is optical characterized with a high resoluted optical measurement system (*S5)*. The measured data from all optical inspection systems is stored in a central database for online statistical analysis. Finally, the robot handles the part to *S6* where, depending upon the measurement result from *S5* and the online analysis, the part is classified as OK, out of control limit or scrap (out of specification). The process

automation also allows for a classroom mode, where the CNC milling machine is absent. Instead, a placement area is present where the teacher can exchange an unprocessed part by a processed part.

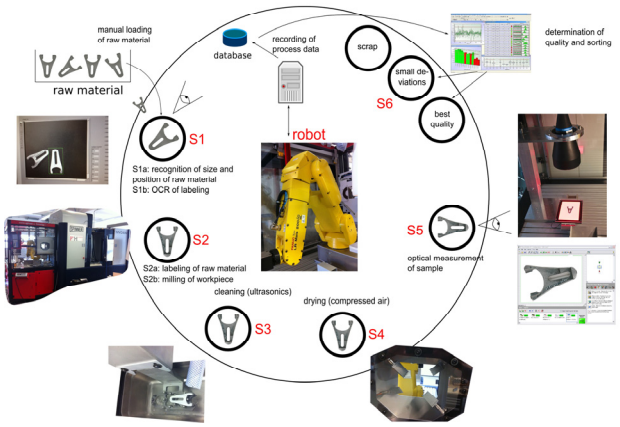


Figure 2: FlexCell Process Automation Overview

III. FLEXIBLE TRAINING

Due to different flexible units designed and used by the project partners a lot of specific training modules have been prepared during the project. Furthermore teachers and researchers in different fields of automation processes can impart their expert-knowledge. In the following Fig. 3 a summary of the prepared modules is given. Note that also an estimated duration for each module was proposed.

Module	Hour
CNC Manual Programing (milling + turning)	8
CNC Cam Programing (milling + turning)	8
Robotics basic	8
Robotics advanced	6
Quality management	4
Optical measurement systems	4
Process control	6
Machine control and embedded systems I	6
Machine control and embedded systems II	6
Safety basic	5
Safety advanced	3
Design I	6
Design II	6
Economics and organization	8

Figure 3: Training modules overview

The red colored part in the figure indicated the modules prepared at CUAS which are described in the next section.

IV. TRAINING MODULES AT CUAS

In this section the specific training parts, which have been given at CUAS are described in details. Basically the training modules where presented to two different target groups. First of all pupils and undergraduate students have been trained, secondly the same topics where presented to teachers in terms of a train the trainer program. It has to be mentioned that the main objective of the second training program was only to instruct our partners about our specific contents and pedagogical methods. This is important because all students are trained in different topics by trainers from all partners. During the project it turned out however that an arbitrary exchange of teachers would need intense and time consuming teaching procedures.

The main focus in all cases was to explain the proposed topics by the FlexCell device. In particular at our location four modules have been trained: robotics, image processing, profibus and statistical process control (SPC). The training program for the modules robotics and SPC are described in details, whereas for the remaining modules just a summary is given.

A. Robotics

Due to the fact that the industrial robot represents the main part of the FlexCell a teaching module “Robotics” is seen as very important. In flexible automation controlling a robot is one of the essential technical tasks an engineer has to cover [7]. On the other hand robotics includes a lot of different sophisticated problem-settings. In details we mean for example analysis of robots dynamical behavior, trajectory planning and robot control. Most of them need some prior technical knowledge or at least some “technical feeling”. Furthermore there is usually the need of an advanced mathematical background if the participant of a training program has to understand and solve robotics tasks in details. In addition some general conditions related to the training program had to be fulfilled:

- The pupils came from different schools in Austria and Slovenia.
- The schools are vocational schools.
- The pupils participated the third year (high school).
- Each training module was given within 4 hours.
- Participants got detailed materials for further discussions in class, no pre-education by the teachers was given.

Furthermore all the contents in class as well as all lecture notes have been presented or prepared in English. Due to the fact that our participants – mostly pupils – did not have the desired knowledge we proceeded in the following way. First of all the main tasks robots in a flexible automation process have to perform are discussed and explained by the FlexCell device, e.g.

- General task description
- Basic elements: joints, coordinates
- Prepare the task for the robot: Trajectory planning
- Make the robot move: Joint control

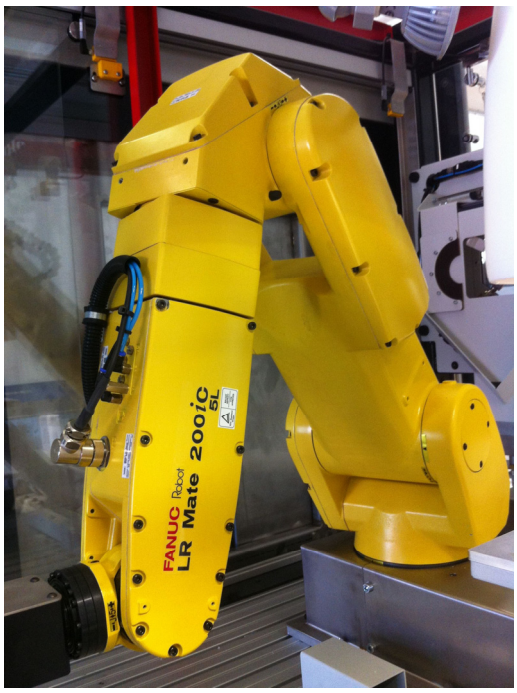


Figure 4: FlexCell Robot

At that point two important aspects have to be discussed. First some theoretical background related to robotics is necessary to understand the robots behavior. Secondly simulation tools are always used to verify a suggested performance or control of the robot. Considering the FlexCell - robot in Fig. 4 was deemed to be too difficult and but also not necessary to achieve our pedagogical goals. Hence a simpler robot-system was selected to elaborate on some selected task in details. In the following figure (Fig .5) the discussed system is given. It consists of two DC-Motors (so two joints) and a flexible link. The main task of the robot system is to command the *x* and *y* position of the tip of the robot.

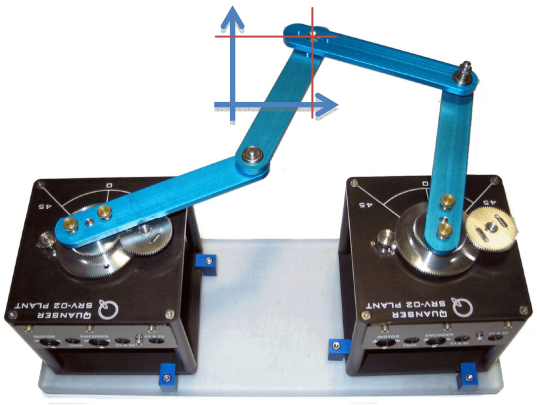


Figure 5: 2 DOF robot experiment

Hence the open tasks which are needed for the FlexCell robot can be discussed and solved in the same way at the simpler system, e.g.

- Kinematic model and dynamic model of a robot
- Control of a joint
- Simulate the robots movement
- Implement the controller to a real word system

In all of our courses MATLAB/SIMULINK is a standard tool for simulation of robotic systems and also for implementation of different control concepts at real world experiments. Deriving and discussing the mathematical model was the first step. Instead of deriving the mathematical model using analytical methods – which have not been well known to the pupils – an experimental setup using SIMULINK given in Fig. 6 was used to get the required data of the system, e.g. gain and time constants. Thereby black box models and suitable system responses (e.g. step response) have been considered. The determined parameters have been associated to the selected black box model (second order system) that describes the dynamic behavior of the joints.

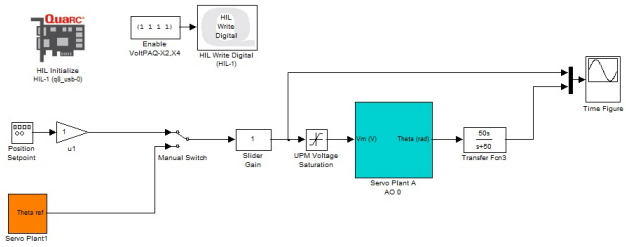


Figure 6: Experimental Setup

The same environment was taken to explain the need of a feedback concept (Fig. 7) and to discuss standard controllers like PID-controllers.

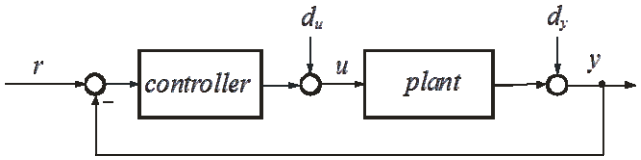


Figure 7: A feedback system

The next step was to control the two joint individually by experimentally found controller. Thanks to the used environment the result from our real world system (robot) are directly shown in class. To control the tip of the robot the connecting links have to be considered. So both the kinematic and inverse kinematics of the robot could be introduced and explained quite naturally. Implementing the kinematic models and performing the control of the tip finalized the robotics training session.

B. Profibus

The third part- profibussystem [3] contained the communication between the stages. Therefore a profi bus is used, which coordinates the communication in the FlexCell. The teacher was talking about the needs and the advantages of the use of a bussystem in an automation process. Furthermore was presented how the protocol is built, which information and messages are included and how the access to the data is managed. A big point was also the safety integrity. There are many norms and requirements for which application the bussystem is used. The profi bus is authorized up to SIL3 (safety integrity level).

C. *Image processing*

The topic of the fourth part was image processing [4]. Due to that the measurement process is done by cameras, the presentation was about how cameras could be used for such an application. One point was to discuss how a picture has to be manipulated so that this information can be evaluated. Therefore the questions about how a colour can be described and which patterns are useful were discussed.

D. *Statistical Process Control*

In addition to robotics and image processing, quality management is an essential part of educational program. Nowadays nearly in every modern industrial enterprise a quality management system is installed and make use of statistical methods to guarantee best possible stability of the manufacturing line. Very common is the concept of "Statistical process control" (SPC) [5,6]. In this section the application of SPC to FlexCell is demonstrated and described.

Experimental setup: after installing and calibrating the optical measurement system, a set of 50 work pieces was produced and measured. Fig. 8 shows three characterized feature sizes of the sample

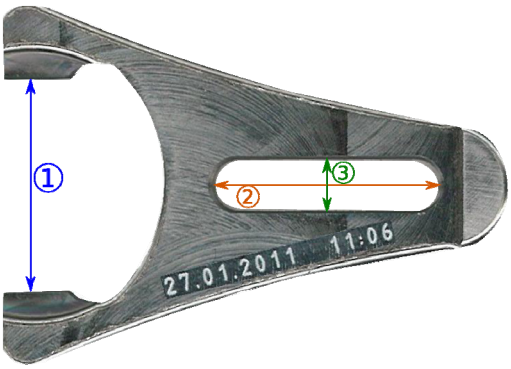


Figure 8: Work piece and measurement points

In Fig. 9 the dataset of feature size 1 (blue arrow) is represented as a cumulative frequency plot for all work pieces.

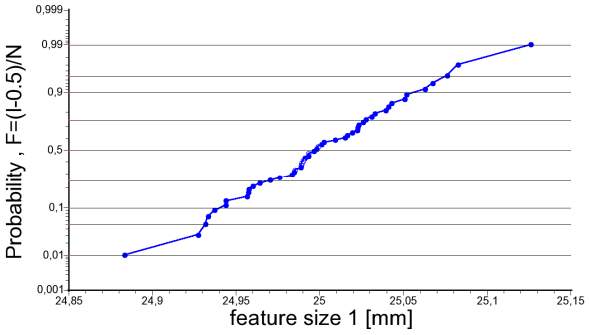


Figure 9: Probability plot of dataset

It is clearly seen from Fig. 9 that the readings are gaussian distributed as expected for a stable production since there is a linear curve in the probability graph. Therefore the prerequisites for the implementation of SPC are fulfiled. Now the upper control limit (UCL) and the lower control limit (LCL) are required for the control chart and can be obtained by the formulas

$$UCL = \bar{x} + 3s \tag{1}$$

and

$$LCL = \bar{x} - 3s \tag{2}$$

In the presented case it is UCL = 25,14 mm and LCL = 24,86 mm. The associated control chart is shown in Fig. 10.

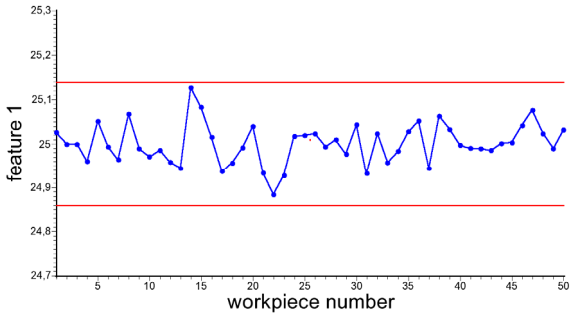


Figure 10: Control chart of stable production

An additional set of 50 work pieces was fabricated. But this time the dimensions were varied intentionally for didactic reasons. The purpose of the variation was to simulate production problems and to check, if the fully automated system is able to detect these problems. The basic criterion for an unstable production is the violation of the control limits. The probobility that these limits are exceeded randomly is only 0,26%. Therefore a violation is interpreted as an instability of the overall process. In this case the production is stopped immediately and a message pops up to the operator. Corrective actions must be taken by the responsible process engineer. Besides the work piece is placed to the scap slot.

In Fig. 11 the violation of the control limits are automatically detected and clearly marked by the software.

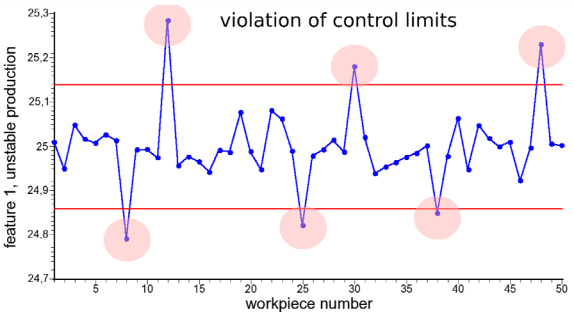


Figure 11: Control chart of unstable productio

V. Discussion and Outlook

In this paper the project FlexCell, which combines various disciplines of industrial automation, like robotics [7], image processing and analysis and CNC-based manufacturing has been described in details. Main focus was to present the essential training modules given at CUAS in Villach.

It has been shown that the device can be successfully used for teaching pupils and students of different levels of education. Furthermore the process will help technical staff of small and medium-sized companies to create modern solutions for their own process automation tasks. An industrial robot represents the central part of the construction. Therefore robot handling and programming is of particular importance in the training modules. Additionally the great relevance of quality management in any industrial process is clearly manifested in the training program of the project.

Figure 12: Training session in Villach

It is important to mention that the manufacturing process is of course not limited to the considered items described in the paper. It can be used to build and to analyze different CNC-based production lines, especially for small-batch production. Furthermore the proposed process device can easily be extended to additional tasks that have to be considered in an adapted or even new automation process.

Fig. 12 shows one of the training sessions given at Villach. Pupils participating at the training program gave very positive feedback to the presented sessions. Most of them pointed out that the combination between basic theoretical knowledge and practical demonstration was well balanced.

The project is continued in the way that both the knowledge - related to the designed automation devices - as well as the training program is extended to education institutions in Turkey and Croatia.

Acknowledgment

Special thanks to the Company Spinner Automation Markgrönigen (Germany) for the cooperative collaboration.

References

[1] www.flexcell.eu

[2] J. Love, *Process Automation Handbook: A Guide to Theory and Practice,* Springer, 2007

[3] J. Weigmann, G. Kilian, *Decentralization with PROFIBUS DP/DPV1: Architecture and Fundamentals*, Wiley-VCH, 2004

[4] T. Klinger, *Image Processing with LabVIEW and IMAQ Vision*, Prentice Hall, 2003

[5] J. S. Oakland, *Statistical Process Control*, Taylor & Francis, 2007

[6] T. Stapenhurst, *Mastering Statistical Process Control: A Handbook for Performance Improvement Using Cases*, Butterworth Heinemann, 2005

[7] B. Siciliano, O. Khatib, *The Handbook of Robotics*, Springer, 2008

[8] C. Madritsch, C. Ungermanns, T. Klinger, W. Werth, *An Innovative Industrial Automation System Showcase for Quality Management and Statistical Process Control Lectures*, MIPRO 2012, Croatia, 2012

An Approach to the Evaluation of Embedded Engineering Study Programs

Ivan Kastelan*, Moshe Barak**, Vlado Sruk***, Margarita Anastassova****, Miodrag Temerinac*
* University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
** Ben-Gurion University of the Negev, Be'er Sheva, Israel
*** University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
**** CEA, LIST, Sensory and Ambient Interfaces Laboratory, 91191 - Gif-sur-Yvette CEDEX, France
ivan.kastelan@ieee.org

Abstract -This paper presents an approach to the evaluation of study programs in the field of embedded computer engineering. The study programs cover the following subject categories: 1) digital system design, 2) computer system design, 3) digital signal processing, 4) computer networking and 5) system integration. The results of the evaluation of hardware, software and instructional materials in embedded engineering learning are presented. Student assignments are evaluated using the proposed tasks taxonomy which consists of three levels: 1) exercises, 2) problems, 3) projects. Students and teachers were asked to analyze the difficulty, lab material, level of student understanding and the amount of individual and team work in the subjects. First research results suggest that there is a large overhead in the number of platforms used in the mentioned courses. It was shown that in learning embedded computer engineering, the students deal merely with doing basic exercises and solving simple problems, while additional work is required in shifting the learning process towards enhancing students' higher-order cognitive skills such as problem solving and creativity, and fostering teamwork. The research also showed that it would be helpful to develop a unified embedded engineering learning platform for multiple courses in the curriculum.

I. INTRODUCTION

This paper presents an approach to the evaluation of a study program in the field of embedded computer engineering. We discuss the rationale for developing three-level taxonomy for evaluating the teaching and learning of subjects in computer engineering, and present the findings from the application of this tool in the field.

Engineering education today has been influenced increasingly by a range of fundamental principles derived from constructivist learning theories [1], [2]. Educators recognize more and more that: knowledge is constructed by learners, not passively transmitted by teachers; learning develops by assimilating new knowledge into the learners' prior knowledge and cognitive structures; and learning is collaborative. Engineering education that adopts these notions must shift from traditional teacher-centered instruction to learner-centered pedagogies such as problem-based learning and project-based learning, which results not only in better content knowledge but also in promoting students' higher-order competences such as critical thinking, creative thinking, problem solving, meta-cognitive skills and communication and teamwork skills [3]. Yet, educators should consider that engaging students in problem-based learning (PBL) cannot take place in a vacuum, and learners must acquire some basic knowledge and skills in a specific area before they can handle complex engineering projects. With these ideas in mind, we sought to develop a constructivist methodology for designing assignments of increasing complexity for engineering students.

The rest of the paper is organized as follows: section II discusses the three-level tasks taxonomy. Section III explains the evaluation method proposed by this paper. Section IV summarizes the findings from the application of the proposed evaluation method in the field. Finally, section V gives conclusions and future work.

II. TASKS TAXONOMY

The three-level tasks taxonomy illustrated in Fig. 1 was derived partially from Bloom's well-known taxonomy of educational objectives in the cognitive domain [4], the revised version of this taxonomy [5], the Problem Solving Taxonomy (PST) in engineering proposed by Plants, Dean, Sears and Venable [6], and Barak's [7] discussion of using taxonomies in engineering and technology education.

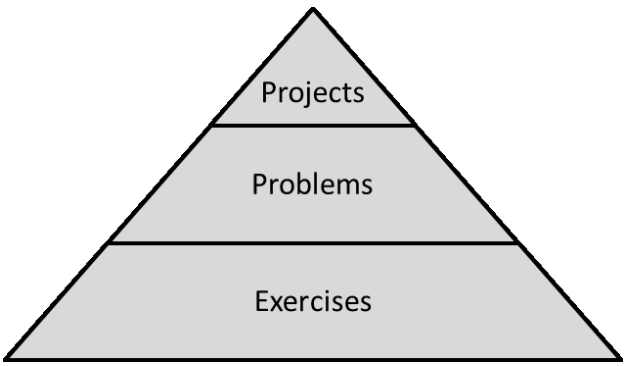


Figure 1. Tasks taxonomy in engineering education

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 317882.

The Tasks Taxonomy consists of the three levels described below.

A. Level 1 - Exercises

Exercises are well defined assignments in which the solving process and expected results are known in advance and the learners can check if they arrived at the correct solution.

This type of task is often used in the first steps of learning a new subject, for example, learning a new software or hardware, drill or practice.

In terms of Bloom's taxonomy of educational objectives (revised version), exercises could be classified within Level 1 – Remembering and Level 2 – Understanding.

B. Level 2 - Problems

Problems are open-ended small-scale tasks in which the students might arrive at different solutions or use different solving methods.

The proposed solution must meet given specifications and constraints, for example, available resources.

In terms of Bloom's taxonomy of educational objectives, problem solving by engineering students could be classified within Level 3 – Applying and Level 4 – Analyzing.

C. Level 3 - Projects

Projects are challenging ill-defined tasks in which the students take part in determining both the objectives and the required resources for a system's development.

Project work is aimed at fostering participants' technical knowledge, collaborative work, aspiration and imagination.

In terms of Bloom's taxonomy of educational objectives, engineering projects could be classified within Level 5 – Evaluation and Level 6 – Creating.

The learning outcomes that are expected from the students are well reflected in the task taxonomy described above. Firstly, we believe that it is important for the students to acquire basic theoretical knowledge and practical skills in learning a specific subject in embedded engineering, for example, digital systems and computer system design (see the next section). In the second stage, we expect that learners would be able solve small-scale problems in which they have to choose the problem-solving method or suggest their solution to a problem. Finally, the program intends to engage students to deal with broad projects in embedded engineering in which they develop not only deep understanding of the subject matter, but also higher-order thinking skills, for example, the ability to use complex non-algorithmic thinking to solve a problem, foster creativity and teamwork.

III. Evaluation Method

The proposed evaluation approach was applied in the field in the evaluation of current study programs in the field of embedded engineering in two Universities. The students and teachers were asked to fill the questionnaire about their current experiences in learning embedded computer engineering.

The five subject areas the questionnaires aim to evaluate are:

- Digital system design
- Computer system design
- Digital signal processing (audio, video)
- Computer networking and interfacing
- System integration.

Both questionnaires (students and teachers) consisted of six questions each. Each question asked the student or teacher to give a quantitative response about a particular statement.

Students were asked to answer the following questions about each subject:

1. In learning this subject we deal with basic exercises, drills and practice.
2. In learning this subject we deal with solving open-ended small-scale problems or design tasks to meet given specifications and constraints.
3. In learning this subject we are engaged in challenging projects in which we take part in defining the problem and the design objectives.
4. In learning this subject we face difficulties.
5. In learning this subject we work in teams.
6. In learning this subject we develop creativity.

Teachers were asked to answer the following questions about each subject:

1. In learning this subject, the students deal with basic exercises, drills and practice.
2. In learning this subject, the students deal with solving open-ended small-scale problems or design tasks to meet given specifications and constraints.
3. In learning this subject, the students are engaged in challenging projects in which we take part in defining the problem and the design objectives.
4. In learning this subject, the students face difficulties.
5. In learning this subject, the students work in teams.
6. In learning this subject, the students develop creativity.

Each question was responded to with one of the following five answers: (1) very much, (2) much, (3) no opinion, (4) little and (5) very little.

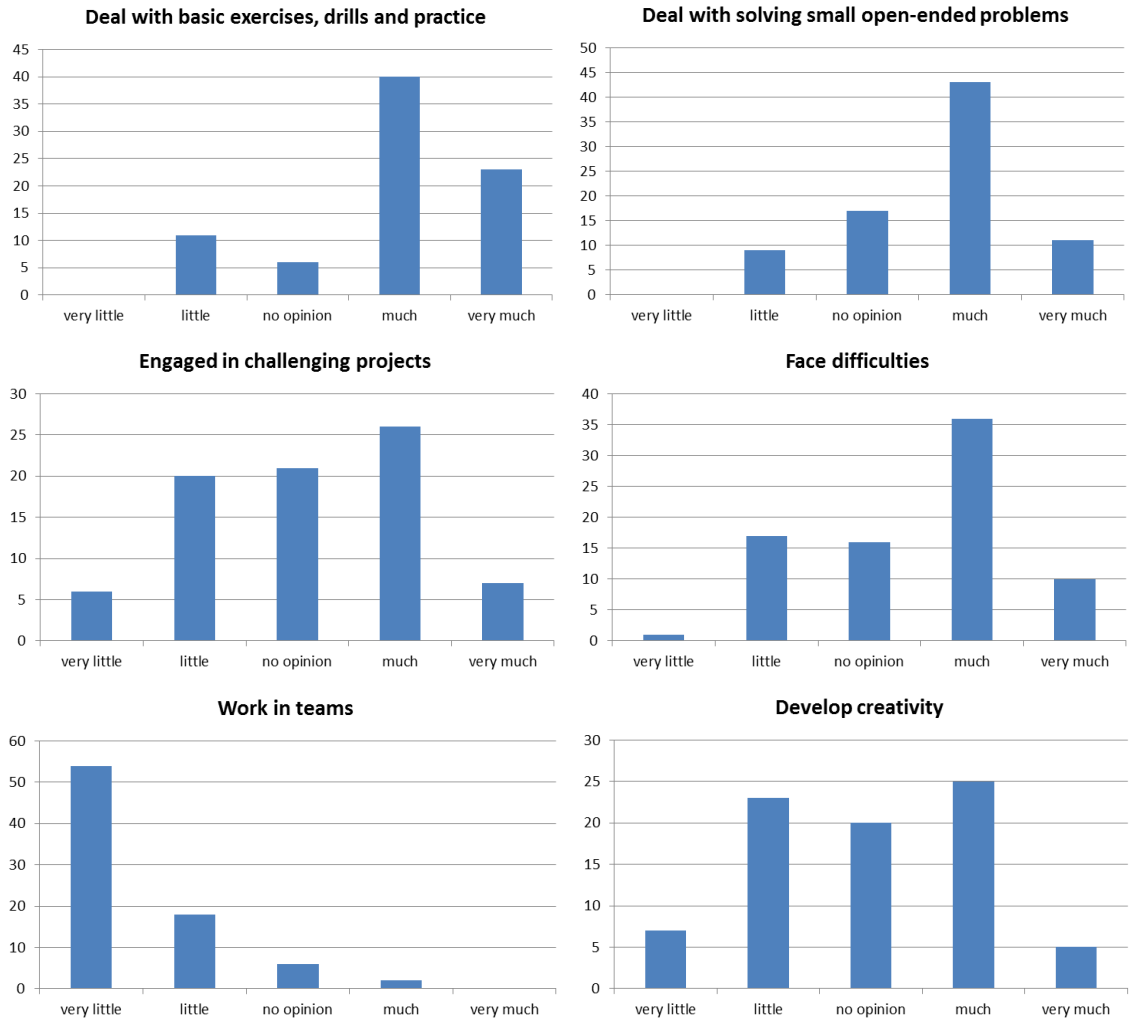


Figure 2. Students' responses about the Digital logic course at the University of Zagreb

IV. FINDINGS

The empirical evaluation was done at the University of Novi Sad and the University of Zagreb. At the University of Zagreb, the number of students who responded to the questionnaire was 80 for the Digital Logic and Computer Architecture 1 courses, 15 for the Computer Architecture 2 course, and eight for the Multimedia Architecture and Systems course (a total of 183 answers). At the University of Novi Sad, the number of students who responded to the questionnaire was 30 for the Digital System Design. Additionally, 13 teachers responded to the teachers' questionnaire in four course categories at the University of Novi Sad.

A. Students' Responses

This section presents the results of the students' questionnaires in five courses from both Universities which were selected for this evaluation. Selected courses are:

- Digital logic (Univ. Zagreb),
- Computer Architecture 1 (Univ. Zagreb),
- Computer architecture 2 (Univ. Zagreb),
- Multimedia architecture and systems (Univ. Zagreb),
- Digital system design (Univ. Novi Sad).

TABLE I. STUDENTS' RESPONSES ABOUT COMPUTER ARCHITECTURE COURSES AND MULTIMEDIA COURSE

No	Students' responses			
	Question	*Comp. Archit. 1*	*Comp. Archit. 2*	*Multimedia Arch. & Sys.*
1	Exercises	much	much	much
2	Problems	much	little	much
3	Projects	little	no opinion	little
4	Difficulties	much	much	much
5	Teamwork	very little	very little	very little
6	Creativity	much	little	no opinion

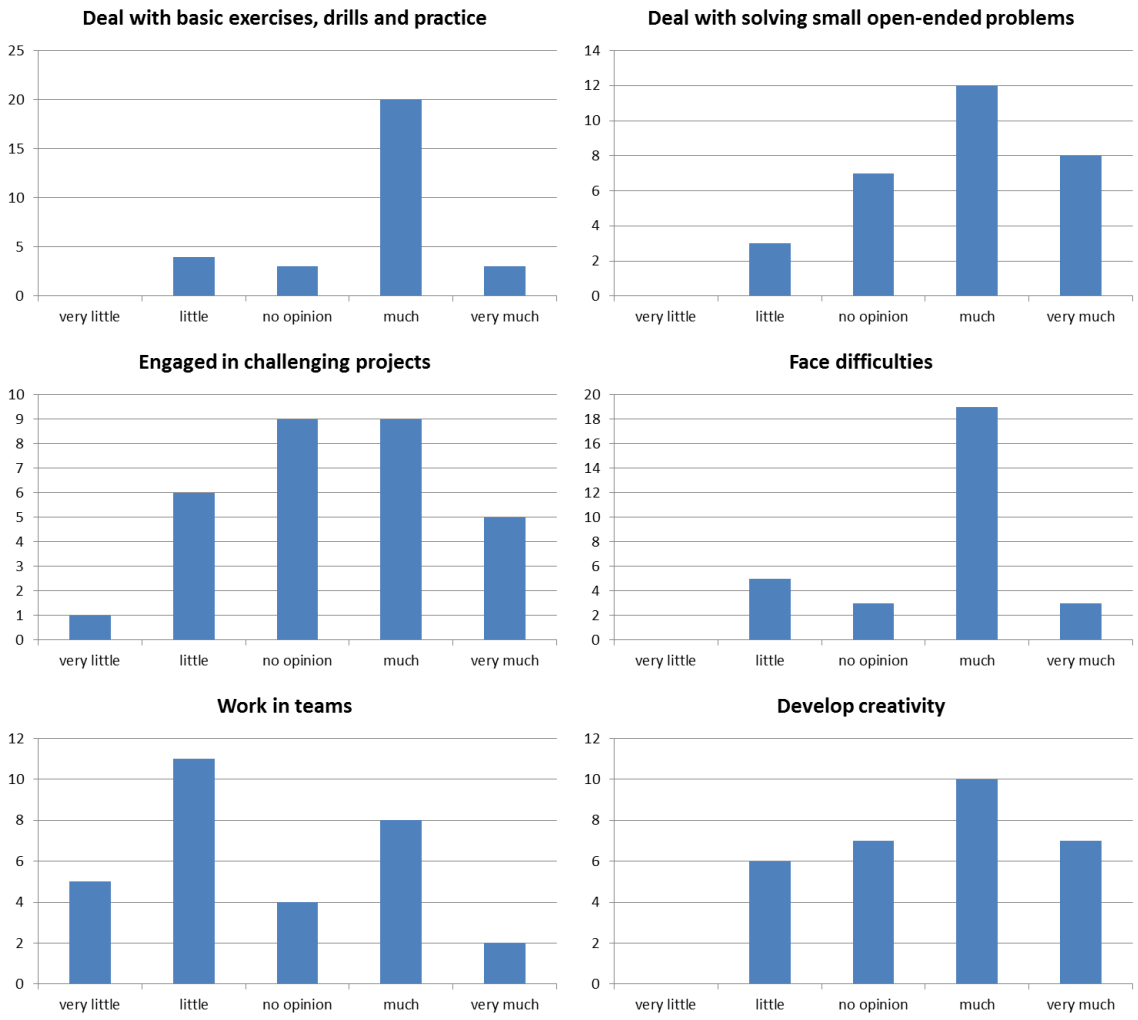


Figure 3. Students' responses about the Digital system design course at the University of Novi Sad

Fig. 2 and Fig. 3 present the distribution of answers for the courses in design of digital systems in University of Zagreb and University of Novi Sad respectively. The answers for the remaining courses are summarized in table 1 which shows the answers which received the most votes for each question.

B. Teachers' Responses

This section presents the results of the teachers' questionnaires in four courses from University of Novi Sad, which were selected for this evaluation. Selected courses are:

- Digital system design,
- Computer system design,
- Digital signal processing (DSP),
- Computer networks.

Fig. 4 illustrates the distribution of answers for the digital system design course at the University of Novi Sad. The answers for the remaining courses are summarized in table 2 which shows the answers which received the most votes for each question.

TABLE II. TEACHERS' RESPONSES ABOUT COMPUTER SYSTEMS, DSP AND COMPUTER NETWORK COURSES

No	Students' responses			
	Question	***Comp. Systems***	***DSP***	***Comp. Networks***
1	Exercises	much	much	much
2	Problems	much	much	much
3	Projects	little	little	little
4	Difficulties	much	little	much
5	Teamwork	much	little	much
6	Creativity	much	little	much

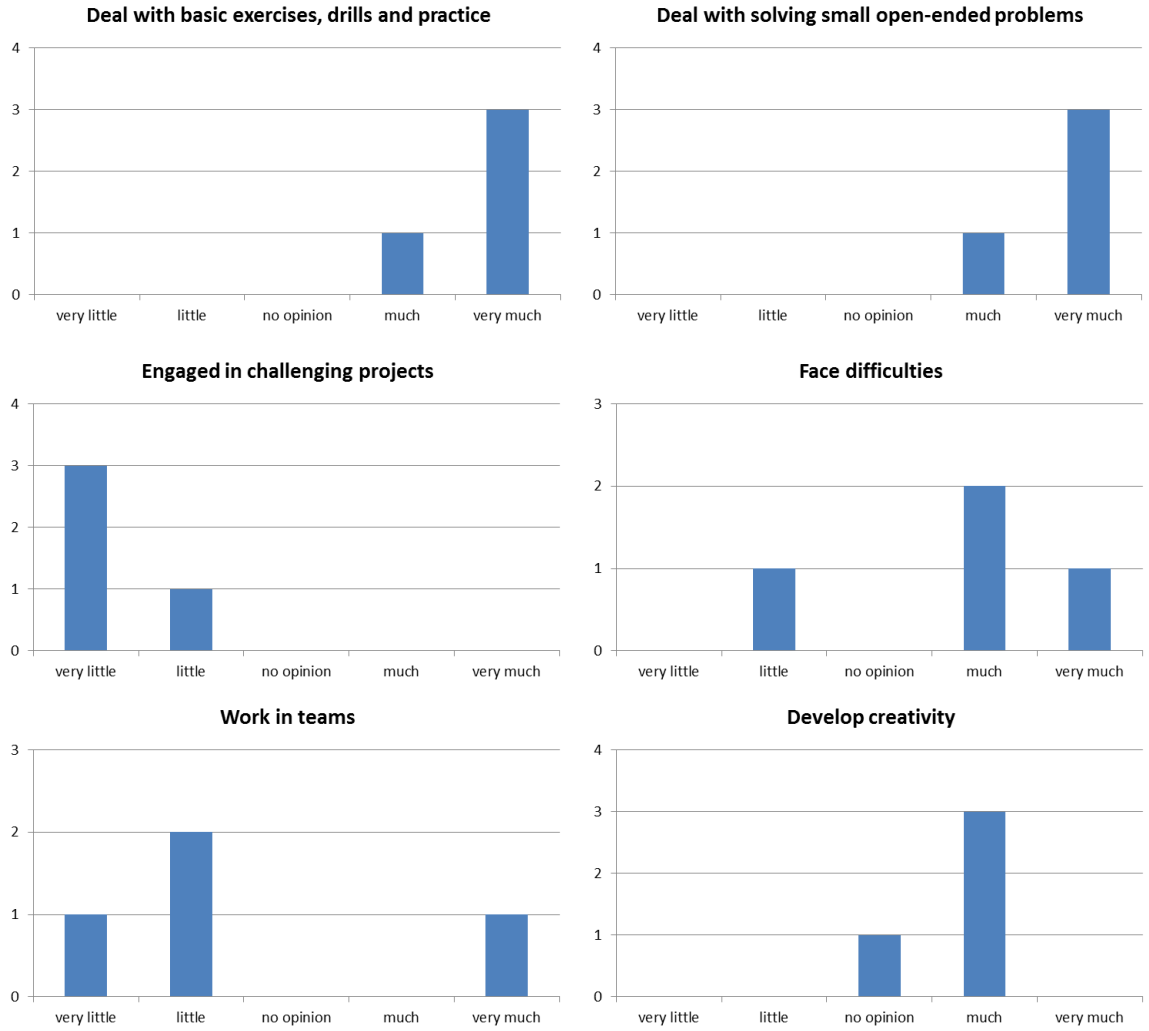


Figure 4. Teachers' responses about the Digital system design course at the University of Novi Sad

C. *Summary of the Findings*

The results from the University of Zagreb, Faculty of Electrical Engineering and Computing and University of Novi Sad, Faculty of Technical Sciences, are summarized below.

To summarize the students' answers, the percentage of students who marked 'Very much' or 'Much' for the six items for each of the four courses was calculated. The outcomes are presented in Figure 5.

It may be seen in Fig. 5 that 71% of the students marked 'Very much' or 'Much' for the item "We deal with basic exercises, drills and practice," which are basic assignments in the taxonomy (scale) we presented earlier in this paper; 61.7% marked that they "deal with solving open-ended small-scale problems or design tasks to meet given specifications and constraints," which are mid-level assignments in the taxonomy; only 31.1% agreed that they "are engaged in challenging projects" which are high-level assignments in the taxonomy; 44.2% marked that they "face difficulties"; only 3.3% answered that they "work in teams;" and 26.2% answered 'Very much' or 'Much' for the expression "we develop creativity."

To summarize the teachers' answers, the percentage of teachers who marked 'Very much' or 'Much' for the six items for each of the four courses was calculated. The outcomes are presented in Figure 6.

It may be seen in Fig. 6 that 100% of the teachers marked 'Very much' or 'Much' for the item "The students deal with basic exercises, drills and practice," which are basic assignments in the taxonomy (scale) we presented earlier in this document; 92.3% marked that the students "deal with solving open-ended small-scale problems or design tasks to meet given specifications and constraints," which are mid-level assignments in the taxonomy; however 0% think that the students "are engaged in challenging projects."

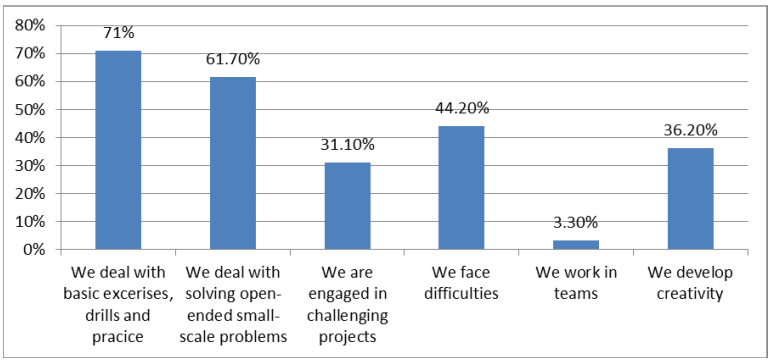


Figure 5. The percentage of students who marked 'Very much' or 'Much' to the six items for each of the five courses in computer engineering

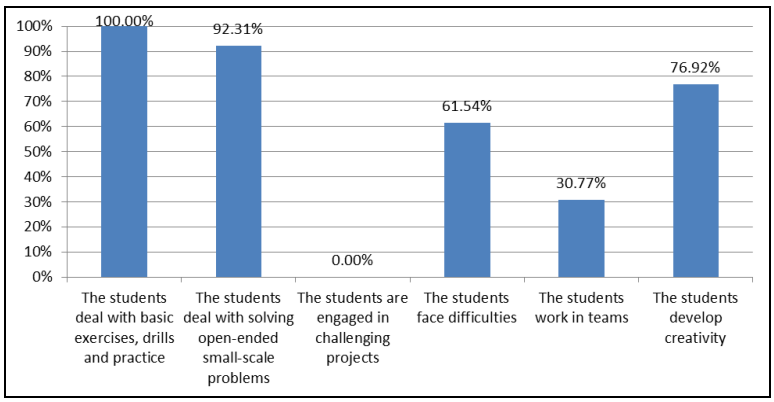


Figure 6. The percentage of teachers who marked 'Very much' or 'Much' to the six items for each of the four courses in computer engineering

In addition, 61.5% marked that the students "face difficulties"; 30.7% answered that the students "work in teams;" and 76.9% answered 'Very much' or 'Much' for the expression "the students develop creativity."

V. CONCLUSIONS

This paper presents an approach to evaluation of embedded engineering study programs and the results of applying the proposed method in evaluation of the embedded systems engineering curriculum across two Universities.

The results indicate that in learning computer engineering, the students deal merely with doing basic exercises and solving simple problems. Much work is required to shift the teaching and learning of embedded engineering and computer science towards enhancing students' higher-order cognitive skills such as problem solving and creativity, and fostering teamwork in the engineering class.

First research results also suggest that there is a large overhead in the number of platforms used in the mentioned courses. In the future, a unified platform will be developed for these and other courses in the curriculum. This platform will extend the previous work [8]-[9] which was used only in courses in digital design. The aim is to reduce the time lost in adapting the students to different platforms in different courses and increase the opportunities for engaging students with more challenging and more creative projects.

REFERENCES

[1] J. Piaget, Judgement and Reasoning in the Child, London: Routledge & Kegan Paul, 1969.

[2] L. S. Vygotsky, Mind in Society: The development of Higher Psychological Processes. Cambridge. MA: Harvard University Press, 1978.

[3] J. W. Thomas, A review of Research on Project-based Learning, San Rafael, CA: Autodesk Foundation, 2000.

[4] B. S. Bloom, and D. R. Krathwohl, Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook 1: Cognitive Domain, New York, Longmans, 1956.

[5] L. W. Anderson, and D. R. Krathwohl, A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives: New York, Longman, 2001.

[6] H. L. Plants, R. K. Dean, J. T. Sears, and W. S. Venable, "A taxonomy of roblem-solving activities and its implications for teaching", in J. L. Lubkin, (Ed.), The Teaching of Elementary Problem-Solving in Engineering and Related Fields. Washington DC, American Society for Engineering Education, 1980, pp. 21-34.

[7] M. Barak, "Teaching engineering and technology: cognitive, knowledge and problem-solving taxonomies", *Journal of Engineering, Design and Technology,* 2012, in press.

[8] D. Majstorovic, E. Neborovski, M. Katona: "A cross-curriculum embedded engineering learning platform", *Proceedings of the 33rd International Convention MIPRO*, 2010, pp. 1039-1044

[9] I. Kastelan, D. Majstorovic, M. Nikolic, J. Eremic, M. Katona: "Laboratory Exercises for Embedded Engineering Learning Platform", *Proceedings of the 35th International Convention MIPRO*, 2012, pp. 1113-1117

BYOD: a challenge for the future digital generation

Ivan Pogarčić, Maja Gligora Marković and Vlatka Davidović
Polytechnic of Rijeka/Business Department, Rijeka, Croatia
{pogarcic, mgligoram, vdavid}@veleri.hr

Abstract - Recently trends in methodology and technology of development of information systems strictly indicate the frames known as cloud computing. Information system located in cloud provides his user with certain advantages, but is also represents a certain challenge. It also requires an increased attention and additional care about the system's safety. Problems are of the general character, meaning they are equal regardless to the business function which information system is being developed. Though several decades have passed from the original efforts of implementing the ICT in educational frames, the real treatment of activities known as eLearning require additional considerations in education. eLearning should be treated as information system with all its features and additional specificities. Independence of place and time of realisation provides an additional possibility. This refers to the possibility of integrating the personal user's equipment in eLearning. Trend in practical information science, referred as BYOD trend, presumes that final users approach to business resources with help of their personal devices: smart phones, tablets and laptops. This paper examines the readiness of student and teachers to adopt such teaching modes.

I. INTRODUCTION

Is it possible for an individual to find instructions explaining how to prepare himself for education, how to find a job or how to efficiently perform it? How much is one's choice of a career influenced by circumstances a person isn't even aware of? Choice of a profession is usually, regardless to aptitudes and acquired education, a matter of moment and circumstances on which individual cannot or doesn't want to influence. However, everybody tries to carry out his business tasks qualitatively, especially if there are possibilities of promotion, appreciation and reward.

Richard Templar emphasizes two things: firstly, qualitatively performed tasks improve organisation/firm and society as a whole. This is achieved by personal intercession which should result in appropriate compensation through which an individual improves his condition and possibilities, and consequently his broader and wider environment. [1] The other side of each business is a bit harder and unspecific since it refers to organisation of business. This is an instance where true problems emerge.

The subject of this paper and research doesn't refer to the history of labour or labour relations, or to the political treatment of labour and labour relations. Still, both history of labour as a phenomenon and labour relations obtrude as unavoidable facts. Another trend has become a cause of this research which authors consider a relevant matter in all aspects of treating the labour as organised human activity. Due to this fact, BYOD (Bring Your Own Device) is being proclaimed mainly within frames of ICT technology, without special underlying of other professions and business activities. The authors represent an opinion of an aggregate involvement of ICT tools and appliances in all human activities. This however doesn't mean such paradigm should be restricted exclusively to ICT.

BYOD can still be considered as a result of development, theory and especially practice within scientific frames which subject of interest are labour and its organisation. This is especially notable with business tasks immanent to information systems. Since the beginning of ICT application as backup to information system, specific professions and certain related tasks have been developed. However, with development and improvement of ICT within hardware and software area, those professions necessarily changed and adjusted. Aside from all the other influences, finances and prices of needed equipment lead to considering IT department as a necessity in business systems, since it wasn't a core business itself. At that time IT has been considered as outsource reality. For information scientists that was considered as "a paradise exile", while information system became an unavoidable necessity in any business, or "the Eastern Sin".

Outsource of IT department as an organisation unit of a business entity brings certain advantages, as well the disadvantages. Further on, the recent paradigms in organisation of information tasks, such as SOA and Cloud Computing, indicate how expensive equipment and applicative solutions can be used as rented resources as well, so significant savings can be achieved in business[2],[3]. A logic continuance of such thinking was BYOD. However, as with other circumstances which accompany information tasks through history, BYOD can also be connected to specific uncertainties in its application.

The original hypotheses of this research indicate that every human activity can be organised in a way so that active participants use their own work appliances, but they only represent one element of the organised work. The second assumption concretises education as the activity in which ICT has increasingly important role, especially in eLearning as the realisation form. [4]

II. BYOD, BYOPC OR SOMETHING ELSE

BYOD – bring your own device is a paradigm famous by its abbreviations BYOT (T-technology), BYOP (P-phone) and BYOPC (PC-personal computer) which describes a possibility for employees to bring along their personal appliances with the mobile possibilities.

The literature usually mentions mobile appliances such as laptops, tablets and smart phones which can be used at workplace. Implicitly, it means that workplace itself can be independent of time and location. Exactly these two characteristics describe eLearning as activity. It also means that both teachers and students can have their own equipment during the educational process. Therefore, BYOD is usually mentioned when describing the practice in which students use their own appliances in education. Let us accept the working definition of BYOD within educational frames in order to understand some attitudes described in this research. It quotes:

Bring your own technology (BYOT) is an educational development and a supplementary school technology resourcing model where the home and the school collaborate in arranging for the young's 24/7/365 use their own digital technology/ies to be extended into the classroom to assist their teaching and learning and the organisation of their schooling and where relevant the complementary education outside the classroom. [5]

Nevertheless, BYOD is equally interesting in working conditions of system's employees where such opportunities exist. Naturally, if there is a will and an agreement to practice such a relation. Within TechRepublic site discussion "„Ten business trends for 2013 - and four ways they affect you", Peter Cochrane, among other, concludes "…IT departments will continue to decline as they are seen as increasingly irrelevant with more capable workforces making a unilateral declaration of independence, adopting BYOD, and striking out as independent workers." and „BYOD will become the default mode with the days of companies being dominant providers of all IT equipment and support begin to decline.."[6]

BYOD trend is a challenge, regardless to hurry and confusion in its application. Advantages of its application are truly enormous, especially in a financial sense since it significantly decreases a price of conducting the business. Employees will use their mobile appliances for different purposes. However, a thorough consideration implies important problems which should be solved or at least considered due to their intensity. Since the employee can constantly apply firm's resources through his appliances, one can expect increased interest in communication. On the other hand, working hours becomes a relative term since the real boundary between efficient work and relaxation becomes invisible. Contents memorised on appliances which refer to firm's data are saved at the same appliance, so there is a threat of mixing them up or invading someone's privacy. Consequently, a problem of management occurs from the aspect of control, updating, report analysis without influence on employee's privacy or property's integrity.

It becomes clear that IT business crystallizes new mode of business/profession - Mobile Application Management (MAM). IT experts should prepare themselves for solutions and modes of managing applications and data installed and saved on a variety of appliances. In same manner [7], they define MAM as a capability of managing applications on devices distanced from the central location. Applications are saved in safe containers from which data can be downloaded, modified and saved according to need and certain rules. Safety is determined by set of rules such as Single Sign On and LDAP check-up of authenticity in accordance to MAM.

Some of MAM functions are: safety management, updating, delivery, general administrating on server position, authentication, authorization, synchronisation, check of the actual version and alike. In the process, a qualitative MAM politics is based upon following principles [7]:

- User manages the applications instead of hardware or software, with emphasis put on data and applications which belong to a company.
- Management is organised according to execution of politics, rules and roles allocated to employees with a possibility of adopting, changing and retrieving those rules.
- Managing the cooperation with the possibility of using the "app stores", where IT administrators can advice and control download of the recommended application from "App Store". This possibility should be regulated by the prior defined rules and needs of BYOD environment and it should ensure integrity of firm's data.
- Actuation of motto "Configure once, apply everywhere".
- Robustness and easy approach from any position – complete visibility and transparency that corresponds to complete insight into employees' activities.

Naturally, the principles will be submissive to certain circumstances which derive from specificities of business activities and certain circumstances in which the employee may work. In other words, within eLearning as a business or activity that can be placed within BYOD environment, described principles will be different for different sorts of learning contents. Respectively, regulation and management will be submissive to teaching goals and content. BYOD can be assumed within these frames as part of the context in which eLearning is being performed.

Mobile Device Management (MDM) is a function which refers to control and management, implemented by IT department and specialists, through software for controlling the functioning of devices, such as cameras, GPS, distanced appliance for saving data such as smart phones, tablets and personal computers. [7]

III. RESEARCH

Further on, authors will describe procedures applied in research, used instruments and sample of examinees who participated in research.

.

A. *Procedure*

For the purpose of the research, a method of questionnaires has been used so to collect data on opinions of examinees/final users regarding their promptness to approach business resources through their personal appliances: smart phones, tablets and laptops. For purpose of data collecting, an instrument has been prepared in shape of online questionnaire, in combination with service of portal innovation.eu for creation of online questionnaires supported by tool LimeSurvey. Questionnaire has been sent primarily via e-mail to employees of Polytechnic of Rijeka and students of Business Department at the same HEI. Examinees have also forwarded it through the social network – Facebook. Together with link to online questionnaire examinees have been provided with several sentences explaining the research and its purpose. The advantage of conducting the research via Internet has been used since it provides fast, economically and simple distribution of questionnaires to the members of targeted sample of examinees. Research was carried out in period November 2012 – January 2013.

B. *Instrument*

Questions in pool have been prepared according to fore analysed researches, in accordance to described principles of MAM/BYOD paradigm. Special attention was directed to coverage of questionnaire so to detain the interest of examinees during their questioning, but also to questions so to keep them clear and concise. Questionnaire included 21 questions, of which 5 were connected to the analysis of demographic structure of examinees, while the rest of them referred to users' readiness for BYOD (closed-type questions) and to opinions' analysis (one opened-type question). Closed-type questions related to claims in which examinees expressed their attitudes. For analysis of collected data authors used methods of descriptive analysis and graphic design which enables data layout in a comprehensible way.

C. *Sample of examinees*

Research included 178 examinees. The first group of questions in questionnaire was intended to analyse demographic characteristics while the second set of questions referred to BYOD related research. Out of total number of examinees, majority made men, totally 54.49%, while there were 44.31% women, and 1.20% didn't reply. The majority of examined, 47 have between 31 – 40 years (28.14%). Also, the majority, or 99 examinees are full-time employed (59.28%), while only 19 examinees are unemployed. Most examinees have higher education (47.31%), while 35.33% have secondary education, and there is only 5.39% Ph.D.'s.

IV. RESEARCH RESULTS

In continuance one can find the research results according to sequence of questions mentioned in questionnaire. The first question referred to firms in which employees work. Most of them work in companies which are not their own property – 123 examinees (73.65%), the least of them working in personal companies – 7 examinees (4.19%). The rest of examinees, 32 of them, work in public institutions (19.16%). When questioned about mobility of their work, 52 examinees (31.14%) answered they can perform their work independently of time, while the same ratio usually does perform their job independently of time (21.56%). 20.36% examinees perform their work in a mobile mode. 16.77% examinees do not perform work independently, while 8.98% does it rarely. When questioned upon sort of equipment needed to perform work, 133 examinees (79.64%) requires laptop, tablet, smart phone and alike, while only on occasion they use specialised equipment, only 2 examinees (1.20%), while 14 participants use other equipment (8.38%). Most examinees own laptop and smart phone (81.46%) and least of them (2.81%) do not own any devices at all. Considering the ownership of the used equipment for performing business tasks, 77 examinees (46.67%) answered they use personal devices, 67 examinees (40.61%) use it occasionally, while only 21 examinees (12.73%) do not use personal equipment. Figure 1 indicates readiness to use personal equipment for business/educational purposes.

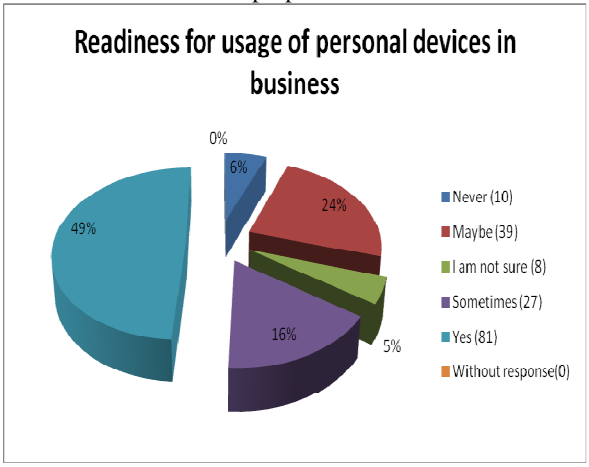


Figure 1 Usage of personal devices in business

Considering their readiness to install application which could control a part of hardware computer resources, 43.64% examinees answered it depends on their work and employer, 12.73% would never agree to it, 19.39% occasionally, while 13.94% are willing to do so. 66.85% answered they use Cloud Computing for personal purposes, while 33.15% use it in performing operational tasks.

Figure 2 displays answers to questions regarding their concern about safety of data saved at PC. Examinees usually consider that aspect of safety.

When questioned about their familiarity with safety politics of firm/HEI where they work/study, examinees provided following answers: 9 (5.45%) were never introduced, 34 (20.61%) knows only basics, while 10 (6.06%) answered yes and no, 45 (27.27%) knows thoroughly the politics, while 4 (2.43%) didn't answer.

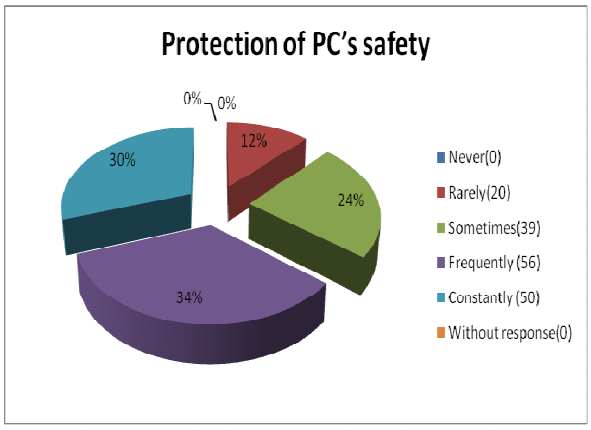


Figure 2 Aptitude to protection of PC's safety

Examinees know very little (30.30%) or just basics (39.39%) of Licence Agreement application rules which they apply in work, with majority (20%) which believes that precise boundaries should be defined so to consider the right over somebody's data. Most examinees believe there is a real threat of endangering privacy (29.70%) by foreign company or third parties if they use applications licensed by company/HEI, with equal portion of participants believing there is a real danger of privacy violation (15.76%), while 17.98% believe there is no such a danger.

Majority of examinees (79.39%) consider they themselves are mostly responsible for data safety, referring both to employees and employers. Figure 3 displays representation of individual characteristics related to BYOD.

A. *Opened-type opinions of examinees*

31 examinees (17.42%) answered the open-type question, while 147 (82.58%) didn't provide an answer. Here are several answers:

- "It is important to determine an expense and amortisation of personal devices used for business purposes, as a percentage of salary".
- "The main issue in accepting BYOD system refers to a fact whether employer is prepare to charge the employee amortisation of his personal device – that is, if I use my own laptop in business purposes, I also spend it, so my property loses its value".
- „Rational compensation in form of amortisation could significantly accelerate implementation of BYOD's system in business processes. BYOD becomes a procedure in which employer is less engaged in usage of core resources or fixed asset which is maximized by his claim how he simplifies users' choice to apply devices at their own pace"

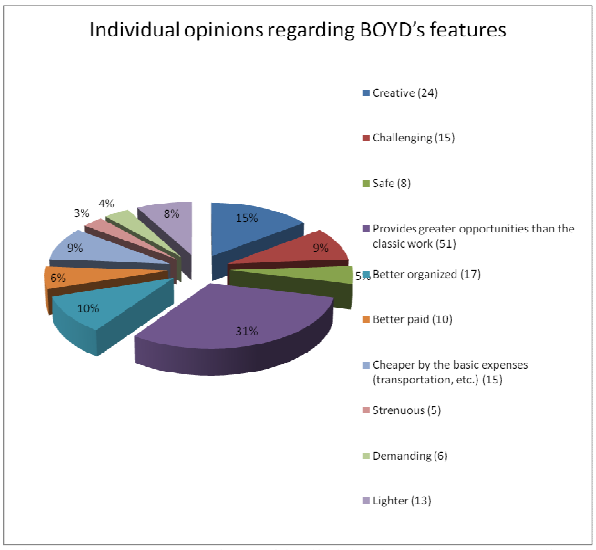


Figure 3 Representation of individual opinions regarding BOYD's features

- "Employer should, in case employee doesn't want to use his own equipment, buy him the needed equipment".
- "The usage of BYOD business mode should be aligned with direct human contact. Contacts between persons should not be neglected, no matter the type of business. Technology does represent a huge help, but only as a tool. When it becomes a purpose of its own, it threatens the business itself. Awareness of BYOD should be spread through media".
- „This is a trend that makes us accept new ways of living and working, while lagging behind means the impossibility of cooperation with the EU and World. We have to follow new trends since otherwise, in a long term range, we will lack behind even more".
- "I have nothing to add."
- "My dilemmas connected to this mode of work: 1. Is it only one way of making work time more flexible, so the employer can even more exploit employees in their spare time without paying them, and 2. which guarantees are provided to employee that his employer won't be given the private information of employee and his activities to which employer isn't entitled to."
- "If BYOD politics doesn't confront with safety of licence aspect of firm/organisation, then BYOD will quickly become a trivial item since appliances used in IT business are highly personalised and adjusted to our personal preferences. That means it is perfectly normal to bring along our own devices, as it is normal to bring own (normally personalised) underwear."

- "Maybe it would be productive to add in questionnaire a question related to compensation company provides in case of damaging employee's personal equipment. For instance, if employee's laptop is damaged, is the employer willing to pay for repair or even to buy a new one".
- "Rules should be defined, especially those related to working hours. Does one have a more flexible choice of daily working hours if he uses own devices?"

V. PROFILE OF BYOD EMPLOYEE

Obtained data helped to create a frame or profile of BYOD employee in some fictive company which is prepared to organise its business in a way to offer such labour relations and terms.

Of totally 178 examinees, more than a half was male (54.49%). This information can serve as a good basis for analysis of connection between gender and readiness to get employment in concordance with BYOD paradigm. Average age of examinees is between 31-40 years with left asymmetry due to certain number of examinees who still study, though such information isn't explicitly registered since the question itself hasn't been posed. Educational level is academic with high 47.31% of examinees. High level is registered also with number of fully employed examinees (59.28%) so that the evaluation is credible. This is also confirmed by a fact that 73.65% examinees are owners or co-owners so they can act within BYOD frames. Related to their opinions on performing work within mobile frames, the answers were uniformly distributed from positive to negative in a five-level raster. This profile of BYOD employee believes ICT equipment to be most suitable for such organisation of work. 46% examinees understand BYOD frames, 81.46% owns laptops which in 47% have been used for education or performing business. 49.09% is ready to put their equipment in function if they are asked to do so, but only 13.94% would unconditionally agree for employer to install his software in employees' computers.

Two thirds of examinees use Cloud Computing services of different types and take care of their own computer's safety. However, they do not recognise a term LA (Licence Agreement). 29.70% consider there is, within BYOD organisation, a relative possibility of data and equipment endangerment, while responsibility over data safety should be shared with company's owner.

Table I. displays a hierarchical summary of the offered ten features of BYOD organisation. According to hierarchy, 31.10% examinees believe BYOD organisation can offer broader possibilities in work, but it is interesting to note how opinion upon creating a business in such mode would be easier or the same is placed on the lowest level of importance. Other interesting data were bolded.

TABLE 1. HIERARCHY OF FEATURES IN BYOD ORGANISATION OF BUSINESS

	Ranking									
	I	II	III	IV	V	VI	VII	VIII	IX	X
Creative (A1)	14,63%	12,20%	**16,46%**	12,27%	15,34%	4,91%	10,43%	8,59%	3,07%	2,45%
Challenging (A5)	9,15%	12,80%	11,59%	**15,34%**	13,50%	14,11%	11,66%	7,36%	1,23%	3,07%
Safe (A4)	4,88%	6,71%	9,15%	6,13%	7,98%	10,43%	**17,79%**	14,72%	12,88%	9,20%
Provides greater opportunities than the classic work (A3)	**31,10%**	17,68%	15,24%	6,75%	8,59%	7,36%	3,68%	4,91%	3,07%	1,84%
Better organized (A2)	10,37%	**17,68%**	10,37%	16,56%	13,50%	12,27%	9,20%	4,91%	3,07%	1,84%
Better paid (A8)	6,10%	6,71%	6,71%	4,91%	9,82%	**16,56%**	10,43%	15,95%	11,66%	11,04%
Cheaper by the basic expenses (transportation, etc.) (A7)	9,15%	**18,29%**	9,76%	17,79%	11,04%	10,43%	10,43%	7,36%	3,68%	1,84%
Strenuous (A6)	3,05%	2,44%	3,66%	5,52%	7,98%	3,07%	9,20%	16,56%	22,70%	**25,77%**
Demanding (10)	3,66%	3,66%	7,32%	9,20%	5,52%	9,20%	6,75%	7,98%	**28,22%**	19,02%
Lighter (A9)	7,93%	1,83%	9,76%	5,52%	6,75%	11,66%	10,43%	11,66%	10,43%	**23,93%**

VI. CONCLUSION

Fast development of information science in all its fields has been accompanied by frequent modifications of paradigm used in information systems development. Specificities of connections between business and information systems are necessary consequence of these modifications, especially due to influences of mutual changes in a mode of business' organisation.

BYOD frame isn't dependant of business itself in the same measure that business can be independent in any segment or form of place and time of realisation. Since eLearning is proclaimed as location and time independent, BYOD organisation offers additional possibilities. Primarily this could significantly cheaper educational process which is the strongest recommendation for such organisation. Further researches could pay attentions to specific forms of business such as eLearning or Cloud Computing.

This research hasn't insisted on specific sort of business or suggested any specificities of business. This supposed to emphasize the generic possibilities of BYOD paradigm.

LITERATURE

[1] R. Templar, The Rules of Work, Expanded Edition: A Definitive Code for Personal Success, FT Press, 2010, ISBN-13: 978-0137072064

[2] T. Erl, Service-Oriented Architecture (SOA): Concepts, Technology, and Design, Prentice Hall, 2005, ISBN-13: 978-0131858589

[3] K. Jamsa, Cloud Computing, Jones & Bartlett Learning, 2012, ISBN-13: 978-1449647391

[4] G. Kumar, A. Chelikani, Analysis of Security Issues in Cloud Based E-Learning: A Qualitative Study, LAP LAMBERT Academic Publishing, 2012, ISBN-13: 978-3848407354

[5] M. Lee, BYOT downloaded from http://www.scoop.it/t/byod-how-and-why/p/ 2339528830 /byot-mal-lee on 16.1.2013.

[6] P.Cochrane, Ten business trends for 2013 - and four ways they affect you, CIO Insights, Techrepublic downloaded from http://www. techrepublic.com/blog/cio-insights/ten-business-trends-for-2013-and-four-ways-they-affect-you on 16.1.2013.

[7] C. Knight, Best Practices For Meeting The BYOD Challenge, Analyst Report Offers Insights, downloaded from http://www.kony.com/ resources/white-papers on: 16.1.2013.

ICT didactics: The new study discipline is needed

Marjan Krašna, Bojan Bedrač
University of Maribor, Faculty of Arts, Department of Pedagogy, Maribor, Slovenia
marjan.krasna@um.si; bojan.bedrac@um.si

Abstract: The educational technology is used at all levels of education. In the USA educational technology studies are already established. In German speaking countries the term "didactics of informatics" is used to cover ICT topics. Other countries rather use "didactics of computer science" to describe the special didactics discipline. It is evident that teachers of informatics or teachers of computer science require different study courses than other teachers despite the fact that all teachers require ICT knowledge. We have analyzed different approaches to the increased demand of required teachers' ICT knowledge. Results reveal the conclusion that ICT in education should become a new discipline since it cannot be covered with the traditional special didactics disciplines any more. Habilitation committee at the University of Maribor agreed that didactics of ICT should become a special didactics discipline and should be an integral part of LLL (Life-long learning). At the Faculty of Arts in The Department of Pedagogy a Center for Information, Communication and Interactive Technologies (CICIT) was established to formally recognize the importance of the discipline. CICIT is to prepare courses for didactics of ICT. The article shows the results of comparative study for the new discipline and the proposed objectives, scope and constraints of ICT didactics discipline.

Keywords

I. INTRODUCTION

Educational technology is an important addition in the TLS (Teaching Study Learning) process. It enables different types of communication and satisfies preferences from different types of people. Educational technology changes in accordance with the advancement of ICT and in the last decade this process was very rapid. Education processes has benefit great with the introduction of electronic equipment but the ICT actually transforms the education. Multimedia learning materials at all levels of education become a reality [1]. We now know that teachers are not the authoritative "knowledgebase on two feet" that transfer knowledge to the students. Each individual with the internet connection and a terminal (pc, tablet or smartphone) is capable of accessing the knowledge which even exceeds the teachers' capabilities. The purpose and paradigm of education therefore changes [2]. It is still possible to get responses from moral authorities of the education that some knowledge still needs to be embedded into the minds of people because of their fear "What will happens if electricity fails or individual in focus do not have an internet connection anymore?" But an increasing number of younger ICT literate educators have a different opinion and they cannot imagine teaching traditionally.

Without multimedia a modern teaching is unimaginable. It has an educational, cultural, technical, and social function. Therefore multimedia elements are selected based on educational goals, acquired knowledge, and psychophysical characteristics of students in one hand but also according to the nature of media, school's equipment and teacher's competences. [3]

It becomes a standard that each classroom has at least one computer and a projector. Traditional blackboard (the green one) and whiteboard are less frequent. Chalk dust and fumes of markers are replaced with the electronic ink on interactive boards. Changes of content become simple and give new dimension into education space. "Undo" was once upon a time just wishful thinking. Electronic materials can be copied immediately and sent to students e-mail or e-classroom for later access. Teachers' work becomes complex and requires more study. Teachers should become producers of e-learning materials and digital competences should be their second nature [4] [5]. But students' demands increase also and aesthetic principles of multimedia learning materials rise in their eyes [6].

Technology is available and only "economic crisis excuse" enables governments to implement austerity measures to slow down or stop the flow of new technology into the classrooms. The effect is actually short-lived and counterproductive since kids have access to the technology. In the developed society where education is mandatory parents give their kids priority over some other spending. Rarely teachers see kids with electronic gadgets as old as their own. It is almost unimaginable that the city kids would not have smart phones yet many teachers still don't have them and they don't know how to use them. This gap between generations should require Life Long Learning approach to the educational technology. Even computer science teachers and computer support employees have a hard time keeping up with the advancement of technology. In general teachers slowly lose in this battle with technology. Recognition of this fact should be taken into account and rectified.

It takes a decade to recognize and formalize the area of ICT as equal to other special didactics [7]. Despite the fact

that didactics of ICT is not a discipline that would be taught in primary and secondary schools it is mandatory for teachers training in the tertiary levels of education. It is interesting that every teacher uses it (more or less successful) but it is not a part of the curricula. In the battle for ECTS these topics lose and regrettably designers often refer to as something that students already know so we do not need to teach them [8]. This deniability gives them more ECTS for professional topics. In the student research of suitability of curricula (topics and outcomes) in the area of ICT among courses an astonishing discovery emerged. Knowledge of ICT, computer literacy and connected topics were mostly addressed in outcomes and not in the topics. These discrepancies are shown in all teachers training programs in all tree universities in Slovenia (Student project 2012; course: Computer application in pedagogy).

In 2012 the Faculty of Arts signed a contract with the interactive board producer Smart. In exchange for equipment we would have to prepare a training facility and conduct education for Smart educational technologies. What was once hidden from our eyes becomes visible and we received equipment that was already in primary and secondary schools. Our students in pedagogical study programs were increasingly discontent with the gap between the topics that were thought in the faculty and the reality that they face in the classrooms in schools. This event and realization that lecturers are unable to use this technology gives a green light to the establishment of the Center for Information Communications and Interactive Technologies (CICIT)

II. Analysis of other educational institutions

The analysis of neighboring disciplines was done with the web research of educational institutions. In our analysis we know that complete analysis is not possible but we want to find as many relevant data to make some credible assumptions. In the first step we wanted to make a list of initial keywords that should be searched and in the progress we include more and more keywords. (*Computer in education, Technology enhanced learning, Educational technology, didaktik der informatik, didactique d'informatique, pédagogique d'informatique, l'ordinateur en l'éducation, éducation assistée par ordinateur, didactique des TIC.*)

For the "computer in education" we have searched through 20 educational institutions and portals. Educational technology was researched in 15 education institutions available on their official web sites.

Results of the analysis show the following facts:

- In the English speaking area (especially in USA) the neighboring discipline is known as Educational technology and is thought on all levels of education. Graduate and postgraduate studies are available and outcomes are (B.Sc, M.Sc, Ph.D and Edu. Dr.).
- The German speaking countries have the same ICT topics covered in the term of didactics of informatics (*ger. didaktik der informatik – DDI*). Didactics of informatics is though in almost all German universities. After a brief search we have discovered a portal for DDI. From the available data we have discovered that DDI is taught in 28 German speaking universities in Germany, Austria and Switzerland.
- Much more problems were encountered in the French speaking countries despite the fact that all universities have study programs of ICT in education. After a disappointing research we contacted an expert for French educational system and found out that we could not find uniform information since each university is a world of its own. We have discovered that in the French language for didactics of ICT could be found with keywords: *didactique d'informatique, pédagogique d'informatique, l'ordinateur en l'éducation, éducation assistée par ordinateur, didactique des TIC*.

III. Establishment of CICIT – Center for Information Communication and Interactive Technologies

Center for Information Communication and Interactive Technologies was established in 2012 at the Faculty of Arts at the University of Maribor. It is a new center with a long time tradition. Conditions were right to make a center that would connect all individuals who worked in the area of ICT, application of ICT and educational technology.

The aims of the CICIT are:

- Research the didactical strategies of ICT and monitor development of educational technologies.
- Testing equipment and prepare the report of didactical suitability of specific technology.
- Become a central competence point for users and researches of ICT in education.

A. Reasons for CICIT

The background for the CICIT was established at the old Faculty of Education where studies for application of educational technologies began. We were the first to implement videoconferences in education and teaching practice monitoring. Pioneering work for preparing e-learning materials and multimedia learning materials were later enhanced to projects funded by the government. At the Faculty of Education the multimedia course for pedagogical students was mandatory and our graduates were more competent and more desirable to their future employers.

In 2006 the Faculty of Education transformed to three faculties (Faculty of Arts, Faculty of Education, and Faculty of Natural Sciences and Mathematics) and experts that were previously together were reassigned to the newly formed faculties. Inevitably the critical mass dropped and projects were halted. In just a few years the drawbacks and stagnation was showing at all faculties. With the CICIT a formal frame is set and we can reestablish the groups of experts to advance the ICT in didactical strategies.

Changes in student populations show that students want new approaches in the TLS (Teaching Study Learning) processes. Contemporary students do not make notes on lectures. They demand e-learning materials that are complete, concise, useful and technically proficient. Their requirements are far beyond the competences of many todays lecturers. Advice and support service is therefore needed.

Learning Management Systems are established and used by students and ICT literate lecturers. There are some lecturers we know that will never use LMS in their work but the time will change that too. At the beginning LMS were used for communication and data distribution purpose but in the last time more and more they are used for electronic testing.

Interactive technologies are used in primary and secondary schools but not in the faculties. Lack of funds and management decisions often bleed the courses that need new technology. Only when feedback from lower levels of education comes the sobering enlightenment shows we are obsolete. This bad impression cannot be corrected only with the purchase of equipment we also need some time to learn how to use it correctly and efficiently.

B. CICIT area of operation

In the beginning the CICIT is set to operate in the following areas:

- Educational technology
- Distance education
- Information support for education and
- Audio and video in education

Education technology becomes elaborate and in the USA reaches such levels of expertise that it is considered an independent scientific discipline. It is possible to achieve even a Ph.D. in the area of educational technology. In our environment we focus to the presentation technique and interactive technology in the classroom.

Distance education covers LMS and multimedia learning materials. Functions of communication, distribution of learning materials, students' supervision, and knowledge testing are covered in LMS. Creation of multimedia learning materials requires software authoring tools. Since even tools are changing fast it needs to be monitored and used appropriately.

Information support for education covers services and software. Videoconference lectures require support in the preparation and operation phases. Project oriented teaching requires project management services. Even web 2.0 technologies do not just work. All these topics require advice and support which is going to be provided by CICIT

Audio and video in education are not new. Only with fast computers and cheap camcorders it becomes accessible to a larger audience. But we need to know that not every video and/or audio is suitable for education. A quick look to YouTube shows that most videos can be categorized as digital junk. Suitable high quality videos are needed in education. Process from recording to editing and finally distribution requires expert knowledge and such services should be available to our lecturers.

C. CICIT cooperation

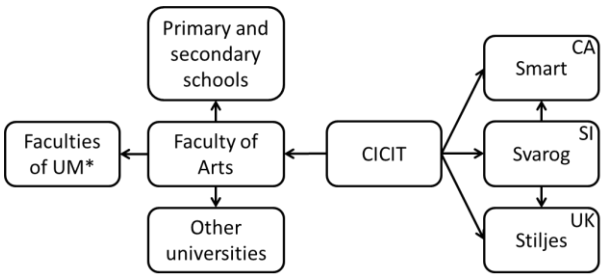


Figure 1: CICIT cooperation structure (UM = University of Maribor)

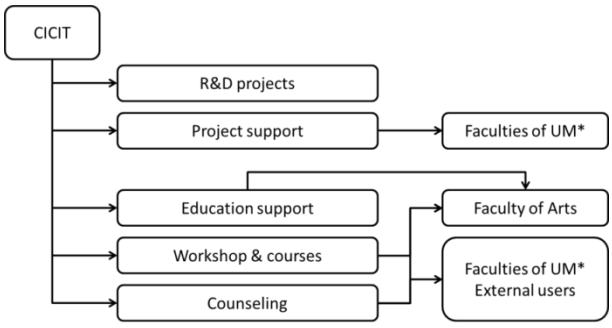


Figure 2: CICIT areas of operation (UM = University of Maribor)

D. CICIT workshops

In addition to the services that are offered individually workshops are also available. Time frames for workshops are different for internal and external users. Internal users are often lecturers and assistants which are occupied with their work and most suitable time frames for them are 2 hours. External users can have even longer time frames but not more than 6 hours with breaks.

Interactive equipment

- Introduction to interactive technology
- Learning materials production for Smart boards
- Responders in education
- PowerPoint on interactive boards
- Lecture recordings
- Interactive boards for brainstorming

Software in education

- Office software
- Moodle
- E-testing
- Collaboration tools
- Microsoft Project
- Electronic surveys
- Audio and video production

Presentation techniques

- Presentation of multimedia learning materials

IV. ESTABLISHMENT OF DIDACTICS OF ICT

It was a long process of unification of habilitation disciplines. This tough nut occupies experts for years without satisfied results. In the last two years we had to prepare three studies how to harmonize disciplines. Some disciplines were accepted but other mostly interdisciplinary ones were disputed. It is evident that disciplines collide and in these collision points sometimes scientific prestige is more important than logic. Since ICT collides with computer science this presents an additional set of problems. Computer science is not a monolith area. Originating from English language this presents additional drawback. In other languages it is simple since we know the distinction between computer science and informatics. Both areas are equally strong and managed to establish a balance. But no other discipline that collides with computer science was allowed. Faculty of Education wants to establish a "special didactics discipline - didactics of computer science" and Faculty of Natural Sciences and Mathematics wants to establish the discipline "computer in education". Both attempts have failed. Faculty of Arts made through research about disciplines in different countries a first suggestion "didactics of informatics" which would be logical and in accordance to the neighboring countries. We have failed to. Only the attempt to establish "didactics of ICT" eventually succeeded because we managed to persuade the committee to hear and understand our reasons.

E. Background for new habilitation discipline

In the preparation of the documentation we have discovered that classifications do not follow the advancement in technologies. Classifications that we used were well established in our society and academic world (KLASIUS [9], ISCED 2011 [10], ARRS [11], CERIF [12], FOS 2007 [13] and ACM [14])

Most appropriate classification is ACMs but it is not used in our university environment.

In the last decade we know that this habilitation should be established. All schools and educational institutions use ICT but training of ICT was not formally introduced. Teachers need ICT no matter which teaching discipline is their specialty. In the past when blackboards and overhead projectors were used in education the pedagogy and general didactics addressed the educational technology. But later educational technology evolved and required more time and effort to cover. A need to separate them from general didactics was satisfied and a new course was introduced. The course of educational technology soon became too narrow to cover ICT and an additional course of Multimedia was introduced. In the Bologna process of reforming curricula most of designers thought that students would have sufficient ICT knowledge upon entering the university study programs. In reality this was not the case and study paradox still holds – students who finish studies are less ICT literate than freshmen. It would be very irresponsible to leave such an important topic to self-education in the teachers training. To address this problem students of pedagogy still have multiple ICT related courses but other studies do not. To make the matter even worse there are constant battles between expert courses and didactic courses. Students from special didactics disciplines - future teachers actually get 60 ECTS of didactics courses in their study and 240 ECTS are for other courses (3+2 system).

In the discussion about the need of didactics of ICT we discovered a problem of understanding. Didactics of computer science is a discipline that is for teachers of computer science in primary and secondary schools. In some segments this discipline overlaps with the didactics of ICT but there are distinct requirements in the social sciences that didactics of computer science does not cover.

The didactics of ICT do not cover topics of:

- Programming and algorithms
- Programming languages,
- Object oriented software development
- Other topics that naturally correspond with teachers of computer science

The didactics of ICT covers topics of:

- Using computers in statistics and pedagogical experiments
- Use of authoring tools for educational material production
- Multimedia
- Didactical strategies of electronic knowledge testing and data gathering
- Strategies for distance education
- Teacher student communication
- Didactical use of interactive technology in education
- Internet safety
- Language translation tools and principles
- Other topics that naturally correspond with pedagogy.

F. Classification of didactics of ICT

The first step is to research the established classifications and use them to specify a new discipline according to the established ones. In our case we used six (6) classifications to classify didactics of ICT. Didactics of ICT is an interdisciplinary area and it naturally overlaps with some other disciplines (computer science, informatics, didactics of computer science and computers in education). Since we were unable to specify our intention with the discipline computers in education we use didactics of ICT.

KLASIUS [9]

level	category	Descriptor
1	1	EDUCATION
2	14	TEACHER TRAINING AND EDUCATION SCIENCE
4	1425	Didactics
4	1452	Training for teachers in social science subjects
4	2135	Multimedia
3	345	Management and administration
2	48	COMPUTING

3	482	Computer use
4	4820	Computer use (broad programs)
4	4821	General computer use
4	4822	Use of word processing software
4	4823	Use of spreadsheet software
4	4824	Use of computer databases
4	4825	Use of software for presenting data and information
4	4826	Use of software for processing pictures and graphics
4	4827	Use of desktop publishing software
4	4828	Use of Internet and computer networks
4	4829	Computer use (other)

ISCED 2011 [10]

1	Education
14	Teacher training and education science – Teacher training for pre-school, kindergarten, elementary school, vocational, practical, non-vocational subject, adult education, teacher trainers and for handicapped children. General and specialized teacher training programs. Education science: curriculum development in non-vocational and vocational subjects. Educational assessment, testing and measurement, educational research, other education science.
4	Science
46	Mathematics and statistics – Mathematics, operations research, numerical analysis, actuarial science, statistics and other allied fields.
48	Computing – Computer sciences: system design, computer programming, data processing, networks, operating systems – software development only (hardware development should be classified with the engineering fields).

ARRS [11]

5	Social sciences
5.01.03	Special didactics

CERIF [12]

S 271	Special didactics
S 272	Teachers education
S 281	Computer assisted education

FOS 2007 [13]

5	Social sciences
5.3	Educational sciences
5.8	Media and communication

ACM [14]

K.3	COMPUTERS AND EDUCATION
K.3.0	General
K.3.1	Computer Uses in Education (*Collaborative learning, Computer-assisted instruction (CAI), Computer-managed instruction (CMI), Distance learning*)
K.3.2	Computer and Information Science Education (*Accreditation, Computer science education, Curriculum, Information systems education, Literacy, Self-assessment*)
K.3.m	Miscellaneous (*Accreditation [**], Computer literacy [**]*)
K.4	COMPUTERS AND SOCIETY
K.4.0	General
K.4.1	Public Policy Issues (*Abuse and crime involving computers, Computer-related health issues, Ethics, Human safety, Intellectual property rights, Privacy, Regulation, Transborder data flow, Use/abuse of power*)
K.4.2	Social Issues (*Abuse and crime involving computers [**], Assistive technologies for persons with disabilities, Employment, Handicapped persons/special needs [**]*)
K.4.3	Organizational Impacts (*Automation, Computer-supported collaborative work, Employment, Reengineering*)
K.4.4	Electronic Commerce (J.1) (*Cybercash, digital cash, Distributed commercial transactions, Electronic data interchange (EDI), Intellectual property, Payment schemes, Security*)
K.4.m	Miscellaneous
K.5	LEGAL ASPECTS OF COMPUTING
K.5.0	General
K.5.1	Hardware/Software Protection (*Copyrights, Licensing, Patents, Proprietary rights, Trade secrets [**]*)
K.5.2	Governmental Issues (*Censorship, Regulation, Taxation*)
K.5.m	Miscellaneous (*Contracts [**], Hardware patents [**]*)
K.6	MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS
K.6.0	General (*Economics*)
K.6.1	Project and People Management (*Life cycle, Management techniques (e.g., PERT/CPM), Staffing, Strategic information systems planning, Systems analysis and design, Systems development, Training*)
K.6.2	Installation Management (*Benchmarks, Computer selection, Computing equipment management, Performance and usage measurement, Pricing and resource allocation*)
K.6.5	Security and Protection (D.4.6, K.4.2) (*Authentication, Insurance [**], Invasive software (e.g., viruses, worms, Trojan horses), Physical security [**], Unauthorized access (e.g., hacking, phreaking)*)

V. CONCLUSION

In a developed society with established social structure even education is highly elaborated. Quality assurance and standards are needed but at the same time they slow down the process of harmonization between technical developments and social structure that categorizes these developments. Study of computer science was established when personal computers were unavailable to the average person. Today each kid in school has a computer yet up to now the only category that could cover social aspect of application of computer in education and everyday life of individuals was computer science. Didactics of computer science was allowed to be established and grow but it is meant for teachers of computer science and does not cover the needs of social sciences. From the recognized demand of a new discipline nearly a decade was needed to be recognized and included into the categorizing structure of (higher) education. Such process was full of

misunderstandings (even deliberate ones) and in most cases so demanding in term of human effort that it has failed. In our case we succeeded and the new discipline is recognized – ICT didactics. This research work and described phases of progress could provide recognition of this discipline even on other institutions that are still holding on to established patterns.

LITERATURE

[1] M. Duh, T. Bratina and M. Krašna, "The role of digital competences in electronic education," *Media, Culture and Public Relations*, vol. 3, no. 2, pp. 131-137, September 2012.

[2] M. Krašna and T. Bratina, "Trends of e-learning in the education of teachers," in *The Fifth International Conference on Informatics, Educational Technology and New Media in Education*, Sombor, 2008.

[3] M. Duh, "Mediji pri sodobni likovni vzgoji," in *Media in education - international scientific symposium 2004.*, Novo Mesto (Slovenia), 2004.

[4] M. Krašna, T. Bratina and B. Kaučič, "Implementation of distance learning materials," in *Mipro 2009*, Opatija, 2009.

[5] M. Krašna, T. Bratina and B. Kaučič, "Differences in e-learning material production according to the learners' age and comprehension level," in *20th Central European Conference on Information and Inteligent Systems*, Varaždin, 2009.

[6] M. Duh and M. Krašna, "Aesthetics assessment of e-learning materials," in *Društvo i tehnologija 2011*, Lovran, 2011.

[7] B. Aberšek and M. Aberšek Kordigel, "Who Shall Guide Education Process - Teacher or Technology," in *DIVAI 2012: 9th International scientific conference on distance learning in applied informatics: conference proceedings*, Sturovo (Slovakia), 2012.

[8] M. Krašna, Multimedia in education, Nova Gorica: Educa, 2010.

[9] Statistical Office of the Republic of Slovenia, "Klasius-p," n.d..

[10] UNESCO, "Revision of the international standard classification of education (ISCED)," 2011.

[11] Slovenian Research Agency, "Field of Research Classification (Classification of ARRS)," n.d..

[12] Slovenian Research Agency, "Common European Research Classification Scheme (CERIF)," 1991.

[13] European Commission eurostat, "SCL - Field of science and technology classification (FOS 2007)," 2007.

[14] ACM - Association for Computing Machinery, "The ACM Computing Classification System (1998)," 1998.

Computer Science Curriculum 2013- An Example

V.Jovanovic
Georgia Southern University/Computer Sciences, Statesboro, USA
vladan@georgiasouthern.edu

Abstract - The paper presents an example of a Computer Science Curriculum redesigned in 2013 to, among other requirements, reflect important expectations of the IEEE/ACM CS Curriculum 2013 Recommendations. Among the advantages of the proposed are professional orientation, balance of fundamentals with an integrated core and advanced elective coursework, integration of design and evaluation aspects across the curriculum, a team based capstone project experience, and the longest chain of courses (including all pre-requisites) of only six semesters. Furthermore the elective CS coursework reflects three general options (possibly overlapping): a) electives grouped for three distinct professional certificates in the CS major (Security, Software Engineering, and Games Programming), b) Internship of up to a year of professional practice, and c) rich choice of contemporary CS electives.

I. INTRODUCTION

The intent of the paper is to present a contemporary model curriculum for undergraduate Computer Science degree in a format of interest to universities primarily preparing students for jobs (in an estimated proportion of under 30% students preparing for a graduate degree immediately affter college). The need for such a model is exacerbated by the recent departure from traditional practice by IEEE/ACM in providing curricular recomedtaions by including model curricula. The CS 2013 recommnedations [1] are based on 18 Knowledge Areas with numerous Knowledge Units (topics), expanding on previous practice [2], [3], [4], the CS 2001 recoomendations [5] and its 2008 interim revisions [6]. Moat of which, especfially the ACM '78 [3] provided more specific course level oriented guidelines. While the curriculum design, presented in this paper, can serve as a model compliant with [1] it is most importantly an example of a deployed curriculum. The fact that any real program evolves in its historic context does not necessarily diminishes its potential to serve as a model curriculum for a degree program.

This paper addresses general requirements (mostly perceived as design constraits, for example from [7]) a year and one third of degree specific coursework, as well as design views serving practical needs such as for advising, graduation readiness checking, catalog program description, and illustrative program of study. Any curriculum design in practice needs to address situation specific constraints, most obvious of which is experience with the existing curriculum and its faculty, but in a deliberate effort to make content of the paper of broader interest those aspects had been excluded.

The paper is organized as follows. The introduction addresses relevant educational standards and curriculum recommendations as well as some of the practical decisions to be reflected in the curriculum design including those not obvious from course descriptions intended for the university catalog (for example choice of primary programming language).

The relevant accreditation bodies are the professional ABET Computing Accreditaion Commission or CAC (in our case for Computer Science and similarly named programs) and an appropriate regional (in our case SACS) higher education accreditation commission. Both emphasize continuous improvement and are based on explicit satisfaction of Program and Student Outcomes i.e. to be defined by the Program in alignment with institutional mission and needs of stakeholders.

The main section of the paper details the curriculum design, elaborate on some individual courses, provides visualization of prerequisites, and presents typical program of study mapped to semesters/years of study.
The conclusion reflects on the curriculum design and suggests some directions for future improvements.

The following quote from [1] reflects general expectations:

> …curricula must prepare students for lifelong learning and must include professional practice (e.g. communication skills, teamwork, ethics) as components of the undergraduate experience. Computer science students must learn to integrate theory and practice, to recognize the importance of abstraction, and to appreciate the value of good engineering design.

Among the requirements from [1] that are really new compared with [8] we recognized primarily:

i. Increased professional orientation on software development in a teamwork setting
ii. Mandated coverage of information assurance
iii. Expected coverage of platforms (such as web)
iv. Coverage of parallelism across the curriculum
v. Two tiered core content requirements (tier-1, basic at 100%, and advanced tier-2 with minimum 80% to be covered by required coursework) separate from CS electives.

Among the CS program specific curriculum requirements, from ABET [7], is a one year and one third

of CS coursework (40 credit hours, counting 3h for a typical one semester course) that must include (verbatim):

1. Coverage of the fundamentals of algorithms, data structures, software design, concepts of programming languages and computer organization and architecture.
2. An exposure to a variety of programming languages and systems.
3. Proficiency in at least one higher-level language.
4. Advanced course work that builds on the fundamental course work to provide depth.

In addition, CS program must provide [7] at the minimum one year of science and mathematics (30h), specifically:

1. Mathematics: At least one-half year that must include discrete mathematics. The additional mathematics might consist of courses in areas such as calculus, linear algebra, numerical methods, probability, statistics, number theory, geometry, or symbolic logic.

2. Science: A science component that develops an understanding of the scientific method and provides students with an opportunity to experience this mode of inquiry in courses for science or engineering majors that provide some exposure to laboratory work.

The [7] also requires the following student outcomes defined as (knowledge, skills, and behaviors) abilities to: a) apply knowledge of computing and mathematics appropriate to the discipline, b) analyze a problem, and identify and define the computing requirements appropriate to its solution, c) design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs, d) function effectively on teams to accomplish a common goal, e) achieve understanding of professional, ethical, legal, security and social issues and responsibilities, f) communicate effectively with a range of audiences, g) analyze the local and global impact of computing on individuals, organizations, and society, h) reach recognition of the need for and an ability to engage in continuing professional development, and i) use current techniques, skills, and tools necessary for computing practice.

Other institutional requirements, faculty qualifications etc. are generally not directly traceable to a program curriculum design and are here excluded from discussion.

Both the ABET [7] and the Regional Commissions for Accreditation of Higher Education Institutions [8], insists on continuous program improvements and measurable outcomes. Current practice utilizes revised Bloom's taxonomy [9] in specifying learning objectives and in defining evaluation rubrics to be used to assess knowledge acquisition and program's success. The assessment aspect is excluded from further discussion in the paper not only in the interest of space, but also because it is mostly situational and inevitably at the very detail level. Nevertheless the Computer Science (CS) curriculum presented here was designed to address:

a) accreditation requirements
b) suggestions by industrial advisory board,
c) assessment of actual achievements of students,
d) faculty perspective including strenghs/research,
e) professional curriculum recommendations, and
f) actionable input from students (i.e. mobile apps).

Examples of design constraints, as requirements include: a) longest sequence of CS coursework of six semesters, b) earlier development of various applications, c) balancing CS theory with practice (showing relevance) in order to motivate and retain as well better prepare our students, and d) resolving findings from assessments (primarily regarding necessary prerequisites).

The course descriptions do not commit to a language but students and industry demands as well as faculty frustration with current disparities in productivity played the pivotal role in enabling the CS 2013 curriculum to be designed as a coherent whole (not just a compromised based set of minor upgrades).

One of the most contentious issues among the faculty was the choice of a primary programming language (Java vs. C#) and a professional grade IDE for instructions in the introductory sequence CS1 and CS2. After months of deliberation a decision was reached to organize a parallel 'experiment' first, namely, the Java/Eclipse in one section (up to 30 students) and the C#/VS 2012 in the other. This comparative trial will take two consecutive semesters (in the Fall 2013 and the Spring 2014) in the Programming Principles I and II courses (i.e. CS1 and CS2) respectively, before an evidence based final decision is to be made. The students will be allowed to choose the sections, and the common final exams will be prepared by the department (the same set of problems regardless of the language). The course descriptions for the CS1 and CS2 will reflect increased expectations (for details see new descriptions later in the text) but will remain language neutral. In the upper level coursework students are anyhow expected to work with different languages, software tools and environments without remedial instructions. The key expectation, from this experiment (besides assuring more time for faculty preparation and buy-in, and itself being an exemplary 'grand' lesson for students) is that eventual benefits should overcome inertia. The experiment will be supervised by the CS Program Assessment and Accreditation Committee and validated by the Dean's office, and a technical report/paper about the results is expected to be broadly disseminated. Our expectation is that by using a more productive environment our students will be able to experience different applications types earlier, as well as to better relate CS coursework with potential applications. All that is expected to have a positive effect on retention (in the major), improve CS graduation rate, and shorten average time needed for obtaining a degree. One of the less obvious constraints was to support transfer among programs within the College (typically from CS to IT, now we hope also from IT to CS and from EE to CS).

It is informative to reflect on past language changes, from Pascal to C++, and C++ to Java, and the related coursework required for a student to absorb the language (including essential algorithms and data structures). The Pascal required CS1 and CS2, the C++ required CS1,

CS2, and Data Structures, and Java currently requires CS1, CS2, Data Structures, part of Algorithm Design and Analysis, and a good part of the O-O Design course- now transformed into a broader Software Design course.

II. Example Curriculum as a Model

Our newly redesigned CS Curriculum is presented (as designed) using several views, starting with a stratified list of CS courses (Table I). There the CS required coursework corresponds to core tier-1 and tier-2 requirements derived from [1]. The Basic-Required set corresponds mainly to tier-1 and our Advanced-Required to tier-2. Our Advanced-Electives provide set of options that will continue to evolve our CS curriculum in accordance with resources, demands as well as important trends.

Professional orientation requirement is addressed in a number of ways: a) requiring teamwork projects in all 5000 level courses, b) providing focused electives covering areas of specializations (i.e. CS Certificates), c) incorporating professional software tools across the curriculum, and orienting Programming Principles II and Software Design courses towards software development (i.e. not just programming and more O-O syntax). Our design was influenced by Cowling [12] who characterized progression from Programming to Software Development (and then to Software Engineering) from two viewpoints: a) size of a product, and b) the scope of concerns. In his classification, Level 1 represents Programming, and deals with functionality of individual programs and Level 2 represents Software Development that deals with small applications, involves teamwork and addresses non-functional requirements. The Level 3 represents professional Software Engineering and deals with larger software artifacts, broad set of concerns and practices and as such is a major expansion of software development that can only be mastered in practice. In our model curriculum, the CS1 course corresponds to Programming, CS2 and Software Design (as well as a number of design/practice oriented courses) to Software Development level, while the Software Engineering capstone provides only a glimpse to professional software engineering (by over engineering at the modest size).

Two courses are positioned to be taken earlier, in the second year instead the third, namely Databases and Data Communications assuring timely prerequisites for two newly required courses the Distributed Web Systems Design and the Computer Security (satisfying Information Assurance requirement).

The advising chart (see Figure 1) visually represents redesigned pre-requisite schema. The decoupling of a Data Structures from the System Software, Theoretical Foundations, and all 5000 level electives, facilitates shortening chain of prerequisite courses from Discrete Mathematics to Software Engineering to only six semesters, and makes room for desirable Internship.

The recommended program of study (Table II) and progress/graduation checklist represent two additional design views (one to prove that a well prepared student can graduate within four years, and the other to provide a checklist complementing CS degree requirements with university specific core requirements). The Checklist is of marginal interest to programs outside the USA and is not included in this paper.

The three elective Certificates are designed as follows:

A. Game Programming Certificate (Complete the Game Programming CS 5439 and any two courses):

5235 - Human Computer Interaction (3)
5437 - Computer Graphics (3)
5438 - Animation (3)
5430 - Artificial Intelligence (3)

B. Network and Computer Security Certificate:

CS 5430 Artificial Intelligence (3)
CS 5531 - Systems and Software Assurance (3)
CS 5532 - Network Management Systems (3)

C. Software Engineering Certificate:

5235 - Human Computer Interaction (3)
5531 - Systems and Software Assurance (3)
5534 - Software Testing and QA (3).

Table I Stratification of CS Courses

Course Number, Name; lecture-lab-total credit hours
CS Basic-Required; 24h
1301 Programming Principles I; 3-3-4
1302 Programming Principles II; 3-0-3
3236 Theoretical Foundation; 3-0-3
2120 Computer Ethics and Society 2-0-2
3432 Database Systems; 3-0-3
3232 Systems Software; 3-0-3
3230 Data Structures; 3-0-3
3332 Data Communication; 3-0-3
CS Advanced-Required; 18h
5335 Software Design; 3-0-3
5436 Distributed Web Systems Design; 3-0-3
5331 Computer Architecture; 2-2-3
5330 Algorithm Design and Analysis; 3-0-3
5431 Computer Security; 2-2-3
5530 Software Engineering Capstone Project; 3-0-3
CS Advanced-Elective; minimum 9-12h
4790 COOP Internship (may be repeated)- hours vary
5132 Data Warehouse Design; 3-0-3
5331 Systems and Software Assurance; 3-0-3
5235 Human Computer Interaction; 3-0-3
5430 Artificial Intelligence; 3-0-3
5437 Computer Graphics; 3-0-3
5438 Animation; 3-0-3
5439 Game Programming; 3-0-3
5532 Network Management Systems; 3-0-3
5534 Software Testing and Quality Assurance; 3-0-3
5537 Broadband Networks; 3-0-3
5538 Wireless and Mobile Systems; 3-0-3
5539 Optical Networks; 3-0-3
5230 Discrete Simulation; 3-0-3
5090 Special Topics in CS 3-0-3

Let us review here, for example, a design rationale for the Software Engineering Certificate. It is designed to provide sufficient content and experience for CS students

to enable them to participate in and/or improve software evaluation practices (complementing general CS outcomes).

Our redesigned curriculum involves important details of course descriptions, course outcomes, assessment plans and methods, rubrics etc. that could not be presented in a short paper but we provide here, for illustration, a sample of the changed course descriptions:

A. CS 1301 Principles of Programming I: Course provides fundamental abilities to develop various types of computer applications (such as: GUI forms, console, office-productivity, mobile/touch apps, simple games, Web and DB connected applications) to solve problems by developing and testing programs using professional tools of an integrated development environment. Concepts include the following: essential programming principles, procedural/visual-component/event/class abstraction design of computer solutions, essential GUI design with elements of graphics and multimedia, algorithm development (using simple data types and control structures, multidimensional arrays, structures, classes and file processing), implementation, unit testing and debugging of programmed problem solutions, modularization using functions, subprograms, API libraries and user-defined classes.

B. CS 1302 Programming Principles II: Course is a continuation of CS 1301. Emphasis is on advanced techniques such as recursion, regular expressions, refactoring, object oriented programming concepts and constructs, reusing components, templates/generics, interfaces and classes (in a data and responsibility driven design), using advanced programming principles and idioms in Multilanguage/Multi-paradigm development, as well as testing, integration, implementation, and deployment of various application types (including rich GUI, services, workflow, cloud based apps etc.) using advanced options in an integrated development environment and experiencing teamwork with shared (code) repositories.

C. CS 5335 Software Design: Introduction to concepts, methods, approaches, tools and contemporary practices in software design and analysis (including object oriented, aspects oriented, component/composition-based, etc.). Topics also include a study of the Unified Modeling Language to model, design and implement software in small group projects. Additional topics include software requirements including nonfunctional, elements of software architecture, software metrics for design evaluation, as well as use of design patterns and application framworks in building contemporary applications.

D. CS 5531 Systems and Software Assurance: Course presents body of knowledge in systems and software assurance and evaluation including security, safety, and integrity analysis. The core part of the course is software assurance where students are exposed to code and architectural analysis, secure coding practices, standards, and tools. Course explores standards in modeling internal security at the organizational level and will involve students in risk assessments, comprehensive assurance planning, as well as an array of countermeasure considerations.

E. CS 5132 Data Warehouse Design: Data warehouse (DW) design principles and technical problems. Topics include: DW architectures, organizing DW design projects, analyzing data and requirements, SQL analytics, materialized views, star-joins and other DW related features, dimensional modeling, physical design and implementation of integrated data warehouses using commercial ROLAP engines.

Courses A and B are examples of core tier-1, C of core tier-2 and D and E are examples of CS electives.

III. CONLUSION

One of the relevant issues, not elaborated in the paper, is differentiation and parallel evolution (expansion) of programs in computing. The [10] and [11] can help one to recognize substantive distinctions among Computer Science (CS), Computer Engineering (CE), Information Systems (IS), Information Technology (IT) and the Software Engineering (SE) types of undergraduate academic degree programs (again mainly in the USA). Among the CS programs we further differentiate: a) mathematics/theory oriented programs, typically preparing students for graduate studies or scientific research, and b) predominantly applied programs, oriented towards preparing students for the workplace.

The new context of an Engineering and IT College, since 2012 hosting the CS, IT as well as traditional industry oriented engineering programs (Electrical, Mechanical, and Civil), encouraged us to revise our CS program towards increased professionalism even before the CS 2013 [1] recommendations become available. In the implementation, the advising of CS students will play more prominent role particularly as students can choose Certificates and/or Internship options. Furthermore we will use here as example the case of Software Engineering Certificate illustrated by a Software Testing and QA course where the content requirements are explicitly drawn from the relevant professional recommendations like SWEBOK [13] and ISTQB [14], and only validated by years of experience and use of common textbooks such as [15],[16],[17] and [18]. A legitimate question raised by some of our faculty, in the context of CS 2013 Curriculum Recommendations [1] move towards software development instead of programming (in the sense of [12]) as a common foundation to both CS and SE Programs [1], was formulated as: how close, for students taking the Software Engineering Certificate, is our applied CS Program to a SE Program? For an undergraduate SE degree program, areas like software architecture, requirements, metrics, process, etc. will require lot more attention but the main differentiator is probably implied/intended application domain. For most SE programs that domain is frequently real-time software for embedded systems (automotive, avionics, etc.) and mainly engineering applications. New graduate SE Programs are demonstrating [19] integration of systems and software engineering perspectives that is operating on

a larger scope requiring more mature students than traditional CS/SE undergraduates, so it seems logical that SE graduate programs will be the venue with a larger growth potential. For the CS undergraduate programs to remain relevant for a majority of work oriented students a professional model program may be just the right option.

One possible direction for improvement is redistribution of the content from the Data Structures into the CS2 and the Algorithm Design and Analysis courses, pending successful CS1 acceleration under contemporary IDEs. When that happen we may be able to include Artificial Intelligence among required courses for the degree; the reason will be to provide background for knowledge representation and better prepare our students for a more knowledge intensive cybernetic future.

REFERENCES

[1] ACM/IEEE-CS,Computer Science Curricula 2013 (http://cs2013.org/ironman-draft/cs2013-ironman-v1.0.pdf)

[2] W. Atchinson et all, Recommendations for academic programs in computer science: a report of the ACM curriculum committee in computer science, Comm. ACM V11, N3, March 1968.

[3] ACM'78 CS Recommendations for the Undergraduate Programs in Computer Science, ACM Education Board, 1979.

[4] A. Tucker, Computing Curricula 91, Comm. ACM V 34, N6, June 1991.

[5] CS 2001, Computing Curricula 2001 Computer Sceince, IEEE and ACM Joint Task Force, Final Report October 2001.

[6] CS 2008, Computing Science Curriculum 2008: An Interim Revision of CS 2001, ACM & IEEE-CS Interim Review Task Force, December 2008.

[7] ABET CAC, Crtiteria for Accrediting Computing Progams, Effective for Review Cycle 2012-2013, October 2011.

[8] CRAC, Regional Accreditation and Student Learning: A Guide for Institutions and Evaluators, Council of Regional Accrediting Commissions 2004.

[9] L. Anderson, and D. Krathwohl, Eds., A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational objectives: Complete edition, Longman., 2001.

[10] V. Jovanovic, and H.Reicghelt, "Comparison of computing programs at the level of intent" ACM SIGITE Newsletter, V2, N1, Spring 2004.

[11] R. Schackelfor, et all, Computing Curricula 2005: overview report, ACM SIGCSE Bulletin, V38, Issue 1, March 2006.

[12] A. Cowling, "Stages in teaching software design" 20th Conference on Software Engineering Education and Training, CSEET 2007.

[13] SWEBOK 2004, Guide to Software Engineering Body of Knowledge, version 2004, IEEE CS Press, http://www.swebok.org

[14] ISTQB Advanced Level Syllabus 2012, Technical Test Analyst, ISTQB 2012.

[15] P. Ammann, and J. Offutt, Introduction to Software Testing, Cambridge University Press 2008.

[16] A. Mathur, Foundations of Software Testing, Pearson 2008.

[17] P. Jorgensen, Software Testing- A Craftsman's Approach, 3ed. Auerbach 2008.

[18] G. Shulmeyer, Handbook of Software Quality Assurance, 4ed. Artech House 2008.

[19] GSwE2009, Graduate Software Engineering 2009, Curriculum Guidelins for Graduate Degree Programs in Software Engineering, Stevens Instritute of technology, September 2009.

Appendix

Table II A Sample Four-Year Study Program

Year I	Fall 16 h	Spring 16h	**Year II**	Fall 16h	Spring 16h
ENGL Composition 1	3		Science 2. in sequence	4	
ENGL Composition 2		3	Physical Education Activities		1
MATH Calculus I		4	CS Computer Ethics and Society		2
Discrete Mathematics	3		CS Database Systems	3	
History	3		MATH Calculus II		3
Science 1. in sequence		4	MATH Statistics	3	
Global Citizens		1	CS Theoretical Foundations		3
CSCI Programming Principles I		4	CS Systems Software	3	
Social Science	3		CS Programming Principles II	3	
Health	2		Principles of Public Speaking		3
First year Seminar	2		CS Data Communications		3
Year III	Fall 16h	Spring 16h	**Year IV**	Fall 15h	Spring 15h
Science 3. with Lab	4		CS Computer Security		3
CSCI Distributed Web Systems Design	3		Social Sciences	3	
CSCI Computer Architecture		3	Social Sciences Elective	3	
Physical education activities		1	CS Elective ** (Systems and Software Assurance)	3	
ENG Technical Writing	3		Free Elective	3	
CS Software Design	3		CS Software Engineering		3
CS Algorithm Design & Analysis		3	CS Elective ** (Software Testing and QA)		3
Free Elective		3	Free Elective		3
Foreign Language		3	Free or CS Elective (Artificial Intelligence)	3	
ENGL World Literature		3	Free Elective		3
CS Elective ** (HC Interaction)	3				
** *Example shown for the Software Engineering Certificate option*			**Total for the Bachelor of CS degree is min 126 hours**		

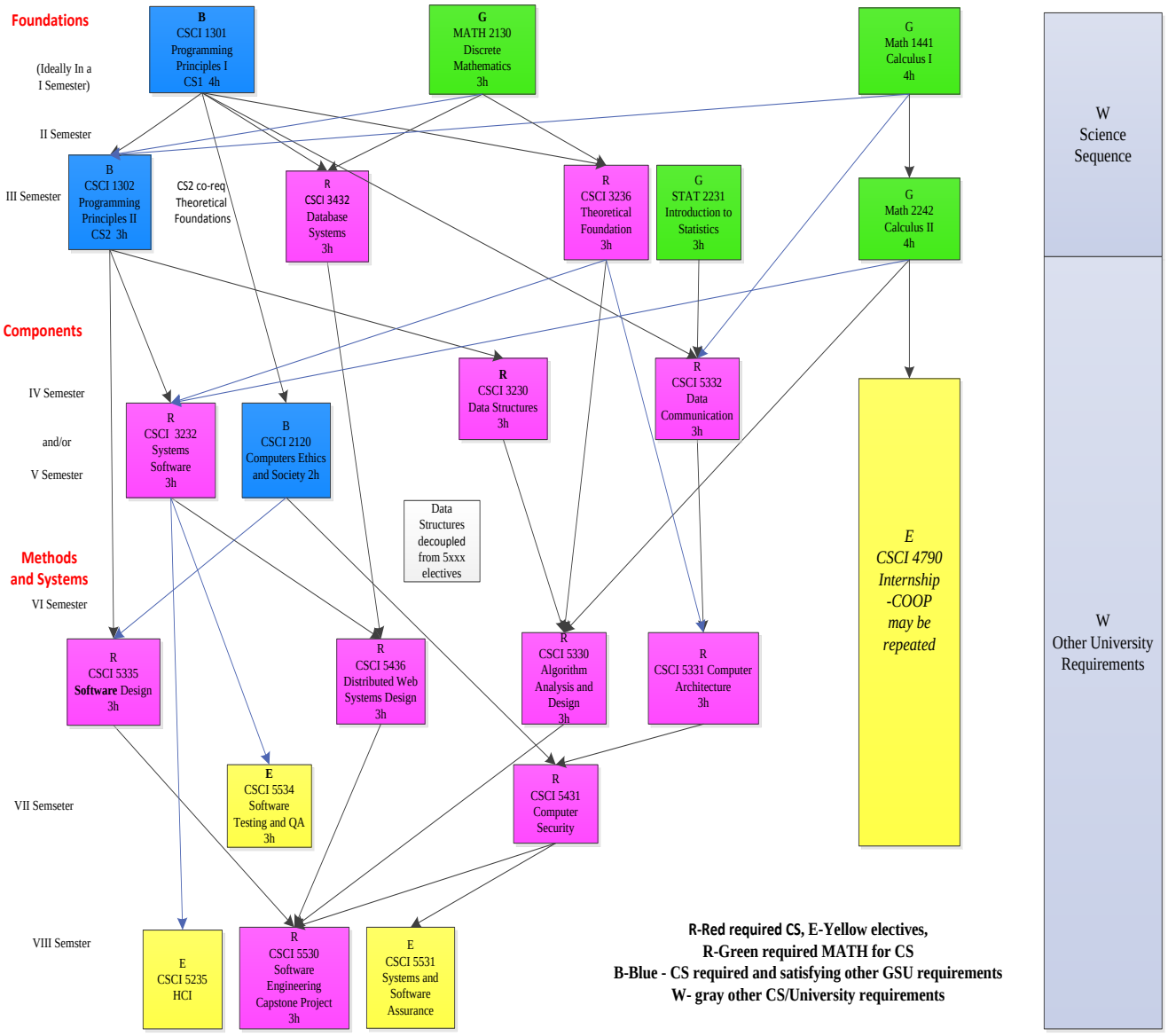


Figure 1. Advising Chart

Determination of the Torque-Speed Characteristic of Induction Motor in Electric Machinery Education

Rade Buinac, Veselko Tomljenović
Electrical engineering department, Polytechnic of Zagreb
Konavoska 2, Zagreb, Croatia
rade.buinac@tvz.hr

Abstract - The paper describes the measuring method for determination of the torque-speed characteristic of an induction machine used in electric machinery classes performed at the Polytechnic of Zagreb. The torque-speed characteristic is determined by recording and differentiating speed signal during the starting. Data is gathered using a measuring system based on a simple digital acquisition card, and processed in custom software, built with LabVIEW, on a personal computer. Advantages and shortages of this sort of measurement compared with other measuring methods are given, and measuring equipment deployment and software development are described.

I. INTRODUCTION

The induction motor torque-speed characteristic gives information on the dependence of induced mechanical torque on mechanical speed. Sometimes this characteristic is called torque-slip characteristic, where slip is defined as speed relative to synchronous speed expressed as a fraction of synchronous speed.

Typical torque-speed curve of a squirrel-cage induction motor given in Figure 1 provides information about the operation of induction motor. The induced torque of the motor is zero at synchronous speed. The curve is nearly linear in normal operating conditions which are between pullout speed and synchronous speed [6]. Maximum possible torque the machine can induce is called the *pullout* torque or the *breakdown* torque and it limits the short-time overload capability of the motor [7]. It is usually two or three times the rated full-load torque of the motor. The starting torque of the motor is slightly larger than its full-load torque, so this motor will start carrying any load that it can supply at full power [4].

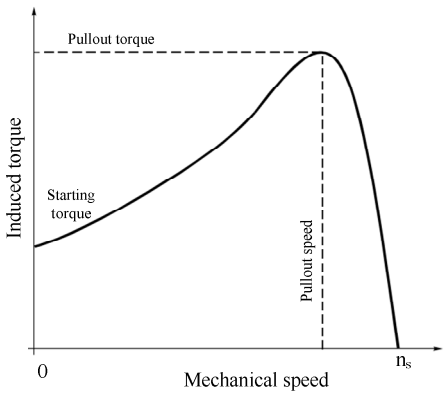


Figure 1. Typical torque-speed characteristic of an induction machine

The torque-speed characteristic is required in order to predict behavior of induction machines in electrical motor drives [1]. It should be measured as accurately as possible, while the selection of which method to apply depends on the characteristics of both the machine and the testing facilities [3].

Main scope of the paper is the implementation of a new measuring method into laboratory exercises and integration of two methods in one exercise. It shortly describes two of the methods outlined in IEEE standard test procedure for polyphase induction motors and generators [3], one of which had been used in the exercises, and the new one, recently included. Main technical aspects of the methods are described in the paper, as well as the deployment of the measuring equipment, development of the software needed for measurements and data processing, and the expected benefits in class.

II. LABORATORY EXCERSISES

Laboratory exercises are performed in the summer semester of the second year of study, after the students have taken the "Electrical machines I" course, and therefore have certain knowledge about electric machinery which enables them to be more self-confident in individual work in the laboratory. The subject is divided in six exercises which follow the classroom lectures, and help students gain better understanding through practical work with electric machinery and measuring equipment. All of the exercises are done in small groups of maximum seven students whose scope is to perform the tasks under the guidance of the teaching assistant. Each individual exercise lasts two academic hours during which students have to take notes which will be the base for reports they have to generate in order to successfully complete the course. Reports are based on the guidelines given in the printed material for all of the exercises. The concept of producing reports documenting work in the laboratory is an old and proven method in engineering disciplines education, especially because of the importance of technical reports in engineering. This was one of the

guiding concepts in the integration of the new measuring technique, described later in the article, into existing program of laboratory exercises.

Torque-speed characteristic is the title and the central part of the report produced after the exercise described in this article. Students use measuring equipment under the supervision of the teaching assistant and, following his directions, and guidelines given in the manual, take notes and finally draw the torque-speed characteristic. As mentioned earlier, there are several methods for determination of the torque-speed characteristic of an induction machine. Four of them are described in [3]. The method that has been used so far is based on direct measurement of the torque and speed. Electric generator dynamometer is directly coupled to the tested engine. Dynamometer housing is made free to rotate and connected to the weight transducer by the torque arm. Speed is measured by inductive transducer mounted on the shaft of the dynamometer. The students control the load and the speed of the motor by controlling the field windings voltage of the electric generator dynamometer connected to a large resistor. Measurement is performed at reduced voltage maintained using a variable-voltage transformer. Equipment setup for this exercise can be seen in Figure 2.

Measuring points are taken depending on speed, ranged from locked-rotor to synchronous speed. Experience has proven a dozen points are enough, with preference of choosing more points in higher speeds range. Alongside speed, at each point the operator, i.e. the students, measure the torque, phase voltage, current, and real power. During the production of the report, in order to establish values of current, voltage and torque at rated values, the students need to calculate them from the values measured at reduced voltage. As the final goal of the report, the torque-speed characteristic, and the current as function of speed characteristic need to be drawn.

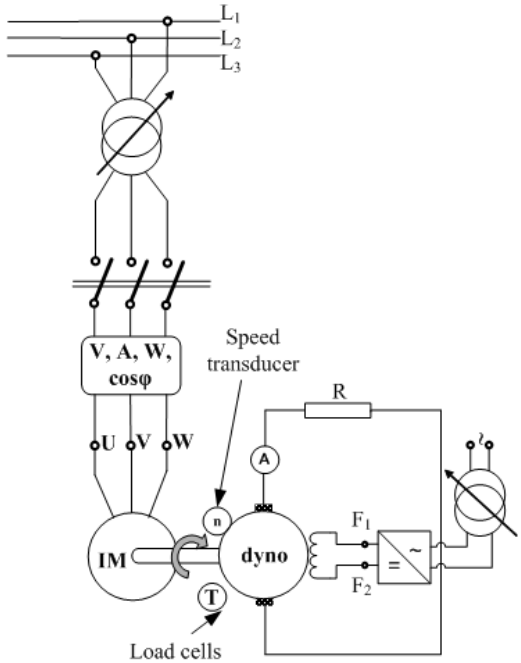


Figure 2. Equipment setup in laboratory exercises for direct measurement of the torque and speed

III. IMPLEMENTATION OF THE NEW METHOD

Another method for obtaining data for a torque-speed curve of an induction machine outlined in the IEEE Standard 112-1996 [3] is the Acceleration method. In this method the value of acceleration is determined at various speeds, and the torque at each speed is determined from the acceleration of the mass of the rotating parts. Torque is calculated as follows:

$$T = J \cdot \frac{2\pi}{60} \cdot \frac{dn}{dt} \tag{1}$$

where T is the torque in Nm

J is the moment of inertia of rotating parts in $kg \cdot m^2$

n is the speed in min^{-1}.

The accelerating time should be long enough so that electrical transient effects in the instruments and in the motor do not distort the torque-speed curve [3], and long enough to ensure successful measurement of the necessary number of data points. Detailed explanation can be found in the literature [1] and [2]. The benefits of acceleration method are short starting period and small heating of the motor, which allow repeating of the experiment, and the ability to perform the tests at rated voltage.

The measuring technique in laboratory exercises is based on the acceleration method, but it is not performed under no load conditions. There are two reasons for such modification of the method. Primary reason for adding the load is increasing the duration of accelerating time to the value ten times larger than electromagnetic time constant which represents the duration of electrical transient effects. To avoid the influence of the transient effect on the torque-speed characteristic, motor should be reversed, meaning it should be rotated manually in the reverse direction to that expected when the motor is energized [2]. This can be performed using the electric generator dynamometer working as a motor. Using the dynamometer as additional load obviates the time-consuming activity of uncoupling the motors after the first part of the exercise, allowing students to be more individual in the laboratory work. Accurate measurements of speed and acceleration are essential requirements [3] for this method, so the data acquisition system should be fast enough, in terms of sampling rate, to record speed, voltage and current, which also may be of interest. Speed transducer used in acceleration method should be a tachometer. It is necessary to have information about the rotation direction because of the reversing before energizing the machine. The measuring system should be able to record the speed, perform mathematical operations including differentiating to determine acceleration and display the results during the measurement and after the processing.

IV. MEASURING SYSTEM

Measuring system is a product of Veski Inc. from Zagreb, consisting of measurement/acquisition device and a personal computer equipped with corresponding software. The system is designed for measurement, preparation, acquisition, storage, processing, and data presentation of voltages, currents and speeds for rotating machinery under 5 kW of power. It is adjusted for measuring the currents up to 10 A, voltages up to 500 V and speeds up to 60000 min^{-1}. Sampling rate is 48 kHz which is high enough to enable obtaining the speed by differentiation, i.e. to avoid the coarseness of quantization. The acquisition device is connected to the PC via a USB port. This data acquisition device is used to measure different physical quantities, some of which are mechanical, and some electrical. Using appropriate measuring transducers this device may be used for measuring any type of quantity, whether mechanical, chemical, etc. The software determines its functionality and the way it emulates traditional electronic instrumentation. This concept is called virtual instrumentation, and it is one of the main advantages of a system such as this. The block scheme of the measuring system including the software is given in Figure 3.

Acquisition device is modular, and composed of modules for signal adaptation (voltage and current transducers, resistors), data acquisition card, and power supply. Device has three inputs, for voltages and currents of each phase, additional input sockets for tachometer signal, measuring signal control sockets, and a USB socket. The deployment of measuring system is simple and intuitive for people with small experience in electrical measurements, so it can be done by the students themselves during the exercises. During the equipment deployment, leads should be connected from the voltage source to the acquisition device. In this method, motor is connected directly to the power grid, so the (3-phase) variable voltage transformer should be disconnected and motor should be started directly on line (DOL).

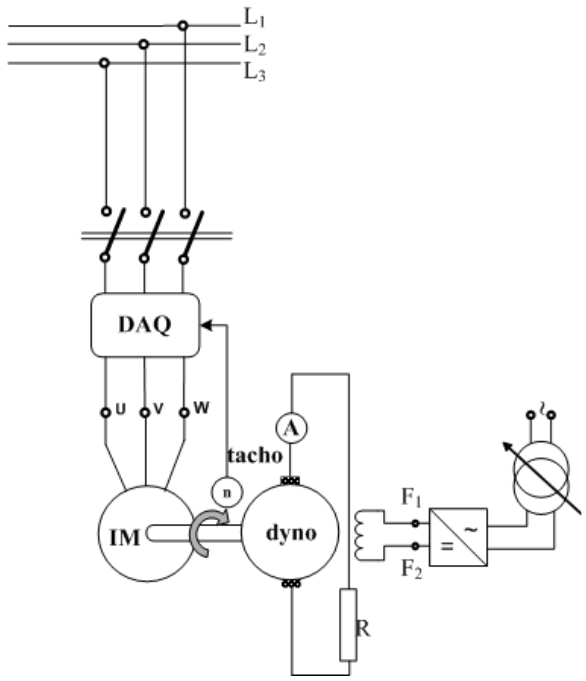


Figure 3. Equipment setup in laboratory exercises for acceleration method

Tachometer should be connected to the sockets provided for speed measurement. Equipment setup should follow the one given in Figure 4. Software deployment, on the other hand, provided some challenges and it should be carefully monitored when performed by the students.

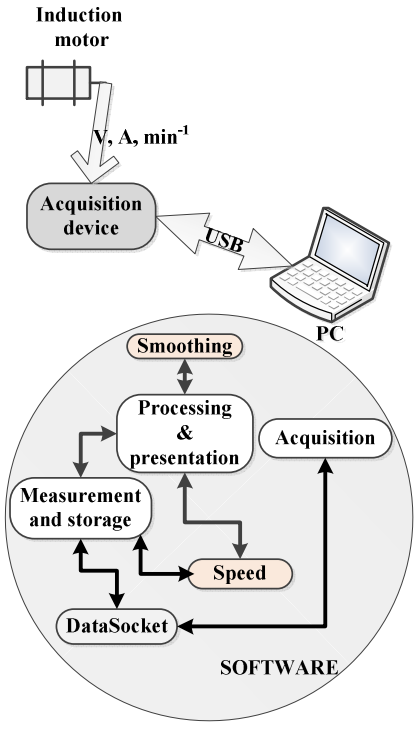


Figure 4. Block scheme of the measuring system

V. SOFTWARE

Software for the measuring system was developed using LabVIEW, programming tool designed for developing applications in test and measurement, automation, instrumentation control, data acquisition and processing [10]. This tool was chosen because applications for measurement are developed much faster in LabVIEW than with *traditional* programming languages. Rich user interfaces imitating parts of *conventional* measuring instrumentation and equipment are included in the program. This is important for electrical engineering students who have some experience in working with electronic instrumentation. User interfaces reminding the students of an oscilloscope, ammeter, 7-segment numeric display, knobs, button, dials, etc. provide the students with familiar surroundings, and hopefully giving them an insight to another possible application of the personal computer, not just the Internet and fun they are used to [8]. Although LabVIEW is declared to be compatible with different operating systems, software installation on unsupported Linux distributions was not simple, and some issues had to be resolved without the official support from the manufacturer, National Instruments. Installation and following software development under MS Windows went smoothly. One of the dominating distinctions between

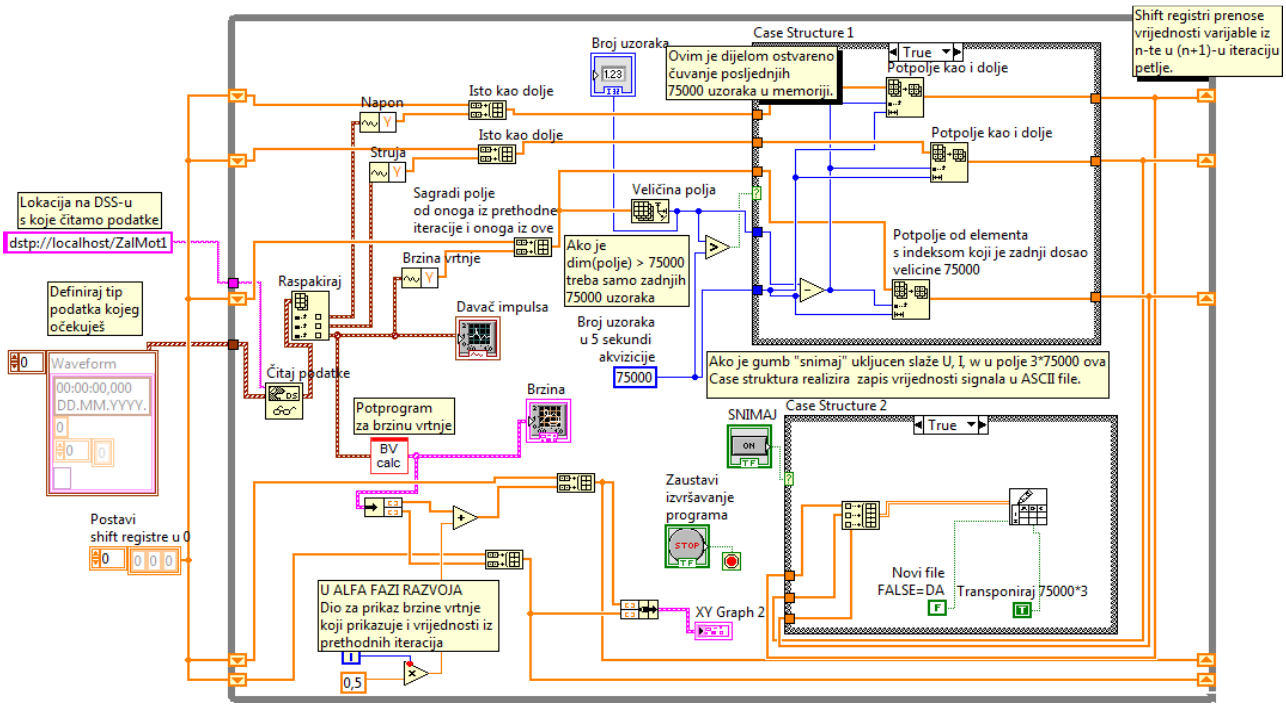


Figure 6. Block diagram of the *Measurement and storage* VI

LabVIEW and other programming languages is the *dataflow* paradigm, a programming paradigm in which functions are not executed in a determined order, but following the arrival of data to a certain function. Thus, the order of execution of functions does not have to be the same every time we start the program. Source code in LabVIEW is written, or better, drawn as a block diagram. Functions are represented by *blocks* and *nodes* interconnected with *wires*. An example of a block diagram is given in Figure 5. Objects in LabVIEW are called *virtual instruments* (VIs) and the structure of an application is often built up of several or more VIs. SubVI is a object large enough to be a VI, but used by VIs of higher rank to perform a specific function. Software developed for the torque-speed characteristic determination purposes has three main VIs and four subVIs, two of which are custom made, and two are integral part of the LabVIEW package. Figure 3 shows the structure of the VIs, with main VIs colored white, and subVIs colored pale red.

The *Acquisition* VI starts communication with the hardware and opens a link to the DataSocket server – a protocol used by LabVIEW for sharing data between programs and computers. This VI uses two standard subVIs for communication establishment and data acquisition. The *Measurement and storage* VI converts signals received, maintaining the last five seconds of the signal during live recording in the buffer [9] and, at the user's command, stores data to the hard drive. Processing and (graphical) presentation is performed in a VI bearing that name. The result of the software analysis, the torque-speed curve can be seen on Figure 6. The result given is smoothed and polynomial interpolated using least square method, so the students are left to discuss precision and accuracy with the teaching assistant. This seems to be a good "trick" to interest the students more for electrical measurements problems.

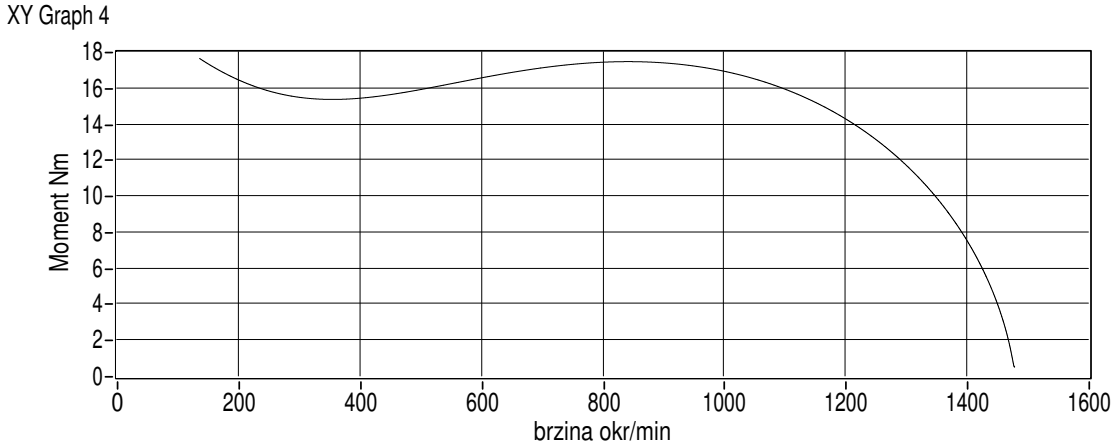


Figure 5. The torque-speed characteristic drawn by the software

VI. CONLUSION

Implementation of the new method in the exercises has several benefits. Students will notice that it is possible to measure the same physical quantity in several ways, get familiar with some of the computer tools used in contemporary electrical engineering, and be able to run the software and make simple modifications themselves. This would be encouraged in order to help better understanding of the electric machinery and the measuring techniques. Teamwork would be important as a desirable quality during the performance of this exercise, preparing the future engineers, at least a bit, for their future work in a team. Students' reactions to the new method are to be tested through a questionnaire and through laboratory reports comparison between several generations in order to qualitatively evaluate improvements in students' knowledge and skills. Without careful investigation, however, it can be stated that their competence in electrical measurements and their skills in computer usage are improved through work with the newly deployed equipment.

ACKNOWLEDGMENT

R.B. would like to thank Boris Meško for his help with the measurement system, Mato Fruk for his advice and support, and Marija Krznarić for her commitment and help with almost everything.

REFERENCES

[1] S. Štefanko, M. Pužar, “Measurement of Torque-slip Curve of Induction Motors during Starting” Automatika, Zagreb, vol. 33, pp. 41-46, April 1992.

[2] B. Szabados, R.D. Findlay, G.M. Obermeyer, R. E. Draper, "Measurement of the torque-speed characteristics of induction motors using an improved new digital approach", IEEE Transactions on Energy Conversion, Vol. 5, No. 3, pp. 565–571, September 1990

[3] IEEE Std 112-1996, "IEEE Standard Test Procedure for Polyphase Induction Motors and Generators", pp. 31-34

[4] Stephen J. Chapman, Electric Machinery Fundamentals, 2nd ed., McGraw-Hill Inc., New York, 1991

[5] M. Fruk, D. Maršić, G. Malčić, “Mjerenje brzine vrtnje pomoću programirljivog logičkog kontrolera”, MIPRO 2008: Proceedings, vol. 2 (CTS&CIS), pp. 102–107, Croatian Society MIPRO, Rijeka, May 2008

[6] I. Mandić, V. Tomljenović, M. Pužar, "Sinkroni i asinkroni električni strojevi", textbook, Polytechnic of Zagreb, Electrical engineering department, Zagreb, 2012

[7] A.E. Fitzgerald, C. Kingsley, S.D. Umans, "Electric machinery", 5th SI ed., McGraw-Hill, London, 1992

[8] M. Krznaric, R. Buinac, "Can online learning help Croatian ESL students achieve better fluency", MIPRO 2010: Proceedings, vol. 4 (CE), pp.876-879, Croatian Society MIPRO, Rijeka, May 2010

[9] R. Buinac, Measuring the torque-slip characteristic of an induction machine from starting, diploma thesis, Polytechnic of Zagreb, Zagreb, 2012

[10] Jeffrey Travis, Jim Kring, LabVIEW for Everyone, Person Education, Upper Saddle River, New Jersey, 2007

RELEVANCE AND REALIZATION OF COMMUNICATION BETWEEN INFORMATION AND COMMUNICATION TECHNOLOGY EMPLOYERS AND EDUCATIONAL SECTOR

Mirna Rotim
APIS IT, Zagreb, Croatia

SUMMARY

Education enables the realization of skills, knowledge, attitudes and values that each individual requires to fulfill their work and social goals. It increases economic competitiveness by ensuring the necessary qualities of human resources. One of the priorities of the educational system in the upcoming period should be the assessment of skills and abilities the economy needs as well as their long term developmental trends.

The speeding rhythm of technological growth deepens the gap between technological requirements and education and it cannot be simply resolved by altering the length of the educational process. Namely, many skills are outdated even before they are realized nowadays. Therefore it is crucial to maintain a continuing and open communication between employers and educational institutions, which is still an exception in Croatia. Considering the fact that educational results and therefore the competences required for employment in a specific area (the focus is on education and employment in ICT sector) derive from (sub) standard education programs, which highly depend on the communication between these two, a research on the realization of communication was conducted. The research examined the extent to which ICT companies communicate with educational institutions that produce the necessary human resources and who should encourage the cooperation.

Key words: education, ICT, cooperation between employers and educational system, human resources.

I. INTRODUCTION

By globalization and spread of new technologies the economies of developed industrial countries are transformed into knowledge-based economies. An important part of the competitiveness of the national economy is the level of knowledge and skills of the labor force.

ICT is the basic infrastructure of knowledge-based society. It is the bearer of competitiveness and economic success in the 21st century. Its development and implementation requires new ways of working as well as innovative and productive application of knowledge and is consequently the groundwork for the successful operation of science and entrepreneurship.

Although higher education should not be approached exclusively from economic impact perspective, its social, cultural and political significance should be considered as well, it must be emphasized that higher education should play the key part in economic development and enable acquisition of knowledge, skills and competences required for successful participation in labor market. Educational system and labor market must be in constant communication.

In this paper, emphasis is placed on the participation of the educational system in the development of society through the prism of achieving a successful and constructive communication between employers in the field of ICT and educational institutions that educate these professionals.

II. EDUCATIONAL PRIORITIES

Information technology plays irreplaceable and key part in development of any society ant its economic progress. The most advanced countries in the world incessantly invest in development of technology, and thus in the experts in the area and educational institutions which educate them. Admission quotas to engineering colleges in developed countries are already incomparably greater than in our country with the tendency of continued growth. The participation of the business sector in the area of technological education system is huge, ranging from financial investments to ongoing concerns about the employability and practical training of their students.

In Croatia, the only thing that follows the modern trend is the satisfying quality of engineering colleges.

To explain, the number of admission quotas is too low as well as the participation of business sector in the work of educational institutions. The only area where Croatia could be compared to the developed countries in the field of educational system is that the educational institutions are perceived as the best 'manufacturers' of future high quality experts.

Research conducted by MojPosao for EduCentar during July and August 2012 on a sample of over 500 employees and 50 employers on educational needs and supply of professional education showed that engineering colleges are considered the best by the respondents.

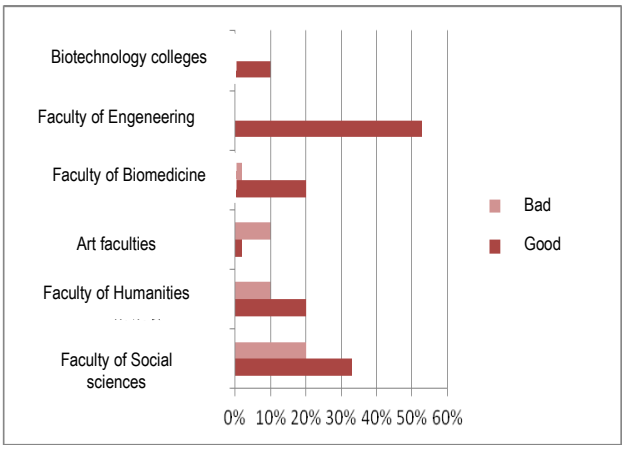


Figure 1 Universities that according to respondents give particularly good and particularly bad candidates

Most of these employers, 53% of them, think that engineering colleges give the best candidates. None of the employers consider that engineering and biotechnical colleges give bad candidates, while the quality of other colleges was estimated as much worse (Figure 1).

Although colleges have generally a triple mission – research, educational and social (participation in development of society) – the reforms conducted in our educational system and which are urgent are mostly directed to the third mission which among other things incorporates bridging the gap between the formal educational system and labor market needs. Labor market has constantly emerging needs and the dynamics of changes in educational system does not interfere with them.

Among the biggest issues in our educational system, the following can be highlighted:

- incompatibility of knowledge and skills of the workforce and labor market needs,
- too high labor costs for current level of productivity,
- low mobility of the workforce

Recommendations to improve the competitiveness of our economy, in the context of the education system are to implement the following measures:

- establish a system of monitoring the demand for knowledge and skills,
- ensure that educational outcomes follow the needs of the labor market,
- increase labor productivity,
- increase the adaptability of the workforce to restructuring of the economy,
- establish better cooperation and achieve a higher degree of responsibility of the social partners for lifelong learning

These recommendations clearly show that educational system and labor market must constantly communicate and that knowledge application, as the basis of success of any educational program, greatly depends on high quality communication between the system which educates the workforce and the one that employs it.

Obsolescence of knowledge is a characteristic of our time and that is specifically true for the field of ICT, which when put in context of educational system, almost completely changes during one generation of students. Educational system, which is much slower than economic system, particularly needs employers in that area because in order to be successful they have to adapt to quick changes in the world of technology.

III. PAST RESEARCH IN THE FIELD OF LABOUR MARKET AND EDUCATIONAL SECTOR CONNECTION

Not much research had been conducted in Croatia covering this issue, but the research conducted as a part of a project or the Institute for education development showed that 52.6% of Croatian students is mostly or completely dissatisfied with available information in higher education that refer to labor market. Information available to students mostly comes from combination of unofficial sources at university (other students and professors) and outside of university (family and other contacts outside of university). Finally, there is no connection between higher education and the labor market in the segment of informing students about employment opportunities after graduation.

MojPosao portal conducted a survey in September 2011 on the compliance of the education system and labor market among 300 respondents. The results showed that 62% of respondents believe that the education system has nothing to do with the labor market. They mostly believe that the system exists for its own purpose.

Instead, they added, there are high admission quotas for occupations which are in excess on the labor market. They also emphasize that universities rarely ensure training which results in not being capable of doing the job they had studied for. Other 27% of respondents believe that the school system is obsolete. They mostly believe that nothing has changed in education for years. They also add that there are no long term plans of workforce development for jobs that will emerge in the future. 9% believe that the school system adapts to labor market needs but not at satisfying pace, and only 2% that the system is well in line with labor market needs.

Distrust of public towards the educational system is also confirmed by the results of Research on educational needs 2011, conducted by EduCentar and MojPosao.

Only 9% of employees believe that formal education provided them with excellent basis to work in their profession, 27% believe it is very good basis and 19% find it insufficient.

Employers showed a somewhat bigger trust in educational system. 38% of them believe that formal education gave their employees a very good basis for doing the job, 3% believe it is excellent and also 3% find it insufficient.

Survey conducted in 2008 on the sample that consisted of 259 employees in ICT companies in Croatia investigated the perception of knowledge and skills acquired through formal education in IT educational institutions, i.e. whether the employees believe that they acquired sufficient knowledge through formal education for the job that they do and how prepared they were to join the

business world once they finished formal education. The results lead to conclusion that the respondents believe that they through formal education did not manage to acquire knowledge and skills related to required fields which are: knowledge and skills related to job within their profession not acquired through formal education by 27.8% of respondent, 52.1% believe they acquired them partially, and only 20% believe that they fully acquired sufficient knowledge and skills for the job that they trained for during formal education. Knowledge and skills required for a specific current job were fully acquired by only 5.8% of respondents, partially by 45.9%, whereas 48.2% received very little or no knowledge and skills altogether. Respondents' perception of the quality of acquired foreign language knowledge is such that 59.1% believe that through formal education the only acquired partial or full knowledge in that area and 41 % believe that they did not acquire this specific knowledge at all or at a very low level. General knowledge and knowledge not tightly connected to their profession was acquired by 63.3% partially or fully, whereas 36.7% acquired very little or none at all. The development of communication skills was abandoned in whole or in part in 58.7% of respondents, while 41.3 % of them believe to have received good education in this area during the formal education. Negotiating skills were neglected in as much as 74.9 % of the respondents, and teamwork skills were not developed during the formal education of 47.4% of respondents. Generally, the respondents mainly agree with that they received sufficient knowledge and skills for job in their profession and that their formal education gave them sufficient amount of general knowledge of foreign languages and mainly disagree that formal education gave them enough knowledge and skills for specific job at which they are currently employed.

Survey participants that the following was at most neglected during formal education: negotiation skills, practical knowledge in the field of information technology and the practical knowledge and skills in the profession. They considered common knowledge within the profession, general knowledge, awareness and knowledge of foreign languages as the least neglected.

To further prove the above mentioned statements, there are Eurostat research results from 2008 where employers estimated that their employees lack many knowledge and skills in relation to business needs. On a scale of least developed skills are managerial skills, problem solving, leadership, communication skills, teamwork skills, and proficiency in English.

OECD research showed that Croatia has more higher education graduates in business world and general population in general than Austria and other most successful transition countries, but that higher education graduated in Croatia mostly graduated from colleges that are not adapted to the needs of labor market and modern economy and that needed professions often lack in required competencies (Figure 2).

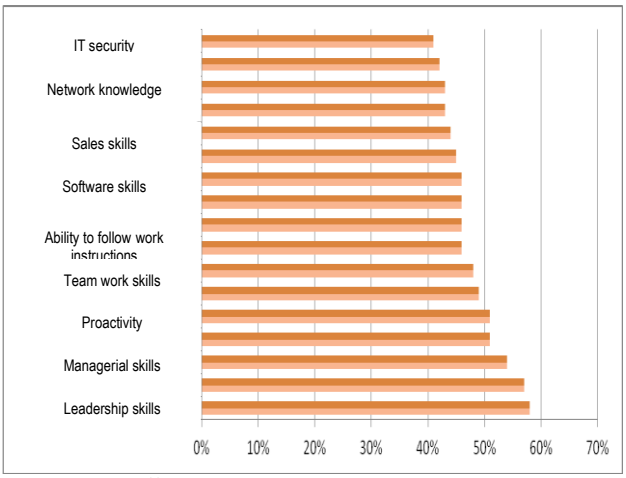


Figure 2: Differences between business needs and employees' knowledge (Eurostat 2008)

Therefore, it is necessary to systematically coordinate education and the needs of the labor market, because skilled labor force by itself is not sufficient development resource.

This research and its unpopular results, as well as the fact that we mostly rely on experts from ICT field in our country's development, has prompted us to research whether employers from this field communicate with educational institutions, if yes what is the quality of that communication and on whose part in their opinion it should be improved.

IV. EMPIRICAL RESEARCH ON EMPLOYERS AND EDUCATIONAL INSTITUTIONS COMMUNICATION

Research goal:

To survey employers for ICT field in what way they cooperate with educational institutions, what benefits they expect and have from that cooperation and who should be the most responsible for encouraging such cooperation

Sample:

Sample consists of 42 ICT companies of different sizes, from 10 employees to multinationals. The research was conducted in November 2012 (Figure 3).

Research tools:

Structured questionnaire was developed for this research. It consists of a scale of assessment of seven particles of which two are dichotomous, and five multiple choice.

Factor analysis shows that there is one dimension that measures the achievement of cooperation between IT employers and educational institutions that educate professionals in the field of Information Technology.

Reliability of the measuring instrument Cronbach-alpha showed that the reliability of the whole scale was 0.88.

Results:

The results have shown that 70% of surveyed employers cooperate with state colleges, and 40% cooperate with other higher education institutions (Figures 4 and 5).

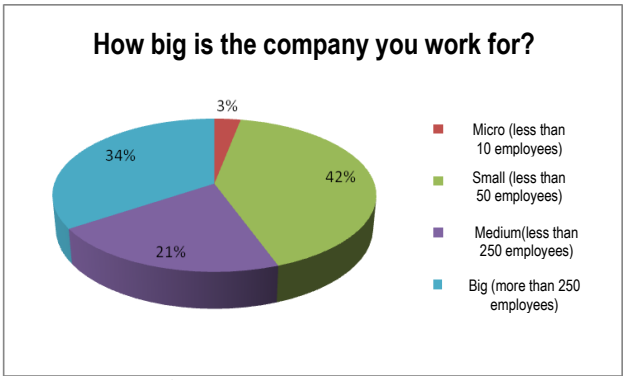


Figure 3 The size of respondents' companies

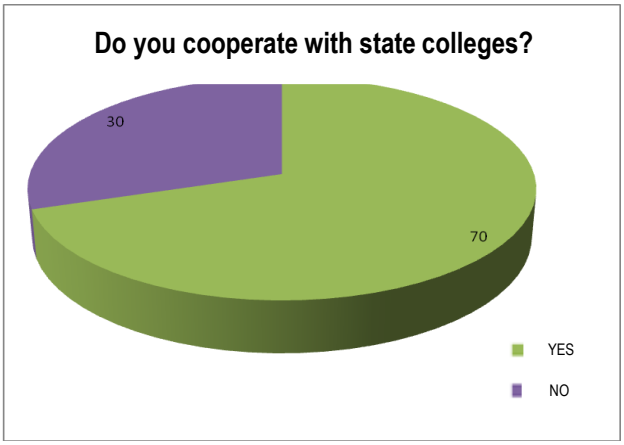


Figure 4 Cooperation between employers and state colleges

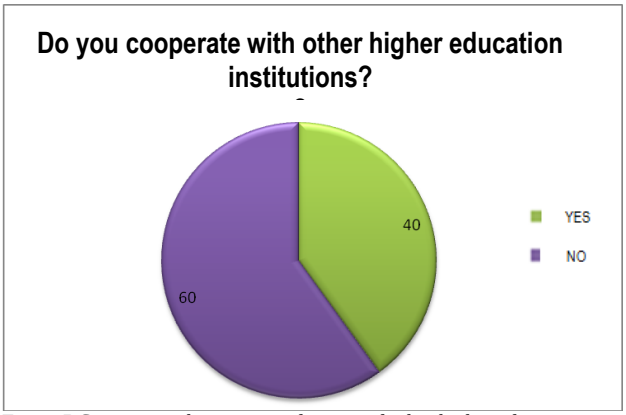


Figure 5 Cooperation between employers and other higher education institutions

Besides general questions asked in this research if employers cooperate with educational institutions, we also asked employers about some details connected to the cooperation. Taken into consideration that previous research showed that respondents emphasized discontent with specific segments of cooperation between educational institutions and employers, which also reflects on later possible employment of students, we decided to research in what ways and in which areas these employers and educational institutions cooperate.

Out of the respondents that cooperate successfully, 30% organizes training and fieldwork, 50% professional training, 27% cooperates in the area of graduate thesis, 33% occasionally organizes student visits in their premises, 24% organizes cooperation in scientific and research area and technology projects, 17% cooperates in the area of e-learning, 25% cooperates in scholarships and 8% financially supports educational institutions (Figure 6).

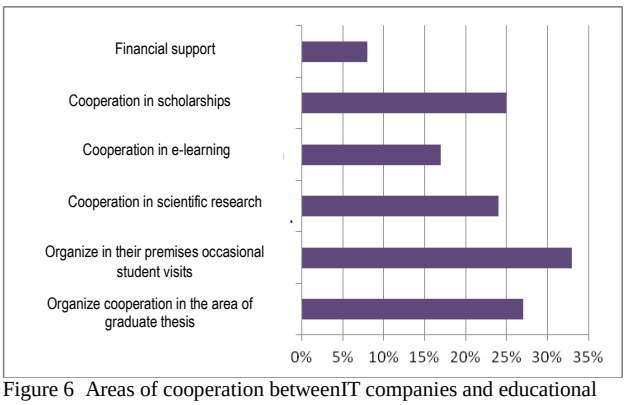


Figure 6 Areas of cooperation betweenIT companies and educational institutions

Furthermore, we investigated what motivates companies of respondents to cooperate with educational institutions. 88% say they can more easily recruit highly qualified employees by receiving recommendations from college professors and 19% are motivated by the fact that in this way they help increase the level of compatibility of educational plan and program with labor market needs (Figure 7).

Figure 7 The reasons of cooperation beteen IT employers and educational institutions

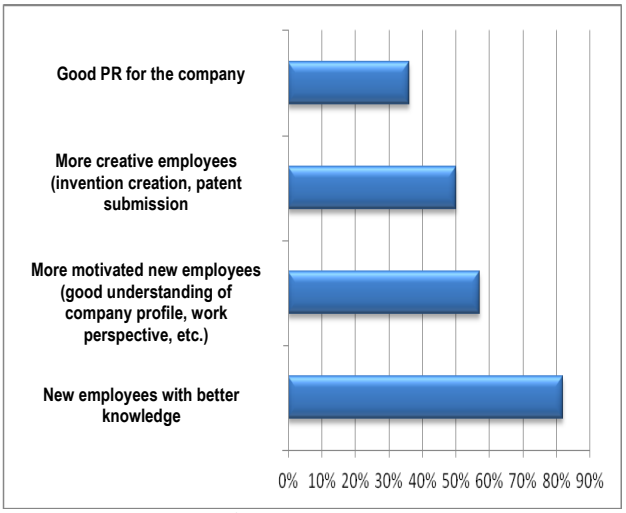


Figure 8 Cooperation benefits

The respondents were asked if they believe that they have so far benefited from cooperation with educational institutions. 82% believe they receive new employees with better knowledge, 57% believe that their new employees are more motivated (because of better understanding of company profile and better work perspective), 50% believe that it raises creativeness of current employees because of participation in inventions and patent submissions in cooperation with educational institutions and 36% believe it is good for the company PR (Figure 8).

We wanted to know what employers think who should encourage that cooperation. Majority, 47% believe that educational institutions should take the initiative for such cooperation. 30% of employers think that they themselves should be the most responsible for initiating the cooperation, the same percentage of respondents think that the initiative should come from economy experts, 20% think that socio-political organizations should take the initiative and only 20% think that that cooperation should be regulated by law (Figure 8).

Furthermore, we asked employers who successfully cooperate with higher educational system if they have their representatives involved in teaching. Out of 42 employers only 3 answered positively to the question!

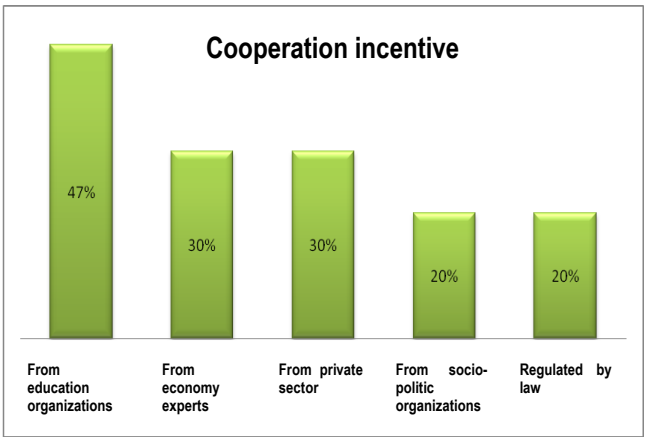


Figure 9 Who should encourage the cooperation between employers and educational institutions

These results are higher in comparison to the average of such cooperation on national level, but in line with world trends to which Croatia is far behind in the matter of public and private investments into education. They are still disappointing, especially regarding financial investments. Even the cooperation that does not require financial investments form employers part is still on a very low level which points to inertia of one system to another and undeveloped awareness of the need for cooperation.

Investigating the motivation for cooperation and answers from employers showed that they are motivated solely by fulfilling their own needs, which means hiring the best students and marketing their own companies, whereas only a small portion is aware that such cooperation should have a much more important task and that is to impact the entire educational system and the economy to come closer in common goal to progress our country and all its inhabitants.

Encouragement for that cooperation, as per most employers, should come from educational system, somewhat less from employers themselves, but interestingly enough, rather small portion believes that such cooperation should be regulated by law (Figure 9).

Are employers afraid that by abiding by the law they should invest much more of their own effort and commitment to achieve this necessary cooperation?!

CONCLUSION

This paper aims to emphasize the importance of connecting employers with educational institutions which educate their future employers. There are many areas of cooperation and commitment to any of the, would raise the quality and the level of knowledge of graduates. Awareness should be raised that both systems have to be active stakeholders to improve such cooperation and not only on the level of meeting their own needs, like specific recruiting or improving financial structure of a certain educational institution, but in a long term strategic way for prosperity of the entire country through increasing the competence of Croatian work force.

One of the biggest challenges for employers is certainly involving economy experts into teaching. Such experts would definitely increase the quality of teaching and bring air of real sector to educational system, make business world closer to students and open doors once they go into employment.

Financial investment from employers in higher education system through different forms and ways is more than welcome and would significantly increase the quality of higher education programs, especially at time when it is not even in the foreseeable future that the state could invest more into educational system.

Training and graduate thesis writing are activities that would not require employers to invest much effort or time, but that would be of great help in training of young people for their future jobs and that would make dull student obligations into something 'alive' and useful.

Educational institutions and employers from ICT sector are on the best way to increase the level of their services from an already high to a higher level.

Opportunities that they provide are diverse, and until such collaboration formalizes in Croatia, it will rest upon their enthusiasm and imagination to help each other and themselves.

REFERENCES:

[1] Bahtijarević-Šiber, F. (1999). Menadžment ljudskih potencijala, Zagreb, Golden Marketing.

[2] Bejaković P (2006). Uloga obrazovnog sustava u postizanju zapošljivosti i konkurentnosti radne snage u Hrvatskoj, Društvena istraživanja, 153; str.401-425.

[3] Hunjak, T. (2008). Obrazovanje informatičara i tržište rada, u: Divjak B.i grupa autora (ur), Ishodi učenja u visokom školstvu, TIVA, Varaždin, str 101-108.

[4] Institut za razvoj obrazovanja, Strateški plan 2011.-2013., Zagreb

[5] Istraživanje o edukativnim potrebama 2012. www.EduCentar.net

[6] Keeley, B. (2007). Ljudski kapital od predškolskog odgoja do cjeloživotnog učenja educa, Zagreb

[7] Marinković, R. (2007). Hrvatski obrazovni prostor i europski utjecaji, u: Marinković R., Nove perspektive u razvoju visokog obrazovanja, Zagreb, A.Z.P.Grafis, str. 25-47.

[8] Marinković, R. (2008). Sustav upravljanja kvalitetom cjeloživotnog obrazovanja, u: Uzelac, V., Vujičić, L. (ur), Cjeloživotno učenje za održivi razvoj, Rijeka, Sveučilište u Rijeci, Učiteljski fakultet u Rijeci, str. 371-377.

[9] MojPosao.net - "Obrazovni sustav šumom, a tržište rada drumom" http://www.moj-posao.net/

[10] Mornar, V. (2008). Uloga gospodarstva u sustavu kvalitete studijskog programa, u: Divjak B.i grupa autora (ur), Ishodi učenja u visokom školstvu, TIVA, Varaždin, str 116-118

[11] Vlada RH. (2008). Zajednički memorandum o prioritetima politike zapošljavanja Republike Hrvatske

The influence of situational factors on accepting e-learning technology by university teachers

Snježana Babić
Polytechnic of Rijeka/Business Deparment, Rijeka, Croatia
snjezana.babic@veleri.hr

Summary – Complex educational environment is one of the key components in considering teacher's acceptance of distance education and can be observed in different scopes in relation to the teacher. This paper deals with an immediate educational environment as a group of situational factors consisting of student characteristics and content/course characteristics and their influence on accepting the technology by university teachers.

I. INTRODUCTION

Even though distance education has been proven as an efficient new education paradigm, its potential is still being underused, as shown by numerous studies.

The quality application of e-learning technology in educational process and creating the virtual learning environment are complex processes and have to be *„rooted in strong pedagogical foundations“* [1].

Therefore, the usability of virtual environments is regarded in two categories [2]: *pedagogical and technical usability*.

There already exist a great number of instructional design models and according to standard ISO/IEC 19796-1 (2005) [3], each type of educational process, including creating virtual learning environment, can be described through the following phases (see: image 1).

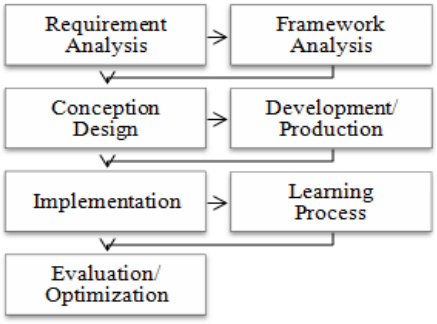


Image 1. Generic process model [3]

University teacher is in the situation to plan the education on different levels: from writing the curriculum to educational units and activities. Regardless of the task level, according to the generic process model, the educational process starts with the analysis of requirements and the work environment.

In the analysis phase, the teacher defines the goals, learning outcomes, analyzes the requirements and limitations of the tasks, examines the institutional factors, as well as those related to the immediate educational environment: content/course characteristics and student characteristics. Based on the analysis results, university teacher opts for the pedagogical model which dictates the choice of e-learning technologies.

Since the university teacher is expected to adapt the educational process to the needs of knowledge society, the e-learning technology should be integrated in the learning process (if there is need for it). However, the research results and practice indicate a problem of insufficient quality of e-learning technology implementation by university teachers.

It is therefore necessary to comprehend the actions of university teachers and determine the factors that influence their acceptance of distance education.

Many authors have determined numerous factors influencing university teachers from various aspects and categorized them in different ways (see: [4]).

Generally, in any type of organization, the acceptance of ICT by individuals is related to the factors which, according to Argawal [5], can be categorized in the following groups: *individual differences, situational factors, social influence, organizational factors, beliefs and attitudes.*

nder the term situational influence, Argawal [5] considers *„the idiosyncratic combination of person and situation that can influence technology acceptance“* while stressing that it is unclear whether the factors including support for the inclusion belong to the category of situational factors.

For that reason, in this paper the term situational *factors* refers to the factors of the immediate educational context divided into two categories: *student characteristics* and *content/course characteristics.*

The theoretical analysis will deal with potential influence of *teachers' perception* of student characteristics and content/course characteristics on university teachers' behaviour and their decision about using e-learning technologies in educational process.

II. SITUATIONAL FACTORS

A. Student characteristics

One of the critical factors in success of the e-learning environment is the student and his/her characteristics [6].

A large number of instructional design models and pedagogical models which list student attributes [7] can be found in literature, for example, according to the Pershing [8], the models indicate the following three major categories of informations:

- *general student characteristics*: grade level, gender, age, job or positions, cultural and socio-economic status, etc.
- *specific entry student competencies*: prior knowledge, technical vocabulary, skills, attitudes, misconceptions, etc.
- *learning styles* (e.g. anxiety, attitude, visual auditory, verbal, logical, etc.)
- *learners motivation*, etc.

The main goal of university system is accomplishing success and satisfaction among students. In distance education the *success of students* is connected to the following attributes (according to Kerr et al. [9]): *self-directed, independent, personally responsible for one's studies, capable of working in an online environment, have developed reading and writing skills, good in time management and motivated for learning.*

According to Sun et al. [10], the key factors influencing *the student satisfaction* in distance education are: *computer anxiety, teacher's attitude towards e-learning, online course flexibility, e-learning quality, perceived usefulness, ease of use and difference in knowledge assessment.*

It is possible to find a large number of studies related to the analysis of *accepting the e-learning technology by students in different contexts*, for example Selim [6] determined the following factors: *student motivation, student technical competency, previous experience with computers, as one of the most critical factors, and interaction among students during team work.*

Student motivation is one of the key factors influencing the student success in distance education. Bubaš et al. [11] conclude that certain students are more affected by factors categorized in group *extrinsic motivation (e.g. being awarded extra points for participating in e-learning course, acquiring a certificate, better grades in the course, etc.).*

Learning style is singled out as a very important student characteristic. To assess the learning styles, different instruments have been developed (e.g. the Grasha-Reichmann Learning Style Scale [12], the Kolb Learning Style Inventory [13]).

Although many authors discuss about learning styles, Wilson [14] emphasizes their importance because the students under the influence of learning styles *„have unique perceptions of their abilities and preferences for learning that may affect their motivation and lifelong learning patterns as well as academic performance levels“.*

Modern e-learning technology and tools enable creating digital educational content supporting various learning styles and present a challenge to the researchers.

Distance education enables student advancing at their own pace with different levels of prior knowledge. Govindasamy [1] states the necessity for analysis of student approaches and information adjustments for groups: *high achievers, average learners and slow learners.*

Based on what was mentioned earlier, we can conclude that a university teacher cannot easily answer the question: *for which of the student characteristics do we need to create an educational process in a digital world and teach by it?*

In practice, different techniques, instruments, interviews, softwares for assessing student characteristics, etc. are being used, but they are not always available to teachers. On the other hand, to use those tools, a certain competencies is required from the teacher.

B. Course characteristics

In the phase of planning and development, according to the ISO standard [3] and instructional design models (e.g. ASSURE, ADDIE, Dick¸&Carey, Kemp), when creating a virtual educational environment it is necessary to identify the following elements connected to the course characteristics:

- identify the educational purpose,
- identify the specific learning outcomes,
- identify the content and analyze educational tasks according to the learning outcomes.

In general, while planning and creating the educational environment the special importance should be given to student understanding of the requirements of content/course defined by *the educational purpose*, *learning outcomes*, *content area* and *learning outcome evaluation.*

Using the *knowledge taxonomies* (e.g. Bloom's taxonomy) learning outcomes are defined and based on these, together with student characteristic analysis, *educational methods and strategies*, as well as student activities which are being determined (for example: team studying, self-regulating studying, simulations, games, problem sloving and similar).

Pedagogy has its foundation in traditional schools, but apart from those kinds of teaching (e.g. behaviorism, constructivism, cognitivism) the literature also lists the use of other theories that have developed from them (such as connectivism) [15].

Depending on the choice of pedagogical model, the e-learning technology can be included in the learning environment through the following components (adapted according to the framework defined by Mishra [16]):

- *educational content* (objective-based course unit; knowledge evaluation: summative and formative)
- *learning activities enable students to participate actively* (for example: forum partaking, team work, research, generating new knowledge)
- *student support* (learner guide, mentor support online, online library, social interaction).

One of the course characteristics is the required level of prior knowledge, and therefore it is necessary to point out the following component of the successful online learning:

- *pre-learning activities* (preparation, student motivation for the environment, knowledge self-assessment, according to Ally [17]).

Anderson [18] states that *„each discipline or field of study contains a world view that provides unique ways of understanding and talking about knowledge"* in which it is necessary to use educational strategies which will, according to Ally [17], answer the question: *how to structure knowledge, motivate students, make students grow as individuals, etc.?*

In the process of using e-learning technology, the following questions are crucial (according to BECTA [19]):

- *Does ICT have a beneficial impact on the results of students, i.e. learning outcomes?*
- *To which extent does curriculum, educational content and knowledge assessment allow the use of ICT?*

Distance education puts students in the focus of education, therefore while considering reasons for using technology in teaching, Ertmer et al. [20] list the following differencies in educational practice (see Table 1).

TABLE I. CATEGORIES OF CLASSROOM PRACTICE (modified according to [20])

Categories of classroom practice	***Teacher – centered Primarily didactic***	***Student-centered: Primarily interactive***
Curricular characteristics	• Breadth – focused on externally mandated curriculum, • Focus on standards, • Fact retention, • Fragmented knowledge and disciplinary separation.	• Depth – focused on student interests, • Focus on understanding of complex ideas, • Application of knowledge to authentic problems, • Integrated multidisciplinary themes.
Technology role	• Drill and practice, • Direct instruction, • Programming.	• Exploration and knowledge construction, • Communication (collaboration, information access, expression), • Tool for writing, data analysis, problem-solving.

It is evident from Table 1 that there are differences between teacher-centered and student-centered practice in e-learning technology application. Both approaches are present in practice, as well as their combination. However, numerous research results show that teachers mostly use e-learning technology for delivering learning material (e.g. PowerPoint presentations, PDF files and similar).

By transforming the teaching material into digital educational content a *multimedia component* is created, commonly called *learning object* [21].

Digital educational materials can be structured as a complete online course, individual lessons or smaller educational units, even presentation of images, video and audio files, graphics, animation, simulation, etc. The idea of learning objects in repositories is based on their flexibility and multiple use [21].

The learning object repositories can be open or closed, and it is important to point out that the teacher can be in position to:

- use someone else's already prepared digital learning materials,
- modify someone else's digital learning materials according to the educational context,
- independently create one's own digital learning materials.

Technology development, especially web 2.0 technology, has enabled different ways of creating and delivering, as well as building new knowledge as a part of the course study with student being able to actively participate in creating digital educational material.

Designing virtual learning environments through project approach can be done by a team of experts which usually includes [21]: *teacher, instructional designer, information technician and administrator*. In the process of designing an online course, one of the teacher's key roles as a member of the team is to be the content expert and he/she is expected to (modified according to [21]):

- identify the available educational resource (literature),
- allow for pedagogical approach to the learning outcomes, educational content and exercises, as well as control and evaluation,
- identify copyright for teaching materials.

Integrating e-learning technology into teaching is a complex and dynamic process which requires careful planning and creating.

Practice has shown that a university teacher most often solves the tasks of instructional design independently, which means that he/she combines the knowledge of pedagogy/didactics and technology while creating virtual learning environments.

It is possible to define unique attributes for every course, thus it is of great importance to determine factors linked to course characteristics and which of them

represent the barrier, i.e. which influence the integration of e-learning technology into the teaching process.

III. THE INFLUENCE OF SITUATIONAL FACTORS ON ACCEPTING E-LEARNING TECHNOLOGY BY UNIVERSITY TEACHERS

Implementing e-learning technology into education is an innovation. Numerous authors have researched various factors which they used to explain teacher's behaviour while deciding about implementing innovation into their teaching process. In their decision teachers have most commonly used, and adapted to their needs, different theories and models of accepting technology and innovation.

As an example, this paper singles out the following applications of theories and models in conducted researches:

- *Technology Acceptance Model* (TAM) [22] is one of the most widely used models whose constructs, perceived usefulness and ease of use, have the proven influence on university teachers when accepting e-learning technology. However, as Schoonenboom [23] states, TAM *"does not make a difference on the level of individual educational task"*. The author questions reasons for the use of different resources of LMS and therefore introduces the construct importance of task into the model and acknowledges that after making the decision about completeing a task, usefulness and ease of use increase the probability that the teacher will use the appropriate LMS resources in teaching.
- *Task-Technology Fit Model* (TTF) [24] is based on the assertment that the probability of technology use and its efficiency will be greater if *technology characteristics* correspond to the *task characteristics on the individual level*. Larsen et al. [25] confirmed that the teachers who perceive the technology characteristics which correspond to requirements of their task use more functionalities of the available technology and develop strong intention for further use.
- *Theory of Planned Behavior* (TPB) [26] introduces the construct *perceived behavioural control* which reflects on the individual's belief about having personal and external resources for a certain behaviour. For example, Sugar et al. [27] state that the teacher's decision to accept the technology is influenced by the attitude towards technology which is based on *personal belief* about the importance of technology in student learning (e.g. [27]: preparing students for their future careers, exposing students to a variety of new technologies, holding students' interest, enabling students to gain additional skills), and as an example of *negative aspect of technology acceptance* they indicate [27] belief (especially Math and Science teachers) that students are becoming too dependent on technology.

Since the construct teachers' beliefs has a very important influence on teachers accepting innovation in their teaching, this paper will now present some of the research results concerning that:

- Kim et al. [28]: *teachers'beliefs about the nature of knowledge and learning* (five dimensions, according to Schommer, 1990: the structure of knowledge, the source of knowledge, the stability of knowledge, the speed of learning, the ability to learn), *teachers' beliefs about effective ways of teaching* (teacher-centered and student-centered) and technology integration are in positive correlation. The authors also confirm that the level of technology use is in significant measure connected with teachers' beliefs about effective ways of teaching.
- Ottenbreit-Leftwich et al. [29]: the results show that *teacher's belief in the value of technology*, seen from the aspect of promoting *student learning*, probably encourages its integration into the teaching process, while listing values to address professional (e.g. [29]: creating customized classroom materials, improving classroom management by engaging students) and student needs.
- Ertmer et al. [20]: states the importance of *pedagogical beliefs*. The research results indicate that *student-centered beliefs* are closely connected with *the change of curriculum* and conducting student-centered practices (authenticity, student choice, collaboration) regardless of the barriers (technological, administrative).

Ertmer et al. [20] state the following *barriers* as the most influential for integrating e-learning technologies into the educational process: *knowledge and skill level, teachers' values and attitude towards technology*. On the other hand, the author [20] states that the highly motivated teachers indicate the following barriers for the use of distance education: *support, country's standards, attitude and beliefs of other teachers, evaluation, availability of resources, as well as the importance of inner motivations*.

British Educational Communications and Technology Agency (BECTA) [19] conducted a research in 5 countries and in their conclusions they state that „curriculum context strongly affects staff expectations of technology, and its use for learning and teaching". Furthermore, they conclude that there is a difference in the use of ICT in relation to the *types of courses* (e.g [19]: ICT is more commonly integrated in Science Technology, Engineering and Mathematics than in the field of humanistic sciences and arts). The emphasis is also put on the lesser use of ICT for the online knowledge assessment. As an example in the part about reflections on the data for England in the framework of Curriculum factors, it is stated [19]: *"My subject area is hands-on practical and has very little scope for the inclusion of ICT"*. It is

interesting to mention the result from Portugal where teachers who are focused on the learning content list the problem of *„availability of electronic learning resources"*[19], as well as teachers from Sweden who emphasize that the development of digital educational resources is the barrier for using e-learning technology in teaching. It is also necessary to mention that the research results indicate the problems of heterogeneity of student attitudes towards technical expertise, availability of internet resources at home, insufficient understanding of internet resources efficiency and other [19].

Gosper et al. [30] point out that while analyzing the curriculum requirements it is important to have a *„clear understanding of philosophical and theoretical underpinnings to ensure that there is compatibility with the chosen software"*.

While inspecting the relation between teachers' beliefs and integration of technology into the educational process, Levin and Wadmany [31] conclude that the teachers who are exposed to innovation in teaching can, under the influence of *„encouragement or even pressure"* alter their teaching styles as well as their basic beliefs about efficient educational practice.

Professional *teacher development* is evident in the development of educational process quality, where student performance has an indirect influence on the change in teacher's knowledge, beliefs and attitudes (see: [32]).

IV. CONCLUSION

The development of distance education depends greatly on the university teachers' approach to creating the quality teaching scenarios. Implementation of e-learning technology depends on the choice of pedagogical model. The factors that influence the teacher's decision about using e-learning technology in practice present a challenge for many researchers.

Considering that the characteristics of key participants in the teaching scenario - student, content and teacher - are specific for different educational environments, the question about the factors which explain the mutual influences is repeatedly being asked.

On the other hand, it should be emphasized that the student and course characteristics change over time and require continuous *professional development* by teachers which is primarily evident in *the change in their plans and programs* [32].

Based on the theoretical research results it is possible to conclude that the university teachers' perception about student characteristics and course characteristics as situational factors influences their belief about the efficiency of distance education or present a barrier (e.g. student motivation, field of study). The importance of *pedagogical belief* [20] should be especially emphasized because it influences the teacher's attitude towards technology, and it is well known that the attitude (positive or negative), according to TBP [26], has a strong impact on the intention to use as well as the use of innovation. On the other hand, it should be said that the perception of situational factor characteristics is also connected to the teachers' knowledge and skills in the field of distance education.

Situational factors present teachers' extrinsic motivation for accepting distance education, which has a mutual influence on inner motivation, so it is assumed that situational factors will have different influence with certain groups of university teachers.

REFERENCES

[1] T. Govindasamy, "Successful implementation of e-learning: Pedagogical considerations", The Internet and Higher Education, 4, 287-299. 92/5/19, 2001.

[2] E. Melis, M. Weber and E. Andrès, "Lessons for (Pedagogic) Usability of eLearning Systems", World Conference on E-Learning in Corporate, Government, Healthcare, & Higher education., (1), 281-284., 2003.

[3] ISO/IEC 19796-1:2005, "Information Technology — Learning, Education, and Training — Quality Management, Assurance and Metrics— Part 1: General Approach", 2005.

[4] R. Agarwal, "Individual Acceptance of Information Technologies", in R. W. Zmud (Ed.), Framing The Domains of IT Management: Projecting the Future Through the Past, Cincinnati, OH: Pinnaflex Press, 85-104., 2000.

[5] S. Babić, "Factors that Influence Academic Teacher's Acceptance of E-Learning Technology in Blended Learning Environment", E-Learning-Organizational Infrastructure and Tools for Specific Areas, Prof. Adilson Guelfi (Ed.), ISBN: 978-953-51-0053-9, InTech, DOI: 10.5772/28682. Available from: http://www.intechopen.com/books/e-learning-organizational-infrastructure-and-tools-for-specific-areas/factors-that-influence-academic-teacher-s-acceptance-of-e-learning-technology-in-blended-learning-en, 2012.

[6] H. M. Selim, "Critical success factors for e-learning acceptance: Confirmatory factor models", Computers & Education, 49 (2), pp. 396‑413, 2007.

[7] J. A. Pershing, H. K. Lee, „Concern Matrix: Analyzing Learners' Needs", Instructional Design in the Real World: a View from the Trenches, 1., 2004.

[8] S. Y. Chyung, „Foundations of Instructional Performance Technology", HRD Press, 2008.

[9] M. S. Kerr, K. Rynearson and M. C. Kerr, „Student characteristics for online learning success", Internet and Higher Education, 9, 91-105., 2006.

[10] P. C. Sun, R. J. Tsai, G. Finger, Y. Y. Chen and D. Yeh, „What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction", Computers & Education, 50(4), 1183-1202., 2008.

[11] G. Bubaš, S. Babić, M. Jadrić, „Motivational Factors Influencing Students' Use of Online Courses: An Exploratory Analysis", Proceedings of 19th Central European Conference on Information and Intelligent Systems, pp. 149-155., Varaždin, Croatia. 2008.

[12] S. Hruska-Riechmann and A. F. Grasha, "The Grasha-Riechmann Student Learning Style Scales", In J. Keefe (Ed.). Student Learning Styles and Brain Behavior (pp. 81-86). Reston, VA: National association of secondary school principals, 1982.

[13] D. A. Kolb, „Learning-styles Inventory: Self-scoring Test and Interpretation Booklet", Boston, MA, Hat/McBer Training Resources Group, 1996.

[14] M. L. Wilson, „Learning Styles, Instructional Strategies, and the Question of Matching: A Literature Review", International Journal of Education, Vol. 4, No. 3, 2012.

[15] S. Babić, "E-learning environment compared to traditional classroom." MIPRO, 2011 Proceedings of the 34th International Convention. IEEE, 2011.

[16] S. Mishra and S. Jain, "Designing an Online Learning Environment for Participatory Management of Displacement, Resettlement and Rehabilitation." Paper presented at the 2 nd Pan-Commonwealth Forum, July 29–August 3, 2002, Durban, South Africa, 2002.

[17] M. Ally, „Foundations of educational theory for online learning“, Theory and practice of online learning, 3-31., 2004.

[18] T. Anderson, „Towards a theory of online learning“. In T. Anderson (Ed.),*The theory and practice of online learning* (2nd ed., pp. 45-74). Edmonton, Canada: Athabasca University Press, 2008.

[19] British Educational Communications and Technology Agency (BECTA),“*Assessing practitioner e-maturity: developing a benchmarking tool to measure practitioner ICT capability in further education: pilot study report.”,* corp creator, *2010.*

[20] P. A. Ertmer, A. T. Ottenbreit-Leftwich, O. Sadik, E. Sendurur and P. Sendurur, „Teacher beliefs and technology integration practices: A critical relationship“, Computers & Education, 59(2), 423-435., 2012.

[21] D. Caplan, R. Graham, “The development of online courses”. Theory and practice of online learning, 175., 2004.

[22] F. D. Davis, F. D., „Perceived usefulness, perceived ease of use, and user acceptance of information technology“, MIS Quarterly, 13(3), 319–340, 1989.

[23] J. Schoonenboom, „The use of technology as one of the possible means of performing instructor tasks: Putting technology acceptance in context“, Computers & Education, Volume 59, Issue 4, December, Pages 1309-1316, 2012.

[24] D. L. Goodhue, R. L. Thompson, „Task-technology fit and individual performance“, MIS Quarterly, 19(2), 213–236., 1995.

[25] T. J. Larsen, A. M. Sørebø and Ø Sørebø, „The role of task-technology fit as users’ motivation to continue information system use“, Computers in Human Behavior, 25(3), 778-784, 2009.

[26] I. Ajzen, The Theory of Planned Behaviour, Organizational Behavior and Human Decision Processes, 50(2), 179-211., 1991.

[27] W Sugar, F. Crawley and B. Fine, „Examining teachers’ decisions to adopt new technology“, Educational Technology and Society, 7 (4), 201-213, 2004.

[28] C. M. Kim, M. K. Kim, C. J. Lee, M. Spector and K. DeMeester „Teacher beliefs and technology integration“, Teaching and Teacher Education, Volume 29, January, Pages 76-85, 2013.

[29] A. T. Ottenbreit-Leftwich, K. D. Glazewski, T. J. Newby, P. A. Ertmer, „Teacher value beliefs associated with using technology: Addressing professional and student needs“, Computers & Education, Volume 55, Issue 3, November, 1321-1335, 2010.

[30] M. Gosper, K. Woo, H. Muir, C. Dudley and K. Nakazawa, „Selecting ICT based solutions for quality learning and sustainable practice“, Australasian Journal of Educational Technology, 23(2), 227-247., 2007.

[31] T. Levin, and R. Wadmany. "Teachers' beliefs and practices in technology-based classrooms: A developmental view." Journal of Research on Technology in Education 39.2 , 157, 2006.

[32] B. J. Fishman et al. "Linking teacher and student learning to improve professional development in systemic reform." Teaching and teacher education19.6 , 643-658., 2003.

The next Generation of University Information Systems – TTI version for XXI century

Boriss Misnevs[1], Nadezda Fila[2]
Transport and Telecommunication Institute, Riga, Latvia
[1]bfm@tsi.lv , [2]nf@tsi.lv

Abstract - The purpose of the paper is to review the evolution of the University Information Systems in the XXI century and to present the future vision and current implementation of TTI University Information System (TTI UIS). TTI UIS supports a large number of information processes and includes a set of subsystems for solving a particular type of university tasks: "Students and alumni", "Time Schedule and Academic Staff Load", "Human Resources", "Accounting", " Hardware and Software Register", and "Library" TTI UIS also provides communication, collaboration and end user support services to system administrative units. The paper discusses existing solutions and problems of TTI UIS and provides a vision of the modern UIS for XXI century.

1. INTRODUCTION

The focus of the given research is the University Information System. Multiple databases successfully help users to solve their everyday tasks – to make decisions taking into account all possible data of study process at the university.

The authors have started research in the area of education services quality control in 1998 [1]. This work has been conducted within the frame of several projects orientated to the development of the Baltic Region such as "Research of Innovative Logistics and Telematics Education in the Baltic Region" [2].

The existing Transport and Telecommunication institute Information System (TTI UIS) was primarily designed at the beginning of XXI century for the collection, storage, searching and retrieval of actual information (as a transactions control system). This system successfully supported the university study process with all tools required for the management of the routine study process:

- scheduling classes,
- the students' enrolment,
- recording results of sessions, etc.

Now TTI UIS supports a large number of information processes and includes a set of subsystems for solving a particular type of university tasks: "Students and alumni", "Time Schedule and Academic Staff Load", "Human Resources", "Accounting", "Hardware and Software Register", and "Library".

For profound analysis of the processes taking place in the university the data should be weeded out, aggregated and presented in the form appropriate for the analysis and use of the University Decision Support System as the part of University Information Systems.

2. UNIVERSITY INFORMATION SYSTEMS: MAIN PARTS

Our University Information Systems provides technical services and enterprise applications to the Transport and Telecommunication institute in Riga and Latgalian Branch in Daugavpils. TTI UIS is committed to excellence in customer service and technical expertise.

Through the development of enterprise applications, TTI UIS supports the University as a whole educational enterprise with systems used by students, faculty, human resources employees, finance employees, and others.

TTI UIS, as shown in Figure 1, includes the following subsystems:

- Human Resources Management System;
- Financial Management System;
- Workflow System;
- WEB site (www.tsi.lv);
- Information System (IS) - intra.tsi.lv;
- University Analytical System (AS-reports.tsi.lv);
- Distance Learning Management System (e.tsi.lv);
- Electronic Research Administration (research.tsi.lv).

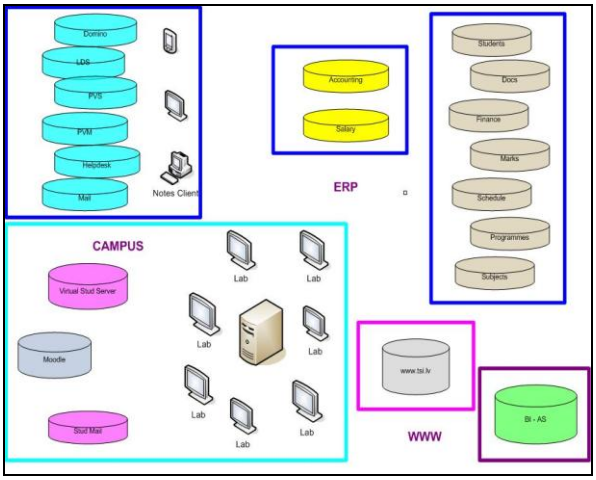


Figure1. TTI UIS architecture.

TTI UIS also provides business applications that support the University's Legal Services department and other departments.

In addition to giving you the tools and applications required, TTI UIS supports you with your technical questions, concerns and issues. Our desktop support staff (helpdesk@tsi.lv) is available to assist you Monday through Saturday from 8:30 a.m. to 10:00 p.m

3. University Information System's Mission and Vision

UIS MISSION: Transport and Telecommunication institute's University Information Systems (TTI UIS) provides the University with a portfolio of secure, reliable, and cost effective enterprise applications. Its enterprise applications supports the following functions:

- Customer Relationship Management;
- Data Mining;
- Document Management;
- Electronic Research Administration;
- Finance Management;
- Human Resources Management;
- Student Information;
- Students Virtual Labs (Stdomino.tsi.lv);
- Distant learning (e.tsi.lv);
- Web Portal (www.tsi.lv);

Additionally, TTI UIS provides communication, collaboration and end user support services to system administrative units.

TTI UIS is committed to exceed or meet service level agreements developed with the University.

UIS VISION: University Information Systems (TTI UIS) will be a valued partner and trusted advisor in the development, delivery and support of enterprise solutions that meet the needs of the University.

TTI UIS provides a big amount of services:

- **Support for desktop hardware:** The repairing and/or replacement of any faulty desktop or laptop that is covered under manufacturer's warranty free of charge, as well as offering assistance in purchasing and installing parts for those computers that are out of warranty. As far as purchasing of new computers, printers, palmtops.
- **Software support:** System Administration department is eligible to receive a great number of software titles and applications free of charge from Microsoft and another software developers . All desktop and laptop computers purchased with university funds have the Microsoft Office, Adobe Reader, Java, MS Visual Studio, MS SQL Server, Microsoft Silverlight, Mozilla Firefox, Winzip, Skype.following software installed:
- **Domain Account setup:** The setup of new user accounts, add them to appropriate security groups, creating a stxxxxx@students.tsi.lv email account and allocating disk space on the network as a personal share drive.
- **Mobile Device Support:** assisting with configuration on wireless and cellular devices such as tablets and smartphones to connect to TTI email, calendaring and contact synchronization.
- **Printers Support:** configuration, installation and troubleshooting, toner replacing.
- **Video Conference Services:** setup of video conferences; troubleshoot connectivity issues, potential audio and video problems.
- **Network Troubleshooting:** performing the first level of troubleshooting of any network problem and fixing those that originate on a client's workstation.
- **Procurement:** desktop support will advise and offer strategic planning to System Administration department on any hardware and software purchases, as well as assist in purchasing IT-related equipment. Department will be responsible for budgeting appropriate funds for those purchases.

4. University Information System's tasks

It is not possible today to control the educational processes at the institute (TTI) without use of integrated University Information System (UIS). The task of such system is the registration of all aspects of the institute's study activity, as well as the management of intellectual, material and financial resources of the institute. Integral part of such system is e-learning system [3].

The TTI information system supports multiple information processes of the university and therefore includes a large number of subsystems each dedicated to the solution of a certain class of tasks.

1. The "Students and Graduates" subsystem is the information subsystem containing information about actual and former TTI students. The subsystem includes two tasks:
    a. «Registration of agreements and payments» - contains data of the Study Department about each student - personal data, information on education and enrolment in TTI; list of orders concerning student in question, list of all agreements with TTI signed by him/her; list of all accounts submitted to him/her by TTI with payment information;
    b. «Registration of progress» - stores and processes data of the Study Department – study programmes; curricula; current and final progress of each student; student assignment to a study group.
2. The "Time-schedule and teachers employment" subsystem is assumed for entering, editing, reviewing of events conducted in the university premises. Generally, these are lectures, practical

and laboratory studies, seminars, consultations, tests, examinations etc., as defined by the specifics of the organization's area of activity. The given information subsystem is designed to simplify search, filter information connected to basic resources of the university as lecture-rooms (computer classes, laboratories), lecturers, student groups planning and optimal premises use.

3. The "Personnel Department" subsystem serves for the storage of personal information about lecturers and employees of the institute.
4. The "Accounts Department" consisting of the "Salary", "Accounts with Suppliers and Recipients of Educational Services", "Fixed Assets" subsystems etc. is for registration of fund and material resources flow in the institute.
5. The "Registration of Hardware and Software Facilities" subsystem is for accounting and management of the hardware and software facilities of the institute used by students, lecturers and employees.
6. The "Library" subsystem is created for registration of library fund at the institute and for registration of multiple library services submitted to lecturers and students as well.

The UIS has been implemented by the employees of the Information Technologies Department of the institute.

Lotus Notes Domino and Applications, as shown in Figure 2, is the main part of the TTI integration platform.

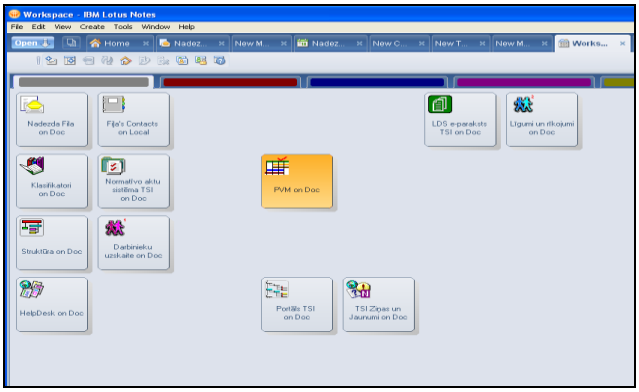


Figure 2. Lotus Notes Workspace.

The application of the modern network technologies and common database within the implementation of computer-aided information system of the university permits to install a simple mechanism of information integration in common information resource and provide opportunity to share information with many users (institute's administration, students, lecturers etc.). Thus the high level of data integrity may be achieved and conditions for the creation of successful functioning of common information environment of the institute appear.

So, the access to time-schedule and other information required by a large number of users is achieved through WEB-interface by any user at any time and at any point from mobile devices.

Also all teachers can see their time-schedule in their e-mail client, such a Lotus Notes or Microsoft Outlook.

5. University Analitical System (AS)

The first version of the University Analytical System (AS), as a main part of TTI UIS, was implemented in the Transport and Telecommunication Institute (TTI) [4] at 2005. The target is efficiency and usability of the TTI managerial decision support system. The algorithms of AS - decision support system and their application in the Transport and Telecommunication Institute are investigated with the aim of processes management improvement and commercial results forecasting. The quantitative and qualitative management of study process, elaboration of managerial decisions and recommendations with the aim of its improvement is particularly important.

The new version of analytical system (AS) for data maintenance and analysis was built on the OLAP principle at 2007. The new technology used by AS allows performing a multidimensional data analysis for decision making purposes. The system uses independent Data Marts, which are constructed using information from existing TTI Information System (OLTP) by filtering and clearing from all historical period of TTI IS operation.

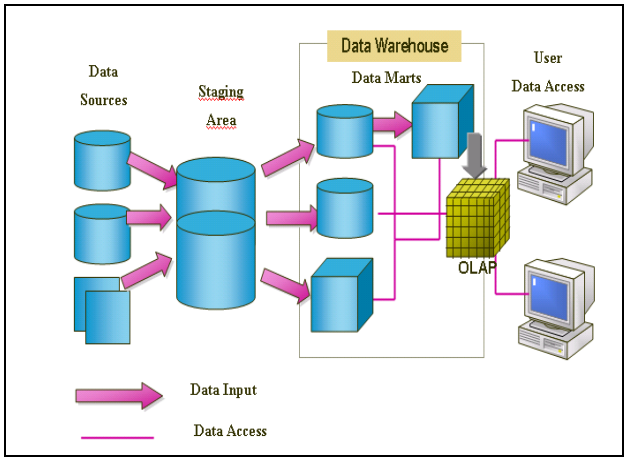


Figure3. TTI Analytical system (AS) architecture.

The OLAP-technologies would help to improve the TTI AS - decision support computer system making use of the existing information systems of the University.

Figure 4. Example of TTI AS (Business Intelligent) interface.

The conducted research and interrogation of top managers of the University has defined the following possible tasks of DSS:

1. Comparative analysis of the study programmes effectiveness (ratio of students enrolled to the students graduated on term) with the aim to define factors influencing on the progress for the study process quality control in the context of the study programmes, study types, study languages.
2. Analysis of the dynamic of tuition fee change in the context of years of enrolment, faculties, study programmes, study types, study languages.

The TTI AS (BI) interface is shown in Figure 4. The example describes distribution of the number of students (in percents from total number) by 5 largest study programmes.

6. DISTANT LEARNING MANAGEMENT SYSTEM

Learning management system (LMS) is responsible for "learning" activities maintenance, while university information systems (UIS) are responsible for handling University managerial activities.

E-learning can be thought of as the learning process created by interaction with digitally delivered content, services, and support. It involves intensive use of information and communication technology to serve, facilitate, and revolutionize the learning process.

Figure 5 shows three main learning models that are used at TTI.

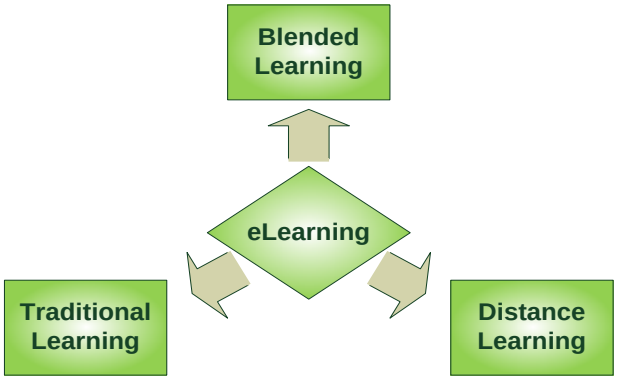


Figure 5. Learning models.

Traditional learning. In traditional learning, students head to a school, college, or other physical space in which to learn. Information and communication technology can enhance the learning process, but is not necessarily included. For example, data show and presentations can be thought of as an implementation of e-learning within a traditional learning institution.

Distance learning. In distance learning, an instructor and students are separated by time, location, or both. Education or training courses are delivered to remote locations via synchronous or asynchronous means of instruction. Distance education does not preclude the use of the traditional classroom.

Blended learning. Blended learning combines multiple models to learning. For example, students in a traditional class can be assigned both print-based and online materials.

The basic e-Learning Management System at TTI is MOODLE.

Developed by an extremely active open source community, Moodle is a popular Course Management System (CMS). That is ideal for creating dynamic online learning communities and for supplementing face-to-face learning. Used in more than 190 countries with support for more than 70 languages, Moodle can scale from a single-teacher site to 50000-student University.

Moodle gives teachers and trainers a powerful set of web-based tools for an array of activities, such as forums, massaging, quizzes, assignments, wikis, blogs, and databases. It has a user friendly interface, as shown in Figure 6.

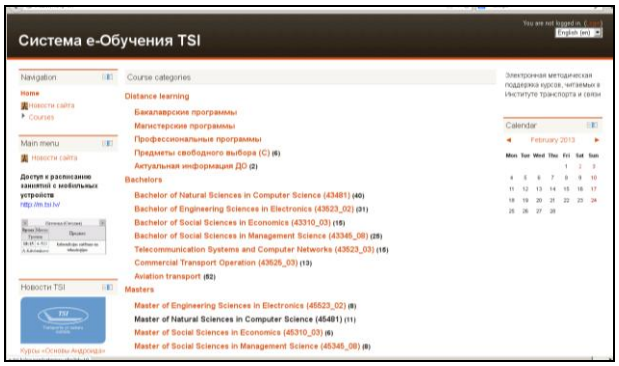


Figure 6. TTI E-Learning System Moodle based.

Moodle is an open source Course Management System (CMS). The name Moodle has two meanings. First, it's an acronym (what isn't these days?) for Modular Object-Oriented Dynamic Learning Environment. Moodle is also a verb that describes the process of lazily meandering through something, doing things as it occurs to you to do them, an enjoyable tinkering that often leads to insight and creativity.

CMSs are web applications, meaning that they run on a server and are accessed by using a web browser. At their most basic, CMSs give educators tools to create a course web site and provide access control so only enrolled students can view it. CMSs also offer a wide variety of tools that can make your course more effective. They provide an easy way to upload and share materials, hold online discussions and chats, give quizzes and surveys, gather and review assignments, and record grades. Let's take a quick look at each of these features and how they might be useful:

We've both spent time researching different CMSs, and we have become fans of Moodle because it is open source, is built on a sound educational philosophy, and has a huge community that supports and develops it. It can compete with the big commercial systems in terms of feature sets and is easy to extend.

Videoconferencing service (else known as video communication) is everywhere used as the tools for e-learning process.

The video communication has also produced a strong impact on education, by this having improved the opportunities of an approach to the most important resources such as education for those living in distant

regions. The introduction of the interactive distance learning in the educational establishment of various countries of the world raises the level of knowledge and unites students, faculties and staff, creates the atmosphere promoting learning, teaching, research and management of the educational establishments.

A videoconference is a unique technology for the distance learning with the presence effect. Lecturers and students are capable to see each other, to look through and discuss various data, documents, and images together in conditions approximated at maximum to a work in ordinary lecture room. The main difference is in the fact that lecturers and students may be in different cities or countries without experiencing discomfort connected with traveling and interruption of their work. In the process of videoconference students may see their lecturer's face expression and gestures language. These things are the most crucial aspects in the process of learning.

Transport and Telecommunication Institute uses videoconferences for the following purposes, including:

- traditional training courses;
- qualification level raising courses;
- connections with lecturers-guests and experts;
- cooperation with other educational establishments;
- exchange of experience and information between lecturers being in different cities and countries;
- virtual traveling to techno-parks, scientific and research centres, museums etc.;
- administrative meetings;
- research conferences conduction;
- special education.

Figure 7. Edge 95 installation in Daugavpils.

Basing of the modern IT a decision about the creation of the videoconference communication infrastructure in the Transport and Telecommunication Institute and in its Latgalian branch in Daugavpils (the city about 250 km far from Riga – capital of Latvia), as shown in Figure 6, has been made [7].

The purpose of the project is to provide students of the branch with a qualitative modern education that will be at the same quality level as the education of students of main campus in Riga.

7. CONCLUSION

As a result of this research it is possible to conclude that: more than 10 years' experience of TTI UIS maintenance allows to define main features of UIS to be implemented as a typical University Information System in XXI century.

The main problem of UIS development is reengineering question of existing heritage subsystems. The question about a platform for all subsystem consolidation still becomes open. As an option for enterprise - based UIS may be considered a kind of "cloud solution", which must be studied separately.

8. REFERENCES

[1] Kabashkin, I.V., Misnevs, B.F., Utehin, G.S. "Using TQM and ISO 9000 Principles in Assuring Education Service Quality", *The Journal of Air Transportation World Wide*, Volume 3, No 2, pp. 70-77 (1998).

[2] Kabashkin, I.V., Misnevs, B.F."Research of Innovative Logistics and Telematics Education in the Baltic Region", *Annual Proceedings of Vidzemes University College "ICTE in Regional Development"*, Valmiera, Latvia, pp. 33-36, (2005).

[3] Misnevs, B.F., Danilov, P.B. "Requirements for the TTI e-learning System". *Proceedings of the XXVI International Conference MIPRO HU*, Opatija, Croatia, pp. 28-32 (2003).

[4] N.Fila, B.Mishnevs, V.Malomins. University Decision Support System: Study of the Implementation. EUNIS 2007, Université Joseph Fourier, Grenoble, French. Proceedings of the 13th International Conference of European University Information Systems, electronic paper No.159 (2007).

[5] N.Fila, B.Misnevs, A.Dunjashkins, E.Abasheva. Simulation Model of Distance Learning Planning and Implementation on the Basis of ECTS Credits. MIPRO 2012, Opatija, Croatia , Proceedings of the 35st International Convention „Computers in Education", ISBN: 978-1-4673-2577-6, pp. 1215 – 1219 (2012).

[6] E-Learning and Management Information Systems Universities Need Both. By Haitham A. El-Ghareeb / September 2009, E-Learn magazine (2009).

[7] Kalachev N., Mishnevs B., Krivchenkov A.:Application of Video Systems for Problem-Oriented Practicums. Bulletin of N.E.Bauman MSTU. Ser. „Natural Sciences". 2010. No 1, Moscow, pp.108-115 (2010).

Preparedness of teachers and learning materials for e-learning

N. Lazić *, J. Klindžić * and M. Odak**
* Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
** Faculty of Philosophy, University of Mostar, Bosnia and Herzegovina
nlazic@ffzg.hr, jklind@ffzg.hr, odakmarko@yahoo.com

Abstract - Time has shown that technology has always had a significant influence on and role in education. The more advanced technology was the greater contribution it made to education. When we look back we can clearly see how great an influence film, radio and television have had on education. The second half of the twentieth century was marked by the results of the use of computers and information and communication technologies in education as well as by extensive research in this field. The intense development of e-learning educational technologies initiated the process of transferring contextual meaning through media. In the paper we will analyze and describe the condition of e-learning at the Faculty of Humanities and Social Sciences in Zagreb, university teachers' competence necessary for this type of teaching, as well as the models and lifecycle of e-learning materials.

I. INTRODUCTION

The transition from the second to the third millennium brought a number of major changes that influenced education and the entire educational policy. It also raised numerous questions regarding the kind of impact and type of transformations brought about by technology, as well as the ways in which it affected teaching, faculty members, students and the quality and form of educational materials. The intense development of e-learning educational technologies initiated a process which enabled the necessary changes and modernization. Over the last couple of decades there have been many debates and studies focusing on the impact of information and communication technology on education.

These transformations are reflected in the social, cultural, economic, and even political context, where knowledge is a strategic social resource. Knowledge society is seen as a just society where human beings as individuals may create, share and use information and knowledge and realize their potential regardless of their economic status. It is not easy to determine the type of arguments in the abovementioned discourses which we may identify as valid and applicable in our conditions. The wide use of the Internet has opened up numerous possibilities for e-learning.

At the very beginning of the introduction of e-learning [1] very few faculty members were willing to adopt this new system due to their lack of knowledge of technology and the need to undergo additional training. One research from that period (2001) clearly shows that there were several faculty members gathered around the projects *Informacijske potrebe korisnika za izobrazbu* (Information needs of users in education) and *Organizacija informacija i znanja u elektroničkom obrazovnom okruženju* (Organization of information and knowledge in an electronic educational environment), who, with the support of the Faculty's IT Division were able to start setting up the Faculty's e-learning platform. Initially these enthusiasts promoted e-learning through various workshops. During this first stage this small group of researchers was closely connected and there was a strong promotion of the new learning system reflected in several papers presented at various conferences. Today there is hardly any faculty member at the Faculty of Humanities and Social Sciences in Zagreb who does not provide some form of e-learning and online educational material as part of his or her course. According to the type of teaching materials that were created during this initial stage it is clear that no pedagogical aspects were taken into consideration in their development. They mostly consisted of lecture notes translated into presentations, accompanied by a list of required reading and possibly some digitized article or a link to a source on the Internet.

Using tools necessary for the development and the sole introduction of e-learning required the fulfilling of certain preconditions, which many in our educational system in this initial stage (2000) were not capable of. This primarily referred to being well-acquainted with and able to use information and communication technologies and the related basic infrastructure. Our first surveys focused on how well our faculty members were acquainted with ICT at that time. After that we aimed our research at students and how well they accepted some of the elements of e-learning, which of the provided materials and resources they used and in what ways. Based on these initial surveys and the results that we obtained from the project *Organizacija i prezentacija znanja u elektroničkom obrazovnom okruženju* (Organization and presentation of knowledge in an electronic educational environment) today we are able to demonstrate how this system developed, how well our faculty members are currently prepared for this type of teaching, what kind of educational material is most frequently provided on the platform and how long is its lifecycle.

The very system that our e-learning platform is based on enabled us to analyze access to the available materials since all types of virtual environments record the frequency and duration of student access to every

individual page of the provided material. However, these elements are not relevant indicators of student work on the material. We learned from the survey forms [2] that many students access the material, print it and then make additional notes and study it in greater detail. We thus came to the conclusion that the statistical data we obtained from the system itself could not tell us much without additional research, which we then conducted using survey forms that students filled out at standard classroom lectures. There were more issues we needed to clarify in order to obtain relevant results, e.g. how long students were able to concentrate in front of the screen as opposed to using paper material (and recent studies [3] have shown that compared to reading material on screen, printed material appears to enable better reading comprehension, and there is the eye fatigue factor – there is less eye fatigue from printed material than from the digital/screen reading materials). Many then answered that they found it more difficult to concentrate in front of the screen, easily wandering off from the platform with an excuse of checking up on something, and in this way 'wasted' more time. In the end we concluded [4] that all these data tell us little about the process of learning and acquiring knowledge by using e-learning educational materials.

Our other goal was to determine the present condition related to e-learning in Croatia and we decided to demonstrate this by focusing on e-learning programs of the Faculty of Humanities and Social Sciences in Zagreb, faculty members' competence at this type of teaching today (ten years after the introduction of e-learning), types of educational materials most frequently offered on the platform and their lifecycle. We chose to use the survey form and interview as methods based on whose results we presented a general overview of the condition in this field at the Faculty of Humanities and Social Sciences in Zagreb. The number of papers focusing on the research of e-learning at the Faculty of Humanities and Social Sciences in Zagreb allows us to state that through them the first and second stages of research - indicating the level of computer and information literacy of both faculty and students, level of acceptance of the system, number of courses included on the platform, number of students using the platform, their satisfaction with it, total usage time, etc. –have been completed. Still ahead of us is the third stage of research through which we have already established that the rate of the Faculty's programs' inclusion on the e-learning platform (as e-courses) is between 20 % and 90 % and that the work needed for developing and preparing materials has increased to such an extent that the enthusiasm and creativity of faculty is no longer sufficient. Research has become divided between technology and pedagogical aspects of e-learning and it clearly points to the need to set up an e-learning centre which would bring together all types of specialists (university teachers of individual courses, computer specialists, teaching methods specialists) in order to ensure quality.

II. THE DIFFUSION OF E-EARNING

As is the case with the majority of innovations, the introduction of e-learning at the Faculty of Humanities and Social Sciences in Zagreb in its first stage (2001-2003) attracted a small number of those who were willing to grapple with challenges involved in the new system. Using the curve of the diffusion of e-learning developed by Zemsky and Massy [5] we presented the stages of the adoption of e-learning based on the number of the available e-courses. In this part we were not yet looking into the number of faculty who were personally devising these courses or the numbers of students using them. Using the abovementioned methodology we divided the courses and statistically processed the diffusion of e-learning at the Faculty of Humanities and Social Sciences in Zagreb. In the period between 2001 and 2003 the diffusion took the following course:

The Faculty had 25 departments, over 120 programs and more than 800 e-courses.

In this period the leading department in the introduction of e-learning was the Department of Information and Communication Sciences, which was not that unexpected since prior to the introduction of e-learning the Department had already run projects focusing on several types of free software and the testing of their use. The significance of the use of tools for the development of e-learning at the Department was reflected in the selection of themes of doctoral theses (2006-2009) focusing on the development of information and communication technologies (Krešimir Pavlina, Mihaela Banek Zorica, Hrvoje Stančić, Sonja Špiranec, Radovan Vrana). Associate departments (early majority) included the departments with a larger number of young associate professors and researchers (Department of Psychology, Department of Archaeology, Department of Phonetics). The category of independent departments, or the late majority, included the Department of the History of Art, Department of Ethnology and Cultural Anthropology, Department of History and Department of Sociology. The category of skeptical departments, or 'conservatives', during that first stage consisted of the departments of languages and literature (ten departments). One specific characteristic of the situation at the Faculty of Humanities and Social Sciences in Zagreb at that time was that the average level of computer and information literacy among the Faculty's teaching staff was very low (10 % of faculty declared themselves computer literate, while information literacy mostly referred to the ability to search and evaluate printed resources). The great majority of university teachers (2002), 70 % of them, were teaching in the way that they were taught, while the technology that they used as their teaching aids included slide projectors, transparencies and photocopied materials. The researchers from the Department of Information and Communication Sciences with the support of the staff of the Faculty's IT Division (initially one professional associate) conducted various surveys and offered solutions which they presented through organized workshops and thus promoted their use (computer, LCD projector, smart boards) (Klindžić, Afrić, Lasić-Lazić in 2002; Mateljan in 2003).

Thus the professional community (faculty), which at that moment – since they were not computer specialists - had the least knowledge of ICT, set about implementing e-learning programs whose application in teaching practice was quite simple. These e-courses mostly consisted of

presentations, several links to reading materials and other related resources available on the Internet. Very few materials were devised through the use of other tools. Libraries, as the underlying infrastructure of education, did not have digital collections, and it seemed that librarians had to provide justification for practically every departure from the established centuries-long library practice. These tendencies, which strongly affected the development of the curriculum, had their roots in a belated and slower development of the library profession in Croatia, caused, of course, by specific and unfavorable economic conditions [6].

Taking as our starting point Zemsky and Massy's claim that at the beginning of the diffusion of e-learning there are innovators - according to them computer enthusiasts - we can note that in our conditions innovators were faculty members and computer specialists, who together developed tools and adjusted them to pedagogical criteria. When we take a look back from a perspective of one decade that has passed until 2012/2013 we can observe that this finding has once again been confirmed. If we analyze the courses which are available today and should constitute the basis for this kind of education, i.e., Information Literacy, Introduction to Social and Humanistic Informatics, ICT Technology, we may see that they are devised in collaboration with computer specialists or are run by faculty members who have better computer skills. The results indicate that the majority of teaching staff at the Faculty of Humanities and Social Sciences in Zagreb have acquired their ICT knowledge and skills through special training courses and self-training, while only 10% of them acquired these competencies through regular education. This 'concept's' only weak spot is its firm orientation towards faculty members' 'free will', i.e. their willingness to implement something or not, as opposed to placing the principal emphasis in this context on the educational institution due to its indisputable role of a real mediator of knowledge and a crucial agent in the implementation of technology in education.

Let us enumerate here only some of the frequently mentioned recent phenomena. The omnipresent digitization of resources, electronic publishing, electronic books, electronic journals, the access vs. possess policy emphasizing accessing as opposed to possessing information, simple and user-friendly interfaces for accessing distributed information systems [7]. During the last decade we saw the surfacing of an entire parallel world of either free or chargeable information sources of high quality offering round-the-clock access which are based on advanced programs for organizing and finding information. Large numbers of documents that can be used in education are daily being published, organized and circulated on the Internet. There is very little important administrative, political, scientific, technical or cultural information that is not available through some public network. The saying that something which is not available on the global network is probably not important is becoming increasingly true. E-learning gave global communication an additional dimension influencing the development of applications and contributed to the reliability of communication and protection of user privacy. The international projects aimed at the digitization of educational materials such as IMS, GESTALT, ARIADNE, EDNA, EASEL, SchoolNET, etc., have deeply penetrated the area of education's supporting system that for centuries was exclusively reserved for libraries.

Electronic publishing is rapidly advancing and arguments in favor of the survival of the printed book must constantly be critically re-examined [8]. The recorded increase of printed material during the last decade, frequently mentioned as a counter-argument to a radical claim that books will gradually entirely disappear, needs to be considered in its overall context. It is a fact that these numbers are almost never compared to an enormous increase in electronic resources in the same period. It is beyond question that electronic publications have an incomparably higher growth tendency.

III. THE LEVEL OF PREPAREDNESS OF UNIVERSITY TEACHERS FOR WORKING IN AN E-LEARNING ENVIRONMENT, TYPES OF E-RESOURCES AND THEIR LIFECYCLE

Under the influence of the previously mentioned changes emerge new educational channels parallel with and independent from those based on printed publications. Traditional skills and knowledge related to providing information and teaching in all fields have maintained their significance and value, but in a different environment, which only resembles the traditional classroom. Some additional knowledge appeared in relation to using and managing technology with the aim of finding, processing and remodeling information for educational purposes, which transcends the use of technology in a traditional education system.

Unfortunately, this new knowledge represents only one segment of the competencies required for the organization and exchange of information at the level of the global network and numerous platforms offering open access to educational materials. The situation we are facing today regarding the organization of information in both classical educational environment and e-learning environment surpasses the domain and standards of any single profession. It may as well be interpreted as the vanishing of boundaries, i.e. walls, of the classroom and the library - as the fundamental support to and source of education – in contemporary technological context. At the same time it may refer to the transposition of traditional skills and knowledge, once exclusively related to the library profession, to the level of universal skills required by university teachers.

The most valuable among these skills are precisely computer and information skills related to the use of teaching materials and their pedagogical modeling for communication purposes [9].

From the context of the previously described changes in the information environment it may be assumed that skills required for future teaching will consist of the skills necessary for developing and using communication and information technology combined with knowledge of the content of a particular course. On the other hand, the quality of e-learning materials will depend on the availability of reusable educational units presented in a

way most suited to a specific content, the selection and amount of value added information and its integration and organization.

Current tendencies suggest that technology should be used in a way which would result in the establishment of better communication between educators and students in stages after the period of instruction, make e-learning material available for subsequent study and research, instead of completely excluding the educator in that subsequent period. When the requirements for the election to educational titles laid down by the Rectors' Conference started including the evaluation of university teachers' work related to devising e-resources, delivering materials on the e-learning platform and setting up e-courses, university teachers' interest and involvement in e-learning grew by 60%. It was also noted that the number of leading departments, or frontrunners, also increased at that time, along with the number of e-courses. The analysis of trends related to different Faculty's departments and their courses led us to the conclusion that this change was also brought about by the demands of students [10], who requested that the teaching material be made available through the e-learning platform. Thus, for example, there is hardly any course now at the Department of Information and Communication Sciences that does not have its e-version.

The research and statistically processed data provided us with the following indicators:

The number of leading departments (innovators) has increased. Despite of a slight stagnation in the development of e-learning, the Department of Information and Communication Sciences is still the leading Faculty's department in this area. There is an urgent need strongly emphasized by all faculty members to set up an e-learning centre which would provide support and assistance in presenting high-quality educational materials and devising reusable digital educational units. In comparison to the initial pattern of diffusion, this leading group – apart from the Department of Phonetics - now includes the Department of the History of Art, Department of Archaeology and Department of Anthropology. At this stage of the diffusion of e-learning the relationship between associate and independent departments remained even, except that both groups increased proportionally. Especially interesting is that skeptical departments practically disappeared. One distinctly negative finding at this stage, when research in this field focuses on the pedagogical aspects of e-learning, is that the Department of Pedagogy was practically the last to take the step towards the introduction of e-learning. A strong interest in the pedagogical aspects of e-learning has been shown by the members of the Department of Information and Communication Sciences. This pattern of diffusion of e-learning is not at all surprising if we take into consideration that precisely the content of the study programs of the departments that featured as leading are especially suitable for devising educational materials in different forms. The same tendency was observed in relation to faculty members' competence related to this type of teaching. The number of faculty capable of using ICT has increased (40%), 30% of them stated that they have completed various specialized courses on using ICT and distance learning platforms, 20% stated that they researched and acquired this knowledge through self-training, while only 10% developed their ICT skills and competencies through some regular education program. We found that there is a different situation concerning the preparation of materials and the actual involvement with the platform. Only 30% of faculty are involved in the preparation and publishing of materials for their courses on the platform and perform all activities related to online teaching. The same percentage of faculty creates and devises their educational materials, but rely on their teaching assistants to maintain and update online content, seeing it merely as a depository open to students. It is interesting that despite being capable of using the e-learning platform, 40% of faculty leave all the work related to e-courses to their teaching assistants while the teaching that they conduct proceeds in the same way as teaching was conducted when they were students. Answers to questions focusing on the type of educational material revealed that 90% of materials are presentations, as was the case in the first research. Recorded lectures together with presentations make 20% of the material on the platform, while films and digital textbooks were reported in around 10% of the answers. It is also interesting that according to 40% of the answers the lifecycle of educational materials was three years, while educational material lifecycle of approximately one year was reported in 15% of the answers. In the remaining 45% of the answers the reported lifecycle of educational materials was longer than three years. Our e-learning platform uses Moodle platform (see Figure 1.), which is very dynamic and enables constant tracking of activities within the system, monitoring of the use of materials by students as well as of any changes of material made by faculty members.

IV. CONLUSION

In order to make further progress in terms of the organization and evaluation in the global context, it is necessary that not only all faculty members possess the required skills, but that the available content be modeled according to pedagogical standards.

Today when practically all members of the teaching and library profession are willing to implement the new technology and adjust traditional tools to the new technological environment, trends in the development of standards for the description of electronic resources (metadata), debates on controlled dictionaries, classification and thesauri all highlight competencies belonging to the field of library profession.

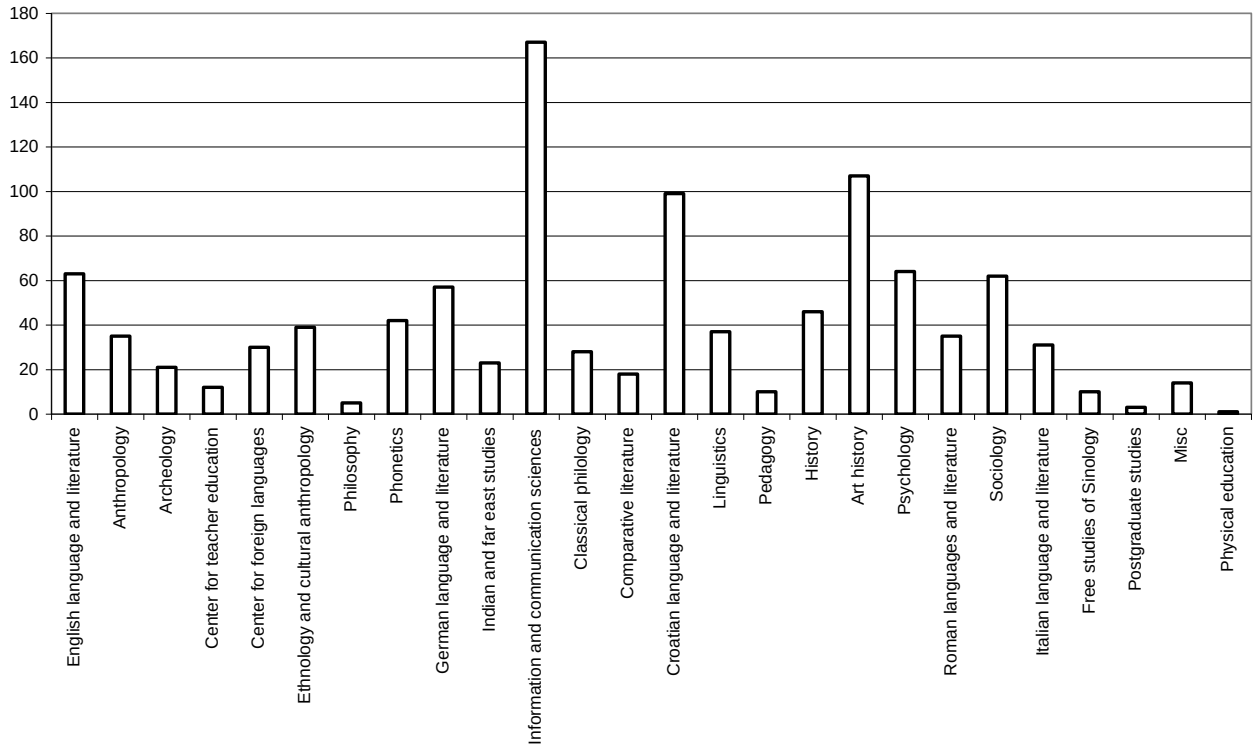


Figure 1. Number of e-courses in our Moodle environment

It has become obvious that traditional knowledge and skills should be taught taking into consideration an entirely new context of the communication of information and its related environment. Information and knowledge have great market value and a prominent role wherever decisions have to be made in relation to managing educational, industrial, financial and social assets. In this context, university teachers' competence at online teaching, devising and distributing educational materials and managing e-learning platforms is relevant for the entire education system. Teacher training and faculty members' ability to keep abreast of the time and changing environment are directly related to the status of education and educational institutions. The results of this research, especially those obtained through interviews conducted with faculty members, confirmed the need to set up an e-learning centre. For the time being the 'centre' exists unofficially, in the virtual world, where it brings together faculty members specializing in various courses and featuring as innovators, pedagogues and computer specialists. However, it is becoming increasingly clear that the services currently provided by the E-learning centre at the Computing Centre of the University of Zagreb (*Sveučilišni računski centar* - SRCE) must instead be provided by our own e-learning centre (that has to be established). The growth of this system, the inclusion of an increasing number of programs and the overall strengthening of the lifelong learning program make necessary a formal and systematic teacher training program. Perhaps the progress that the Faculty's Department of Pedagogy is currently making will have a favorable impact on research focusing on the pedagogical aspects of e-learning

REFERENCES

[1] J. Lasić-Lazić, V. Mateljan, M. Banek Zorica K. Pavlina, "Implementing blended learning in higher education". Proceedings of papers of the third international conference on informatics, educational technology and new media in educationa / Soleša, Dragan (ed.).Sombor : Faculty of education in Sombor, 2006., pp. 212-217.

[2] J. Lasić-Lazić, M. Banek Zorica, S. Špiranec, J. Klindžić, "Using Open Source Learning Management System for Educating Information Professionals" Current Developments in Technology-Assisted Education. Mendez-Vilas, Antonio. (Ed.). Badajoz: Formatex, 2006, 88-92.

[3] H. Jeong, "A comparison of the influence of electronic books and paper books on reading comprehension, eye fatigue, and perception", Electronic Library, The, Vol. 30 Iss: 3 (2012). 390 – 408.

[4] M. Banek Zorica, S. Špiranec, N. Lazić, "Usage of Distance and Blended Learning in Educating School Librarians" Cyberspace, D-world, E-learning: Giving Libraries and Schools the Cutting Edge. Taipei: National Taiwan Normal University, 2007.

[5] R. Zemsky, W.F. Massy, "Thwarted Innovation: What Happened to eLearning and Why". Edupoint, 2006, No. 47.

[6] Lasić-Lazić, J; Slavić, A.. Organization and management of information professionals' education. Informatologia, 33, 3-4(2000), 198-202.

[7] W. Crawford, M. Gorman, "Future libraries: dreams, madness & reality". Chicago : ALA, 1995.

[8] K.H. Veltman, "Understanding New Media: Augmented Knowledge And Culture". Calgary: University of Calgary Press, 2006.

[9] S. Špiranec, M. Banek Zorica, "Information Literacy 2.0: hype or discourse refinement?", Journal of Documentation, Vol. 66 Iss: 1 (2010), 140 – 153.

[10] N. Lazić, J. Klindžić, "Libraries and Learning Management Systems: Peaceful Co-existence or Vital Necessity?", QQML2010: Qualitative and Quantitative Methods in Libraries / Katsirikou, A. (ed.). Chania, Crete : Techn. Univ. of Crete, 2010.

The Strategic Oriented Information Science Education in the Function of Gaining Strategic Knowledge

Marko Čičin-Šain*, Slavomir Vukmirović**, Marina Čičin-Šain**,
*Primorska Banka, Rijeka, Croatia
**Faculty of Economy, University of Rijeka, Croatia
marinacs@efri.hr

Abstract - ***This work presents an analysis of the meaning of strategically oriented information science education and the possibilities of using modern information technology in gaining strategic knowledge. The framework of methodology used in information science education is defined. Variety of knowledge levels are analysed from the viewpoint of the transfer of strategic knowledge. The contents of information science education are grouped according to the levels of complexity which are defined by the degree of the variety of knowledge. The activities of strategically oriented information science education are defined and analysed from the standpoint of creating the guidelines for the development of informatization of education on the principles of creativity and innovation.***

I. Introduction

In the system of strategically oriented information science education a selection of contents should be made and teaching plan models should be created in the area of information technology. Strategically oriented information science education is best organised in the form of interdisciplinary specialised courses which can outgrow into entire specialist postgraduate studies [11]. Information science education in the function of gaining strategic knowledge diffuse from informatics education to entire education process [9].

Robust, intelligent and user-oriented information systems needed in information economics cannot be build upon capabilities gained from the knowledge of traditional data processing. Users and information personnel should learn to think in a new way. Mentors and team work are of crucial importance for gaining strategic knowledge based upon rapid development of methods of object analysis and design and a new era of development of applications that are a result of the initiatives of business re-engineering. The key guidelines for, that is, the fundamentals of, strategic methods in information science education are:

[1] synchronized system notions, modeling and simulation
[2] user-oriented methods of information science education
[3] self-managed team leadership
[4] object-oriented software engineering

In place of the traditional development of specific applications leading towards one solution, object-oriented systems enable multi-dimensional development of applications and create a framework for the introduction of a creative and innovative approach in solving strategic problems, which is based on continuous upgrading of applications on previously created infrastructure and superstructure.

Strategic knowledge enables the integration of the business modelling process and the development of applications. It is based on object notion in line with gathering and developing of knowledge and capabilities necessary for business re-engineering. Structure-oriented methods request defining new business flows and seeking new solutions for each new task. Object-oriented learning assumes the assignment of tasks to objects beyond the business flow limits. By the adoption of new, object-oriented notions, it is possible to adopt a non-sequential way of learning which is useful in many areas, and, of course, particularly flexible for managing the syntax, the grammar and the complexity of object-oriented tools and techniques that match the set task [5].

II. The Goals of Strategically Oriented Information Science Education

Each educational program is aimed at achieving certain specific goals which provide an idea as to what should be learnt by the person that completes that program and what that person should be able to perform upon its completion, or how that person should behave. These goals must satisfy the general criteria of "quality" goals, ie. they must be precise, realistic and measurable [1; 739]. The goals of strategically oriented information science education are aimed at increasing knowledge and abilities of persons that partake in forming the strategy of the development of informatization, that is, in the strategic processes of introduction and usage of information technology.

Information technology is delivering an extraordinary improvement to all areas of human activities, it is opening up the possibilities of improving the quality and the quantity of production and the standard of living, and it is generally broadening the scope of new opportunities in all areas of life [8; 12]. Achieving the goals of strategically oriented information science education is a key factor in mastering the relevant strategic knowledge and skills. In the aim of achieving more qualitative and continuous information science

education, it is necessary to define educational goals. The most important teaching goals in information science education can be defined as the following [7]; [11; II-3]:

- a clear conception of the possibilities offered by computers and computer applications for a given problem area,
- the ability to accept the information science thinking process and the understanding of computer logic in problem analysis,
- an understanding of the logic and the advantages of this new way of solving tasks by the use of modern information technologies,
- the ability to apply knowledge gained at university to the development and the use of applications in dealing with strategic problems,
- the ability to create, sustain and develop one's own user programs (applications) for performing given work tasks,
- a positive attitude about the introduction of information technologies to solving strategic problems and a positive influence on the work environment (for example, groups, teams, management),
- to master the use of a computer and all of its units,
- to master information technology,
- to gain basic information science literacy to the level of solving complex problems in non-structured situations, with the application of information technology,
- to gain and to develop logic and creative abilities in selecting and writing programs in solving non-structured problems, given for a specific problem situation,
- to get introduced to the goals of the society and the dimensions of informatization and information resource management,
- to get introduced to the possibilities and the advantages of network communication,
- to develop the correct relationship towards the use and the protection of programs and data,
- to observe the role of team work in information science.

The achievement of goals in strategically oriented information science education is a fundamental factor in mastering relevant strategic knowledge and skills on a university level. The benefit of such education and achievement of skills in understanding and using information technology can be seen in the example of a transition from a traditional office employee to an employee that uses information technology. Quality information technology education is a fundamental factor in achieving presumptions for such a transition.

III. Grading the Variety of Knowledge in Strategically Oriented Information Science Education

In the process of learning, every individual, indeed the entire society, passes through different stages that can be best recognised according to the variety of knowledge levels [2]; [7].

1. The interpolative level: the mechanical exercising and the practical adoption of knowledge and facts by the use of simple computer programs and commands based on fact-oriented knowledge, mechanical exercises, the practical usage and the forming of concepts and principles using computer tools such as taxonomic keys and identification programs,
2. The operative level: application of facts, concepts and rules in solving specifically structured problems using computer tools such as search devices, data dictionaries and electronic bibliography, application of facts, concepts, rules and principles in solving specifically structured problems (for example, lower levels of case studies explain "why and how" to use computer programs in communication and capabilities of searching for information),
3. The extrapolative level: problem analysis and defining the best solution from the available solutions by the trial and error method using computer simulation and expert systems (for example SimCity) that present results for each of the ideas that are being tested, after problem analysis and defining the possible solutions by the use of computer simulation and expert systems (for example, flight simulation), the intellectual processes learned from the computer are transferred on to the actual performance of assignments and the gain of capabilities.
4. In team or group environment, the problem that has a large number of possible solutions is analysed by working together, so that a solution that is acceptable to all can be found. Computer communications are used in presenting problem cases and in coordinating the group. Different computer tools can be used in the search of the optimal solution. In line with the case study, computer games on the Net are an example of this level of learning.

Variety of knowledge levels are a starting point in the process of gaining new, at times very complex, knowledge and skills. Table 1. shows the variety of knowledge levels [2]; [6; 148-152]. Extrapolative variety of knowledge assumes skills in reorganising data, information and knowledge in the function of non-stereotype problem approach based mostly on divergent notions. The capabilities on the extrapolative variety level are an assumption for solving strategic problems.

Strategic knowledge is manifested in new cases where complex cognitive activity enables the extension of directions and tendencies outside the scope of previously adopted reaction models, with the purpose of determining or predicting certain consequences, effects of follow-up occurrences and similar events under these new circumstances which show specific new links or coincide with the previously studied conditions and known effects. Extrapolation - as intelligent operation of

data, information and knowledge - can be gained by systematically-formed and computer-supported methodology of organising processes of information science education in the function of achieving such levels of knowledge.

Students that stand out due to their high degree of intellectual ability development can further increase the gained knowledge relatively easily and successfully to much higher levels (extrapolative and strategic) and use them with expertise and dexterity not only in known, but also and new cases. But the goal is to bring information science education closer to an average student by improving methodology. Computer-supported methodology of information science education enables gaining of the higher variety of knowledge for average students, that is to those who are not especially talented in programming. It is therefore a condition that each formerly adopted variety of knowledge level is a prerequisite for the next, more complex, variety of knowledge.

Table 1. VARIETY OF KNOWLEDGE LEVELS

Variety level	Learning processes in education
1. Interpolative	Search, identification, classification, discrimination, explanation, systematic arrangement of data, simple computer processing, measurement, presentation
2. Operative	Comparison, differentiation, interpretation, interpolation according to problem logic, organisation of data in new relevant forms
3. Extrapolative	Discovery and documentation of links, elaboration of concepts, identification of input variables, proving, deduction, extrapolation, integration, transfer, predicting, definition of order, reorganisation, synthetising
4. Strategic	Critical discussion, formulation of hypotheses and experimentation, generalisation based on data, improvisation of the process for preparing the explanation models, definition of guidelines for filtration and refinement, derivation and definition of logic arguments

A comparison can be made between knowledge levels in the computer science area and knowledge levels in the learning of the mother tongue, which are much closer and easier to understand to all. Certain analogies can be found if informatics classes are compared to other subjects, for example to mathematics classes (the logic in solving problems) and even to physical education classes (in those moments when the teacher is training the protege for a competition), but analogies to learning how to read and write are mostly expressive. Learning the letters, learning how to read and write is similar to learning certain commands, understanding how the program works and writing the program [4; 5].

It is possible to compare the knowledge levels from the computer science area to the knowledge levels when learning the mother tongue, much closer and easier to understand to everyone. A hundred years ago, only the most intelligent people in a certain country could read and write, while at this day, it is "expected" from, not only the average person, to be literate, but also from those with special needs or difficulties (mentally retarded). This has been made possible through continuous improvement of the learning methodology, which is, in such cases, adjusted according to the specific case in question, and works on the principles of increased repetition and undertaking of smaller steps. The analogies of learning programming and learning literacy in the mother tongue are presented in Table 2 [3].

Table 2. THE ANALOGY OF LEARNING PROGRAMMING AND LEARNING LITERACY IN THE MOTHER TONGUE

Learning the mother tongue	**Learning programming**	**Comparison points**	**Level variety**
Letters, signs, syntax, grammar	Formats, commands, rules	Learning of details	Interpolative
Spelling	Use of software on the basic level	Limited use on the basic level	Interpolative Operative
Reading	Understanding the way that the program works	Passive use of basic elements	Operative
Writing	Programming	Active use of basic elements	Operative Explorative
Creative writing	Advanced programming	Active use of all elements	Explorative Strategic

If viewed separately, basic details in informatics courses are relatively simple. Problems arise when many details need to be linked into a unity. Informatics students go through the spelling stage, when they can comprehend specific details, but they cannot grasp the unity. Memorising does not help in this case, just as it does not help in the mother tongue, because it cannot replace the skill of reading or writing. This is the reason why some students which have very good and excellent grades find informatics to be a problem for them, because they cannot master it by "cramming", and why some average or below-average students interested in computers excel in informatics.

Table 2 shows that a creative approach to writing is based on active and systematic use of all relevant elements as an encircled unity so that a student learning the mother tongue can clearly describe anything that he/she wants to. As an analogy, a creative approach to computers is necessary so that the student can use the computer on the basis of learned facts, and so that he/she can easily deal with unknown or poorly structured case studies in the future and resolve them by applying information technology [4; 6].

Information science education bears similarities and differences to other sorts of education. There are some analogies to those of learning a mother tongue, analogies to a creative approach to problem solving in mathematics, analogies to training how to achieve better results in sports, etc. [3].

There's a great difference in relation to other subjects in the field of student motivation. Students are generally better motivated for subjects relating to information

technology than for other educational matters. It's sometimes the case that:

- in average, male students are more interested in computer subjects than female students [10]
- the students that are excellent in other subjects show average or less than average success in computer subjects, or
- average or less than average students in other subjects simply thrive in computer science.

IV. The Guidelines and the Effects of Strategically Oriented Information Science Education

The basic paradigm of active operation in the area of information technology is programming. Namely, the computer is a universal machine which can be transformed, by an adequate program, into a tool for performing the most divergent scope of tasks. The program provided for the computer consists of a series of steps that need to achieve a certain conceived algorithm (process). Each artefact in information and communication technology has been created by the writing of the program. It is obvious from Table 2 that advanced programming is positioned on explorative and strategic variety of knowledge levels and is comparable to the level of creative writing. Both activities are based on active and systematic use of all elements which are important for solving tasks, which is a key factor in explorative and strategic variety knowledge levels.

Table 3. THE EFFECTS OF INFORMATION SCIENCE EDUCATION AND THE VARIETIES OF KNOWLEDGE

Educational goals	**Variety of knowledge**
Knowledge of details	Fact-oriented
Knowledge of paths and ways of treating details	Interpolative
Analysis, Application	Operative
Evaluation	Extrapolative
Understanding	Strategic

In the aim of developing variety of knowledge, a classification scheme of educational goals in the cognitive area is recommended and elaborated: knowledge, understanding, application, analysis, synthesis and evaluation, whereby knowledge was classified into the following sub-categories: knowledge of details, knowledge of paths and ways of treating details and knowledge of general notions or universal ideas in a certain area, which would correspond to fact-oriented and interpolative knowledge level. Understanding, application, analysis and evaluation of the system are classified into the category of intellectual expertise and skills. Furthermore, operative knowledge level is in the application category, and strategic level of knowledge is in the understanding category.

V. Strategic-oriented Information Science Education Activities

Strategic education assumes independent use of information technology in dealing with unknown problem cases. This level also requires self-management programs for learning with the aim of better understanding of the functioning and the possibilities of information technology. The educational methods that are crucial for gaining skills for solving strategic problems are in the function of gaining experience methods, increasing the depth of knowledge and the generic development in forming problem-oriented concepts. The forming of problem-oriented concepts develops in interaction with mentors under the conditions of new problem situation. On this level, attention is re-focused from the operative details and the analysis of problem elements to the problem process.

In the process of adoption of strategic knowledge, thinking skills broaden outside the boundaries of the existing rules. The adoption of content and the solution for problems is self-explanatory with continuous learning and with a mentor. In adopting strategic knowledge the student gains the expertise to use new or complex applications, to find alternative or creative solutions and to strategically develop the systems. The achievement of this knowledge level is based on the ability to use a wide scope of applications and to rapidly learn detailed application possibilities and their mutual linking in line with the problem that is being dealt with.

Strategic education implies defining tools for documentation of goals, activities, methodology and measurement of the results of learning. Learning prompts up in all situations and it includes assignment of work tasks, training methods and team work.

The objective approach to strategic education with a transfer of knowledge about the usage of tools and techniques, creates an evolution of thinking in the function of adopting new methods for solving problems. Strategic-oriented information science education consists of activities that are mutually intertwined and can be reduced down to:

1. gathering information
2. education
3. training

Gathering information. Information technology is rapidly developing. In the stage of gathering information it is necessary to recognise the possibilities and the functioning of new technologies in line with the needs and possibilities of a business system in which new technologies will be used. It is not only necessary to inform the student, it is more important to train the student so that he/she becomes skilled in permanently being informed about the novelties by using all available sources.

Education. Widening the scope of knowledge, discovery, creativity and innovation are the key tasks of education. Education includes formulation of concepts

and adoption of principles of new paradigms. Conceptual learning in the process of education is a key presumption for achievement of knowledge and expertise during the training stage. As in gathering information, in the education stage it is also necessary to enable the student in self-education, because it is clear that no knowledge in the area of information technology is final, and learning should be a process which should be kept alive throughout the whole career.

Training. Training is a key factor in gaining skills for effective application of information technology in performing business tasks. In the training activities, a work group is defined that creates an object model for solving the defined problem aided by object-oriented computer applications (for example Smalltalk, Access). The starting object area can be set on a web page where the problem is solved by team work method of students and mentors. Throughout training, besides solving particular tasks, it is good to train students to become self-informed and self-educated having in mind that the student will need these processes even after graduating.

Education processes are based on pre-defined educational activities. Education contains three key processes:

1) knowledge formation and acquirement,
2) knowledge organization and
3) technical processing.

Knowledge formation and acquirement is based on designing the knowledge through collecting relevant information from various sources. By knowledge organization, the knowledge acquired in the process of knowledge formation and acquirement is sorted through and systematized, as well as joined with the existing knowledge. Critical information is put in the focus and is being filtered for a closer selection with respect to the theme of the educational matter. The knowledge and information are prepared for processing by being systematized and modeled. The results of processing create new knowledge and information that clarify and solve the given educational problem, but they also refer to collection of additional knowledge and information for further analysis and development within the educational matters being studied.

The transfer of knowledge should be incremental, iterative and innovative. The aim of successful knowledge transfer is the ability of independent performance in a new technological paradigm based on the use of object methods. The key factor of successful transfer of knowledge is in the method of small, development steps. In Table 4 an object-oriented education process in the function of gaining strategic knowledge is shown.

Table 4. STRATEGICALLY ORIENTED INFORMATION SCIENCE EDUCATION ACTIVITIES

Learning stage	**Education processes**	**Variety of knowledge**
Gathering information	Formation Organization Technological process	Fact-oriented Interpolative Operative Extrapolative Strategic
Education	Formation Organization Technological processing	Fact-oriented Interpolative Operative Extrapolative Strategic
Training (implementation)	Formation Organization Technological processing	Fact-oriented Interpolative Operative Extrapolative Strategic

VI. CONLUSION

In the process of computer education, students should be motivated to use computer applications in order to maximise the chances of gaining strategic knowledge. The methods of strategically oriented information science education should be focused on providing suggestions and hints rather than providing direct answers to the questions, as well as on dealing with real-life problem cases, that is, with actual practical situations.

Strategy-oriented education process activities are not sequential, they are intensively multi-linked. The principles of strategic-oriented education are based on incrementality, iterativity and innovativeness. In the process of education it is necessary to define and group the contents according to their degree of complexity. The global division of content according to the degree of complexity can be defined according to the degree of variety of knowledge. For example, in line with that criteria, in a teaching program, teaching contents and tasks that have interpolative, operative, extrapolative and strategic weight should be defined and separated.

The transfer of knowledge unfolds by the method of solving analogous problems. After presenting a method of solving problems and explaining on the example of a solved problem a similar problem is set. When the student solves the problem, another new problem is set although with a lesser degree of similarity. The dynamics of transfer of knowledge can be managed and measured by the method of defining analogous contents. The decrease in the degree of analogy between problems that are being solved can be managed by student's success in solving such problems. Successful solving of problems that have a small degree of similarity compared to the reference problem is proportional to a greater transfer of knowledge.

Informatics education criterias need to leverage according to informatics centers which enable most

quality informatics education and with specifical requirements of this center (6). Teachers have to identify the varieties of knowledge in the forms of subject requirements. By the programs which enable the gaining of operative knowledge, educational process have to diffuse with programs which include the subjects of extrapolative and strategic varieties. That is the key factor of strategic oriented informatics education and capabilities achieving of matters understanding and problem resolving on the extrapolative and strategic level of variety.

References

[1] Bahtijarević-Šiber, F., Human resources Management (org. Croatian Management ljudskih potencijala), Golden marketing, Zagreb, 1999.

[2] Bloom, S., A Taxonomy for Writing Cognitive ObjectivesSeven Levels of Complexity, http://gfb.cas.psu.edu/EdTech/writtax.htm#top

[3] Čičin-Šain, M., The Teacher of Computer Science - the Analogies and Differences to other Teachers, MicroCAD 98, Miskolci 1998., str. 11-18.

[4] Čičin-Šain, M., The teachers' role in preparation of students for a creative approach to computer (org. Croatian: Uloga nastavnika u pripremi učenika za kreativan pristup računalu), 21. međunarodni skup MIPRO 1998., Zbornik radova Computers in Education, Opatija, 1998., str. 4-8,.

[5] Fingar, P., Object-Oriented Knowledge Transfer home1.gte.net/pfingar/knowtransfer_h.htm

[6] T. Grgin, The School Assignment of Knowledge (org. Croatian: Školsko ocjenjivanje znanja, Naknada Slap, Jastrebarsko, 1999.

[7] Gugić (et all), The manual of methodic of teaching computer science (org. Croatian: Priručnik metodike za nastavu računalstva i informatike), Pentium, Vinkovci, 1997.

[8] M. Marinović, M. Čičin-Šain, S. Čičin-Šain,The Pre-school child and the computer (Org. Croatian: Predškolsko dijete i računalo), Računala u školi, 2000 MIPRO, XXIII međunarodni skup, Opatija, Hrvatska, 2000., str. 11-13.

[9] The team working for the president of Croatia, e-Croatia, The proposal for the informatization strategy in Croatia (org. Croatian: Radna skupina Predsjednika Republike, e-Croatia – Prijedlog strategije informatizacije Hrvatske), http://www.predsjednik.hr/radne_skupine/strategija_i_hrvatske/datoteke/strategija_i_hrvatske.html

[10] Skyllingstad, D., An Exploration Into The Lack Of Female High School Students In Computer Science, www.cs.swarthmore.edu/~newhall/Sky.pdf

[11] The strategy of development of Republic Croatia: Croatia in 21 sentury (Org. Croatian: Strategija razvitka republike Hrvatske: Hrvatska u 21. stoljeću; Projektni zadatak: Informacijska i komunikacijska tehnologija) http://www.open.hr/buducnost/vlada/strategijav50.pdf

[12] S. Vukmirović, J. Šehanović, R. Zelenika, M. Žugaj The importance of information staff in the office information system development , (org. Croatian: Značenje informacijskog djelatnika u razvoju uredskog informacijskog sustava), V međunarodni simpozij, Zbornik radova, Informacijski sustavi '94, Fakultet organizacije i informatike, Varaždin, 1994.

E-learning controversy in practical application

Tomaž Bratina*, Matjaž Duh*, Marjan Krašna**
*University of Maribor, Faculty of Education, Maribor, Slovenia
**University of Maribor, Faculty of Arts, Maribor, Slovenia
tomaz.bratina@uni-mb.si, matjaz.duh@uni-mb.si, marjan.krasna@um.si

Abstract – In last few years we witnessed huge growth in e-learning materials production. Hardware development and its availability managed to shatter the computer world. Last year more tablet computers were sold than all others computers combined. Smart phones are in the reach of children. Huge market was created for software application and e-learning materials. Groups of experts cooperate in the production effort. On the other hand we see that advancement is mostly evident in the technical perfection and much preferred aesthetic view still lag behind. It is good that the amount of text is reduced and multimedia elements grow in lower levels of education. But animations, video clips, graphs and drawings are still below desired aesthetical quality. In the natural sciences quality of technical drawings and visualization achieved desired aesthetical level but the same is not true in social sciences. Research conducted between digitally literate users confirms the desire to enhance aesthetical quality of e-learning materials and review the gap between generations. Students generally have mixed opinions about lecturers' presentations' aesthetical value.

Keywords: e-learning, aesthetic, visalization, learning material desing, multimedia

I. INTRODUCTION

Since e-learning materials generally are audiovisual interactive learning materials the visual communication is paramount. The combine aesthetical effect of speech, text and image is required. The aim of learning materials is to stimulate intellectual and emotional response in the learners (Hembree 2006). Therefore the design process of learning materials should include this requirement. In the process of learning we need to know not only how student acquire information. All students have exactly the same process of receiving information but their interpretation varies. A two stage process occurs from stimuli to perception: (1) sensation in the sensory organs and (2) detection that means selection, organization and interpretation of stimuli. The perception of individual sensory inputs converts the content to a specific meaning "*Perception is therefore psychic process in which we become aware of or we recognize a thing, object or phenomenon in order to connect our sensation with our previous knowledge.*" [1]. Such perception is also present in the recognition and understanding of computer screens of e-learning materials. Three components that are closely related are relevant in this process: attention to stimuli; organization of sensations; and their interpretation. According to these findings the design of learning materials should be such that organization of sensations enables direct interpretation. Relations between visual elements have high influence to their perception. Visual elements in e-learning materials should draw immediate attention of the learners. Criteria for quality assessment of web pages [2] should be applied to e-learning materials also. According to the quality assessment of web we could grade e-learning material on topics of:

- content;
- usability;
- user support; and
- visual sensation.

In the development of e-learning materials author mainly regard only first two topics. The third topic (user support) is limited to the user manuals but the last topic is practically not regarded. Visual sensation should harmonize colors, images, and content. The factors that affect organization of visual space into separated visual domains are: [1]

- proximity law (visual objects are grouped together based on proximity)
- similarity law (if visual stimuli becomes from many objects then we start to group them based on their similarity like: content, color, shape)
- close shape law (it is natural that closed shapes is easier to perceive)
- common fate law (parts that have the “common fate” form a unit much easier than others); and
- experience law.

Multiple approach and methodologies for assessment of e-learning materials are known based on reviewers and size (number of screens) of e-learning materials. Like aesthetic assessment strategies we can use: Linear assessment (reviewers assess items one by one), (2) Map assessment (a viewer has all items in front of their eyes), (3) Sort assessment (viewer decides between two items), (4) Mix assessment (a combination of the previous assessment types). [3].

II. E-LEARNING MATERIALS DESIGN

Printed and e-learning materials require different design approach. Text and images on the computer screen have different more indirect impact on learners than textbooks. Images are easier to manipulate and adapt to different circumstances. Principles of printed materials design are not applicable to the web materials. On computer there is not turning the page but scrolling. Up until the tablet pc where touch and drag on the screen represent the same feeling as we experience on books. Digital technology evolves new aesthetic where basic

principles of contrast and tempo still applies. [4] Tempo and contrast are essential for preserving readers interest with their diversity. This is particularly true for learning materials were a lot of images are almost mandatory. Guidelines for reader to different information should be considered carefully. In the long unbroken text in the e-learning materials reader needs more time. But on complex screen with text and images of different sizes reader may just scan the page and stops at specific area where information of his interest are. Content and space dictates tempo but contrast guides it. If design of e-learning materials wants to dictate the tempo visual highlights are added in form of font size and images or their unusual relationship. To achieve good tempo with contrast and visual rhythm on multiple computer screens one have to be aware that too complex design becomes counterproductive and sequence of similar arrangements becomes boring. The risk of decreased attention becomes high. [4]. Therefore graphic design of computer screen of e-learning materials should be focused on best transfer of idea or message to the student. E-learning materials should be visually appealing and clearly understood by students [5]. In the graphics design even contemporary products follows existing pattern, shapes and genre therefore it would be wise to follow the same principles even in the design of e-learning materials. Traditional patterns represent visual communication structure - visual language that evolves and expands. As each verbal or written statement relies on its grammar for correct interpretation visual design should rely on best-proven practices. [6].

III. E-LEARNING CONTENT VISUAL DESIGN

Authors of e-learning materials have important decision to make in the design process. Should they use photos, illustration, text or all three of them? "*The harmonious appearance is a trademark of good product design, its final appearance highly dependents on proper use and integration of images into a whole*." [4]. Digital art techniques erase boundaries between photographs and illustrations. To use realistic photos or illustrators interpretation or their combination requires thorough decision. In the e-learning materials visual elements with highest degree of quality information transfer should be used. Once this can be achieved with video others by photo or high quality illustration. Metrics in this account should be the visual quality and transfer of information. In cases of many graphics elements on one computer screen their size becomes relevant. Larger image attracts more viewers and get more attention [7]. Illustrations values in contrast to the photos are also influenced by their colors and not just position. Style of illustration should be adequate and consistent throughout the e-learning materials. Size, direction and movement on the illustrations should match neighboring content. In its primary role images (photos or illustrations) are symbols and symbols are language of common understanding between author of e-learning materials and students. Symbols can be abstract forms, geometrical shapes; letters of alphabets; numbers; or anything that carries some meaning [8]. The use of Lightbox should be considered in the position of photos and video. Lightbox is a simple script that enables image enlargement on mouse click but the rest of the computer screen darkens when users preview images or video.

IV. THE IMPORTANCE OF TYPOGRAPHY IN THE SELECTION OF E-LEARNING MATERIALS

Typography is an important element in the visualization of e-learning content. It is the art of selecting and editing fonts. But on the other hand typography is the art of fonts usage in graphic design to achieve a desired effects. Fonts are usually taken from a collection of files that can be found in any computer system. The authors of e-learning materials usually select fonts from the most commonly used fonts. Fonts are divided according to the shape and meaning. Some are more formal and elegant, others simple and relaxed. Ones are reminiscent of the past, while others are modern. Font selection is very important since different fonts provoke different reactions in people. [9]. Font designing actually means the creation of signs that compose words [10]. In the e-learning material design selection of fonts opens different aspects: font face, font size, color, line spacing, … It is mandatory that we use right typography in e-learning materials. This is not just aesthetical requirement but also the technical limits of devices that are used to visualize the computer screen. Typography is a visual language used to record speech or text [10] The role of typography is to effectively transmit the written messages. The selection of different font types, sizes and color on the same screen presents the visual confusion - screen fog, poor legibility and readability [11]. Legibility is the degree of distinction between signs. These signs are designed to represent clear and concise distinction [12]. Two distinct design values have high impact to the legibility: space between letters and line spacing. In the e-learning materials legibility means recognition of information value of content represented with alphanumeric signs (words, statements and paragraphs). Readability of computer screen depends on space between words and sentences; margins and paragraph spaces; but not so much on individual letter or signs. [11].

V. SIZE AND BALANCE OF TYPOGRAPHY

Font size in e-learning content must be adjusted to the display size. In the Eyetrack III study results shows that smaller size font encourages users to read text in more details. Larger size fonts on the other hand enable fast overview of content. It is recommended that the headings should be 1,5 times larger than body text. Today it is no problem preparing justified aligned text but interestingly the research from 1997 shows that left alignment is most readable [7]. Even others readability research confirms these findings.

Environment and space that letter occupy needs to be properly defined. Letters need white space. Too much white space diminish information but too little white space choke the information. From the beginning of printing of text we have problems with the letter spacing. Each letter occupies its space but sometimes positions of two neighboring letters are distracting. Optical alignment can correct these distractions and bring harmony between words and lines. [13] [14]. With optical alignment we can achieve dynamic spacing — better position of white space between letters. Kerning is available for better position of

white spaces between some letters but it comes to effect only in large letters. For legibility and readability space between words is also important. In some research it is suggested that space between words should be at least a quarter of a point or in other words as wide as letter I with the white space [10] [15]. On other research suggested white space was as wide as space of minuscule n [12]. But in the end viewers perceive a computer screen always as the combined composition of elements that influence the quality of e-learning materials. E-learning materials inevitably have a lot of text. The composition of the text should help readers and makes reading more fluent. This is particularly important in the transition from one line to another and paragraph starts and endings. The text in e-learning materials should be uniform from the beginning to the end. Vertical alignment between lines should enable reader to follow the text in the transition from one line to another. [12]. Typography does not only reflect the shape of letters for good readability but also transmit another information like mood, posture and tone. [16].

VI. COLORS, ELEMENT POSITION AND VISUAL COHERENCE OF E-LEARNING MATERIALS

In the e-learning materials design use of colors and their harmony need to be considered. Colors can raise certain emotional response without the need of text or speech and those emotional responses can be predicted [17]. Under assumption that e-learning material wants to achieve best possible educational and upbringing efficiency the topic of colors cannot be easily dismissed. From the start of the design process of e-learning materials it is wise to define color structure into clear and logical description. Authors must fulfill the colors' messages to combine with the intended content [7]. On the other hand authors should also create desired mood, message transfer and provoke desired reaction [17]. The final goal of creation of color scheme is the effectiveness of color combination for knowledge retention.

In the concept of e-learning materials screen design the position of visual elements is important. Eyetrack research results were clear. Users have specific pattern of scanning the computer screen. Usually this pattern has the shape of letter F. At the beginning users focus their attention to the upper left corner of screen. For a few moments eyes rest on this position and then starts moving to the right. After first imaginable line is scanned eyes follows to the next line below first scanned line and continue to the right. In the next stage eyes continue to the bottom of the page – letter F is complete. From this point on scanning of the eyes follows the bottom of the screen to the right and up to the upper right corner [18]. Knowing this rule it is wise to position elements on the screen according to this pattern.

Let not be misled by previous findings. There is also another fact we need to obey in the design of computer screen – consistency. Consistency leads to the uniform style that needs to be fundamental design part of any e-learning materials and computer screen in general. Therefore a healthy conservatism in the selection of shapes; sizes; and white spaces in text is advisable [7]. These rules can be generalized to the position of empty space, clip art, style and size of any screen elements. A quality design e-learning materials needs to harmonize conflicts between consistency and diversity. The aim should therefore be the creation of high-quality e-learning material that is internally consistent, but not boring. To achieve the internal dynamics of the computer screen it is possible to use different frames, lines, shadows and gaps at the top. Emptiness, which is not necessarily the white empty space without text or images, is one of the most underrated design elements. With the empty space we achieve contrast while the reader's eyes rest there. [7].

VII. RESEARCH METHODOLOGY

To acquire the insight to the students' aesthetic perception of learning materials we conduct the survey. The survey was performed on the sample of 60 students from the study discipline of elementary education in Slovenia.

We were focused on the level of student's estimation of likeability of learning materials used in different courses. Under the term "like" in its all variations we assumed the student's general impressions of all learning materials used during the courses. Teachers with different proficiency level of aesthetics learning contents design prepared the learning contents.

The 5-point Likert scale with values from "*I totally dislike them*" to the "*I like them very much*" was used for student's emotion estimations.

Since it was not our intention to get results from any particular study course we rather joined courses to their natural distinct disciplines: Social sciences, Natural sciences and Fine arts. Later we also analyzed the impact of the particular study course to the general emotional assessment of the particular discipline.

The studying of the impact has undergone the assumption that assessment of the aesthetic perception is a very subjective matter and may be influenced by many factors or stimuli, like colors, artifacts, complexity etc. [19]. Based on our experiences the aesthetic perception can also be affected by the respondent's mood. We assumed that the difference between the levels of estimation *like* and *like very much* cannot be reliably separated using our researching method therefore we joined them together.

VIII. RESEARCH RESULTS

TABLE 1: SOCIAL SCIENCES' LEARNING MATERIALS LIKEABILITY EMOTIONAL RESPONSE.

Social sciences courses	%
Totally dislike	5,6%
Dislike	13,9%
Neither like/nor like	30,7%
Like	41,2%
Like very much	8,6%

In general the likeability of the learning materials for social sciences is good. But nearly the third (30,7%) of

students have no particular emotional response and remained neutral. The reason for such results may be the content, which is generally perceived as difficult for nice presentation. Around 14% of disliking assessment is a good indication for our possible explanation. Historical background of teaching of these disciplines and new "out of the box" thinking that requires a lot of effort is probably too much for teachers to prepare the learning contents in a way to raise pleasant aesthetic impression of the students. It would be wise to take the advantages of accessible multimedia elements and spice up teaching process.

TABLE 2: SOCIAL SCIENCES' LEARNING MATERIALS LIKEABILITY BY TEACHING SUBJECT

Social sciences courses	Like & Very like
Pedagogy	17,3%
Psychology in Education	15,8%
Selected Topics from Social Sciences - Geography	14,3%
Didactics	13,5%
Introduction to the Methodology of Pedagogical Research	11,3%
Developmental Psychology (middle childhood)	10,5%
Sociology of Education/Moral Education and Ethics	10,5%
Selected Topics from Social Sciences - History	6,8%

The impact portion of the particular social sciences study course definitively brings the confirmation about specific nature of social sciences learning contents. The impact portions of all study courses are very similar with slightly higher impact of learning contents in Pedagogy. Since the students are in the disciplines of pedagogical studies the result is almost expected. The lowest level of pleasant emotional response from the learning contents (around 7%) is found in history course. This may probably be influenced with the notoriously large number of facts required in this course (dates, timelines, persons etc.). Again is to emphasize that preparing of the aesthetically likeable learning contents at any field of social sciences is not a simple task.

TABLE 3: NATURAL SCIENCES' LEARNING MATERIALS LIKEABILITY

Natural sciences courses	%
Totally dislike	1,8%
Dislike	13,7%
Neither like/nor like	22,6%
Like	48,2%
Like very much	13,7%

The learning contents in the field of natural sciences are easier to present since they are in most cases connected with real life experiences. The nature itself has many aesthetically pleased phenomena or objects. The learning contents are far easier to be presented with multimedia and ICT. The likeability of the presented learning contents for natural sciences is estimated reasonably high. The neutral estimations dropped to less than a quarter (22,6%) in contrast to the social sciences. In general the trend is positive and students have much better aesthetic impressions.

Table 4: Natural sciences' learning materials likeability by teaching subject

Natural sciences courses	Like & Very like
Physics	31,7%
ICT	22,1%
Mathematics	18,3%
Chemistry	15,4%
Biology	12,5%

It is certain that the influence of the application of multimedia and ICT in the process of preparation of learning contents consecutively raises the level of likeability and/or aesthetical pleasure. In the Physics the multimedia elements and ICT have long tradition and they are required and widely used. Therefore the highest impact portion is the logical consequence. The teaching subject's objective of ICT is to reveal the possibilities of the implementation of ICT in learning process. So this teaching subject's higher impact portion is not surprising. The three other important natural science's teaching subjects were still engaged with their impact portion but a bit lower than expected.

TABLE 5: FINE ARTS LEARNING MATERIALS LIKEABILITY

Fine arts courses	%
Totally dislike	0,0%
Dislike	0,0%
Neither like/nor like	13,3%
Like	53,3%
Like very much	33,3%

The level of knowledge and expertise of lecturers in understanding and using the principles of aesthetics is clearly expressed in the field of fine arts. Students assessed their learning materials as pleasant or very pleasant in 86,6% (like and like very much). Just about 13,3% were neutral and not a single assessment was negative. These results are very encouraging and indicating that students are aware or sensitive enough to the learning material's aesthetical appearance or aspect.

TABLE 6: FINE ARTS' LEARNING MATERIALS LIKEABILITY BY TEACHING SUBJECT

Fine arts courses	Like & Very like
Didactics of Fine Art I	36,5%
Didactics of Fine Art II	34,6%
General pedagogy for fine art	28,8%

The impact of the courses didactical of fine arts are very high and more than two thirds (71,1%) of students find these learning materials as pleasant. Such high impact is very satisfying and bring us to the conclusion that knowledge and expertise in the application of aesthetics principles will be transferred to the future generations of teachers. Regarding to the results of social and natural science's study course appropriate training of lecturers is needed. Such workshops are also advisable to the active teachers at primary and secondary school level.

IX. COMPARISON TO OTHER STUDENTS

Primary research group was students from elementary education. These students have study courses from all three observed disciplines: natural sciences, social sciences and fine arts. They can compare learning materials from different lecturers and make better assessment. To verify their finding we have also made roundtable with the students of physics and students of general pedagogy. First are strictly natural science students but later are completely social science students. Since we could only get 9 students from 2nd year of physics and 14 students from 5th year of general pedagogy we could not make the same test as we could with the students of elementary education hence roundtable. But we got some interesting feedback from them which generally break apart these two groups.

Students of physics are much more visually oriented and graphs, schemas and concise short text is much more favorable in their eyes. Their lecturers often follow these rules. On the other hands students of physics are much more fond of classic blackboard for writing equations doing math. They state that math cannot be thought with the slides. Slides are mostly regarded as images but writing equations on the blackboard is the "video" that is slow enough for them to copy and process the mental iterations of lecturer.

Students of general pedagogy are older and more experienced. They are not very easily impressed with the nice electronic presentation. They do like it but they could easily sacrifice the nice screen composition to more textual computer screen as long as they do not need to search for other study literature. They are opportunists by trade. The term like is just as a strawberry on the cake - it looks good but the cake is good even without it. We were little disappointed with their attitude but on the other hand we tried to think what we would like to have on the LLL courses. Our own time constraints actually bring us much closer to students of general pedagogy than others but we are still more graphic than textual oriented.

From these small groups we could make some interesting conclusion. Aesthetical pleased learning materials have bigger and better effect on younger students. Experienced students do not depend so much on aesthetically pleased learning materials. Some analogy can be found from the past research of neuropsychology (Jaušovec, xxx) where multimedia learning materials have best impact to the average and slightly below average students but the best students do not need them. If we make comparison between years of study that present experience and better experienced students are the best then the analogy is confirmed.

X. CONCLUSION

The likeability of learning material varies depending on the areas and learning contents. Specific learning contents are not easy to be prepared to fulfill the majority of aesthetics principles and guidelines. The learning contents in the field of social sciences is not basically suitable for be nice. Much easier is to prepare learning contents in the field of natural sciences since multimedia element and application of ICT have a long tradition. The highest level of likeability in the fine arts is almost certainly achieved due to higher level of knowledge and expertise of lecturers and their congenital or inherited sense of beauty. It is therefore advisable that aesthetics principles and guidelines should be presented to all of the teaching personnel. Aesthetically pleased learning materials have positive feedback in the students mind. Even if in some students does not affects their learning potential they do not harm it either. Since we are unable to prepare all learning materials in this manner it is wise to start preparing them for younger students.

LITERATURE

[1] M. Butina, Prvine likovne prakse, Ljubljana: Debora, 1997.

[2] L. d. Chernatony, Blagovna znamka, Ljubljana: Založba GV, 2002.

[3] M. Duh, B. Bedrač and M. Krašna, "Assessment of e-learning materials aesthetic design," in *MIPRO 2012*, Opatija (Croatia), 2012.

[4] D. Dabner, S. Calvert and C. Anoki, Grafično oblikovanje, Ljubljana: Tehniška založba Slovenije, 2011.

[5] R. Landa, Graphic Desing Solution, Albany, NY: Delmar - Thomas Learning, 2011.

[6] Q. Newark, What Is Graphic Design?, Meis (Switzerland): A RotoVision Book, 1997.

[7] R. C. Parker, Grafično oblikovanje (Looking Good in Print), Ljubljana: Pasadena, 1997.

[8] P. Rand, The Symbol in Visual Communication: A Designer's Art, New Haven & London: Yale University Press, 2000.

[9] A. A. Berger, Seeing Is Believing, London, Toronto: Mayfield Publishing Company, 1998.

[10] P. Baines in H. Andrew, Type and Typography, London: Laurence King, 2002.

[11] M. Duh and K. Marjan, "Aesthetic and creativity in the e-learning material," in *MIPRO 2011*, Opatija (Croatia), 2011.

[12] D. Jury, What is Typography?, Mies (Switzerland): RotoVision SA, 2006.

[13] J. I. Biegeleisen, Classic type faces and how to use them: Including 91 complete fonts, New York: Dover Publications, 1995.

[14] R. Bringhurst, The elements of Typographic Style, Vancouver: Hartley & Marks Publishers, 1999.

[15] J. Tschichold, Treasury of Alphabets and Lettering, London & New York: Lund Humphries Publishers, 1992.

[16] G. R. Dowling, Corporate reputations: Strategies for developing the corporate brand, London: Kogan Page Limited, 1994.

[17] B. M. Whelan, Barvna Harmonija 2, Priročnik za ustvarjanje kreativnih barvnih kombinacij, Ljubljana: Rockport Publishers, 1995.

[18] R. Skrt, "Eyetracking raziskave – Kako si ogledujemo spletne strani?," July 2006. [Online]. Available: http://www.nasvet.com/eyetracking/. [Accessed 13 02 2013].

[19] C. J. Cela-Conde, G. Marty, F. Maestu, T. Ortiz, E. Munar, A. Fernandez, M. Roca, J. Rossello and F. Quesney, "Activation of the prefrontal cortex in the human visual aesthetic perception," in *PNAS - Proceedings of the National Academy of Sciences of the United States of America*, online, 2004.

COB: Online educational tool for collaborative ontology building

Mile Jovanov, Marjan Gusev and Darko Martinovikj

Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering,
Rugjer Boshkovik 16, PO Box 393, 1000 Skopje, Macedonia
{mile.jovanov, marjan.gushev}@finki.ukim.mk, darko.martinovikj@gmail.com

***Abstract*—The Semantic Web technologies can really contribute in a better and more effective education. Building a collaborative ontology tool can encourage students to learn and improve their knowledge in a specific area. However, it is not an easy task.**

There are some existing tools for collaborative ontology building, like Collaborative Protege, Hozo, OntoWiki, which offer user collaboration, but their focus is not in education environment, which can affect and lead to harder and more complicated usage in educational purposes.

In this paper, we present a new tool, called COB (Collaborative Ontology Building) tool, built on our proposed model. It matches the requirements for collaborative ontology development and management using the learning content management system Moodle. The presented tool was used at the Faculty of Computer Science and Engineering in Macedonia and very high students participation was achieved. Also, a rich ontology on the domain of Compilers was developed.

***Index Terms*—model, collaboration, ontology, e-Learning**

I. Introduction

Education has been improving and changing recently due to the existence of a lot of new sophisticated technology supported solutions, especially with the rise of Internet. E-Testing is technology used in both teaching and learning and demands extensive dataset of questions. The problem of generation of a question dataset should be delegated to machine by appropriate human support. Semantic webs and ontology play a crucial role in integration and combination of data extracted from different web sources and are possible solution for this problem.

In this paper we continue the results from our recent paper [1], where we have presented a new model for automatic question production that uses Semantic Web ontology (OWL document) as input. This model allows very fast production of large question set for the purposes of e-Testing.

Using the presented model in the process of creation of questions for e-Testing purposes required finished ontology on some domain presented in OWL form. We have presented a new model that allowed collaborative ontology building on a given domain, as part of on-line e-Learning collaborative activity [2]. In this paper we will present a new tool for building a ontology on a given domain that covers the entire curricula of a given course, or a part of it, which is based on the presented model.

The rest of the paper is organized as follows. Section II presents the related work and Section III briefly describes the tool features for collaborative ontology building. The new developed tool is explained in Section IV. Section V presents the results of the evaluation of the tool and Section VI presents conclusion and future work.

II. Related work

Various technical and pedagogical methods for computer and electronic education usage, standards for the creation of electronic course and multimedia textbooks, and trained staff are required within the learning process [3].

Four different steps are identified in the e-Learning cycle model [4]: Skill analysis (The learning manager analyzes the learner's present skills and skills that are set as a learning goal), Content development (The developer creates exercise questions and the content structure linked with explanatory pages), Learning process (The learner engages in learning process suited to the needs and skill analysis - individual learning for knowledge acquisition, or collaborative learning for workshop-type learning), and Evaluation (The learner carries out exercises and takes examinations using questions designed according to the learning goal. The learning manager makes the evaluation of each learner, using results of exercises and examinations).

LMS (Learning Management System) or LCMS (Learning Content Management System) is a platform for on-line courses, represented by a (password protected) web-site containing files such as word processing documents, sound, pictures, videos, etc, allowing students to access them and interact between each other and with the teacher [5]. Mostly used platforms include Moodle [6]; Fle3 Learning Environment [7]; The Manhattan Virtual Classroom [8], ATutor [9],.LRN [10], Blackboard Learning Platform [11] and others.

The technology changing world we live in drives the collaboration as a process with raised intensity. Currently it is a respected method in learning, leading to research and innovations in the area of collaborative learning [3]. There are several definitions about collaboration supported learning, and we will refer to it, as a recursive process where two or more entities work together to realize shared goals by sharing knowledge, learning and building consensus [12]. Wikis have been used for online teaching and learning [13].

As with other processes, the current technology development has great impact on the collaboration. A numerous collaborative tools ease the process of collaboration, giving

an ”e” in front of ”e-collaboration”. Depending on the level and type of collaboration, we may mention sets of web tools for communication, tools for conferencing, a collaborative project management tools and general collaborative management tools. One accepted example is the set of Google tools named ”Google Apps”, now partly included in the new ”Google Drive” [14]. In this way, more and more collaboration is possible. Hence, there is a greater interest in the essence of the process.

III. Features of a tool for collaborative ontology building in e-Learning environment

A tool for collaborative ontology building has to support concurrent ontology building, allowing multiple edits to be made to the ontology by different users. These users will make insertions (entries) to the ontology, so the question ”who will do the insertion evaluation?” must be answered. The possible answers are: the privilege of insertions evaluation will have all of the users or a specific group of users, chosen previously.

We shouldn't disregard the fact that some users produce far more accurate and better insertions compared to the other users and globally the ontology will consist of terms inputted by those users. In the end as a result of a tool of this type a ontology on a given domain will be built.

A right decision is not always brought If all of the users In an e-Learning environment have the privilege for ontology evaluation that can lead to mistakes and inconsideration. This is why we proposed a new model [2] where the evaluators are the teachers of the course with the domain in question, and have great experience and knowledge in the area so they can evaluate the insertions properly. Also in our model we introduced a grading and rating formula of the users, to ease the process of evaluation.

In comparison, today, there are some existing tools for collaborative ontology building like Collaborative protege [15], Hozo [16], OntoWiki [17]. They all allow collaborative ontology building, but as far as evaluation is concerned, all of the collaborators are the evaluators of the insertions.

So, these tools are not suitable to be used for collaborative ontology building in a e-Learning environment, and we concluded that a new tool needs to be built. Further on in this paper we described our realization of the solution, based on our proposed model.

With this model, even with the additional checking of produced questions, the production speed is 10 times bigger than in the process of manual production. Good results of the model were also shown on good question formulation quality. We showed that the set of automatically produced questions doesn't significantly difer from the set of manually produced ones, in the sense of question solvability.

IV. Description of the COB tool

Our choice for Moodle was brought based on the fact that it is a free source LCMS widely used at the university courses worldwide and also used at our University. We refer to the tool itself as COB (Collaborative Ontology Building) tool.

A. COB as Moodle activity module

As described in the development documentation for Moodle in [18], Moodle is structured as an application core, surrounded by numerous plug-ins to provide specific functionality. It is built in PHP scripting language [19], allowing modules to communicate with the Moodle core.

Every activity module in Moodle must follow some basic rules for file locations and mandatory file names. Following the Moodle specification for activity module creation, we have realized and included the next four mandatory files:

- *install.xml:* File in XML-Schema format for defining all of our 12 database tables. This file is executed automatically when the module is installed for the first time.
- *view.php:* PHP file that is executed when the user views an existing activity for ontology collaboration. This is the first opened view. Depending of the role of the user (if he is a teacher or a student) the system shows appropriate actions. We have also files for other views which are opened consequently from the user's actions.
- *mod form.php:* PHP file that is executed when the teacher opens a new activity for ontology collaboration.
- *lib.php:* PHP function file that consists of functions for database writing/reading. These database functions are called from the views when a user requests one view.

Fig. 1 shows execution of files in our module based on user's action.

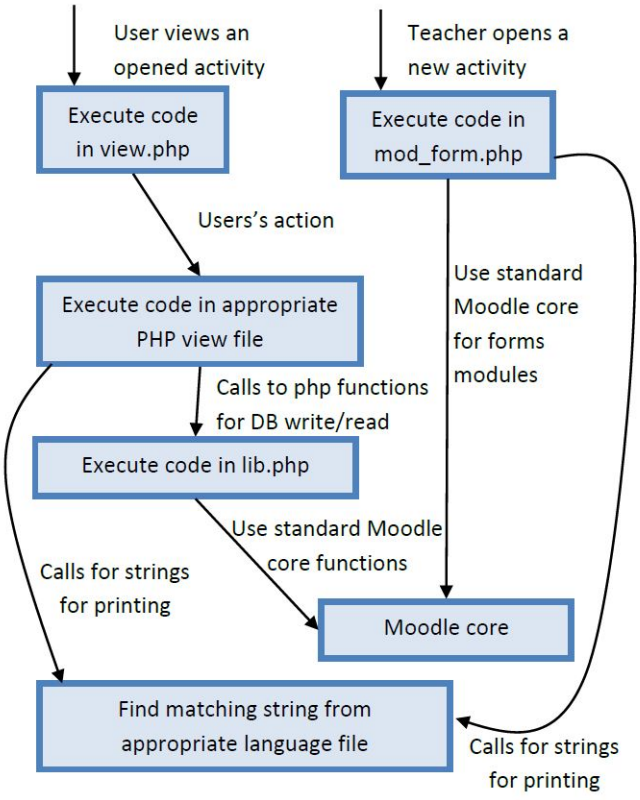


Fig. 1. Activity diagram for executions of files based on user's action

This execution is similar to the standard module activity execution with a difference that all php functions that are called from the views are stored in one file lib.php and the database read/writes are executed from that file.

B. Functionality and realization in Moodle

The functionality that this tool offers is reflection to the proposed model [2].

As we explained in the proposed model we have two types of users: users that will make insertions, i.e. students, and users that will evaluate the insertions, i.e. teachers. The process of collaborative ontology building in Moodle will take place on the appropriate courses. Therefore, the roles of the users are determined by the existing users' roles in Moodle. So when a user requests a php file, the user's role is checked by Moodle and the file is opened if the user has the appropriate privileges.

The process of ontology building is composed of several sub-activities opened gradually after each part of covered matter in the lectures. The opening of one sub-activity is executed by the teacher (Fig. 2), where the teacher specifies the starting and ending date of the activity. During that time, the students can make new insertions to the ontology.

The insertions were divided by their type, forming four tabs for browsing: classes, object properties, data properties and individuals and all types were marked with a unique color for easier recognition. For each sub-activity the user gets a table where all of his insertions are shown together with their grading information (Fig. 3).

When the ending date of a sub-activity passes, the teacher, or teachers - if there are more, need to evaluate the insertions. For every student, starting from the students with the highest rating, every insertion is checked and appropriately marked as correct, incorrect or incorrect with penalty (Fig. 4). After finishing with insertions checking, there is a set (more precisely the multiset) of insertions, acceptable for the inclusion in the ontology.

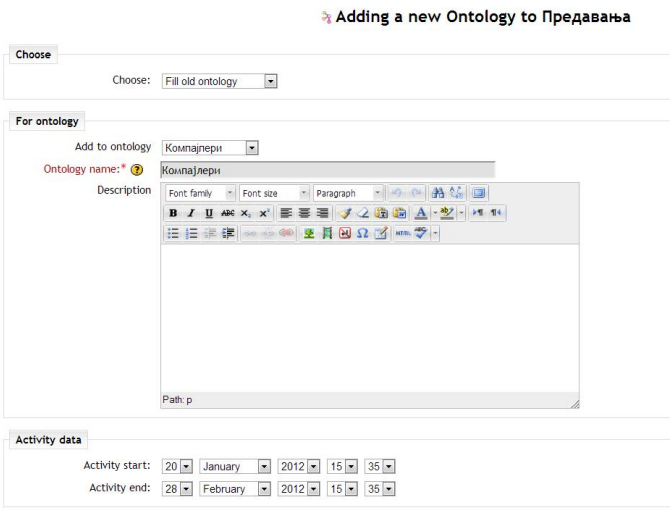


Fig. 2. Teacher opening a new activity

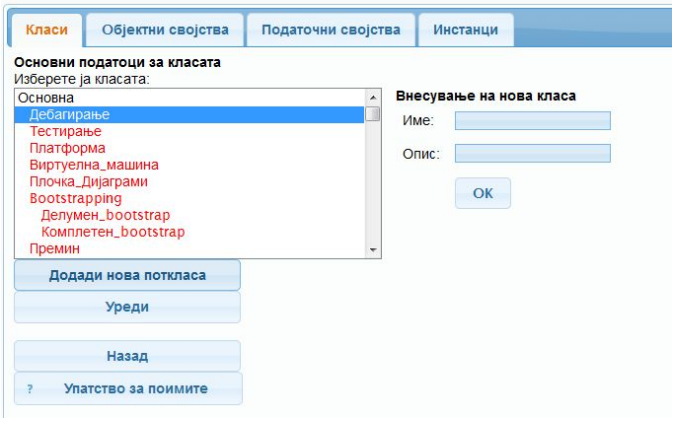


Fig. 3. Student making an insertion

After checking the insertions, the teacher starts to evaluate the set of insertions. If the insertions are named with the same name (same concept, by two or more students), they are already mixed as one term. At the end of every sub-activity automatic grading and rating are calculated for every student that participated in that activity, with the formulas that we presented in [3].

Our activity module is multilingual, so when a view is needed, it has to be translated to the user's selected language. If there isn't a translation on that language then the default language translation is shown. Currently, we have successfully added support for two languages: English and Macedonian. Adding a new language is not a hard task, all that needs to be done is translating the used strings.

C. The realization process

The presented module was gradually built, adding new features and improvements with each new version. When a stable version of the tool was built, which supported the specified functionalities, the testing and quality assurance phase was executed in parallel with a group of 15 volunteers. Following the specification of our model, the volunteers participated in 4 sub-activities, building an ontology in the domain of Object-Oriented Programming. Their task was using this tool and making insertions, to find and report issues and faults. At

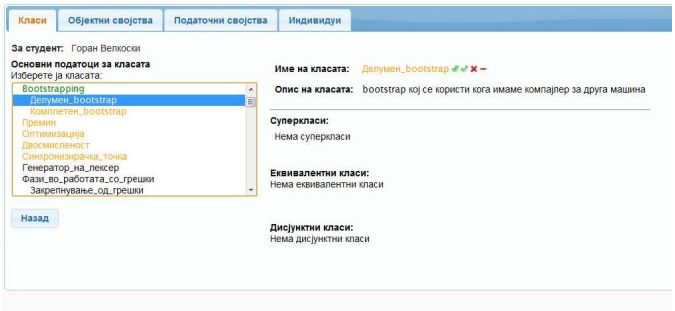


Fig. 4. Teacher checking the insertions of a given student

the beginning not all of them had previous knowledge in the domain of the Semantic Web, so a lecture was held explaining the crucial terms. Interesting improvements were made thanks to this group.

The layout of the views was slightly changed so the number of required clicks for browsing and making new insertions was reduced to a minimum. Additionally, a help views were added containing the explanation of the crucial terms used for Ontology Creation, one example shown in Fig. 5.

The main question and issue that was considered during the testing was the writing of the ontology restrictions. Some of the volunteers were making mistakes while writing the restrictions, and others even avoided writing them. Because of that, we developed a predictive push-down automaton so users will write the restrictions only by clicking and choosing. As far as the syntax for writing restrictions in concerned we chose the Manchester Owl Syntax [20] because of the popularity and usage in other ontology building tools. So the users can write the restrictions with expressions using predictive push-down automaton.

Additionally, in the realization we have used technologies like Ajax, for better user experience. As far as installation is concerned, this tool needs to be installed on server with an existing Moodle system. The installation is the same as installing any other module for Moodle: the administrator should copy the contents of this module in the appropriate directory where all of the modules are located and after that following the standard procedure should install the module. We have tested this plug-in on different versions on Moodle, starting from 1.9 to 2.1. The tool production was finalized at the end of 2011 and it was ready to use in regular courses.

V. Evaluation of the COB tool

We have used the tool in a regular faculty course in order to evaluate the tool and the model itself. As far as the tool is concerned, we examined the following: is the tool user-friendly and does it allow easy use (easy access to the activity, visibility of already accepted insertions in the ontology, quick insertions, etc); do the users continue to use the tool over the next sub-activities; and, do the activity and the tool allow production of comprehensive ontology on the domain in question.

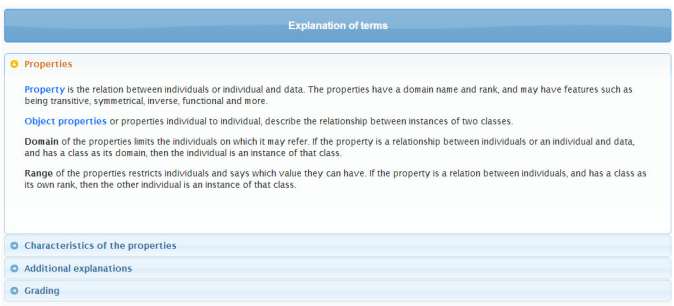


Fig. 5. View explaining the crucial terms

A. Experiment description

As proposed in our solution [2], building an ontology for the domain of a course consists of multiple collaborative activities. Our tool was used in a regular one-semester course Compilers held at our University. The number of participating students was 62. They had the role of making insertions in the ontology, as their own proposals. The evaluators of the insertions were the two teachers of the course. When a sub-activity finishes, the teachers had to check the validity of the insertions. At the end of every sub-activity, the automatic grading and rating took place. Numeric data were gathered and analyzed for all 4 sub-activities.

B. Experiment results and final remarks

Even though this activity was not mandatory, students have shown great interest in participation. During the execution of the experiment a total of 5980 insertions were made. This is a huge number of insertions and it shows that the tool was easy to use and that the insertions were easily done. This was also concluded during the testing stage of the tool (mentioned in the previous section), by the interviews with the testers. Having in mind the number of insertions and the limited time for the activity and the activity evaluation by the teacher(s), it can be concluded that the tool also allows quick evaluation with its additional features of combining the insertions and sorting the participating students by their previous rating.

Having the data, we have also determined which type of insertions were popular among the students and which insertions students avoided. A distribution of the insertions by type is given in Table I. The types of insertions are based on standard OWL Web ontology language constructs, noting that the expressions are compositions of standard restriction constructs in OWL language.

As we can see, the most used type of insertion was class. 57% of the insertions in the final ontology were classes as shown in Fig. 6. The most written restrictions were for the domain and range of the properties. Instances were also popular.

TABLE I
Distribution of total insertions by type

Insertions type	**Accepted in ontology**	**Accepted not in ontology**	**Refused or Penalized**
Classes	160	2560	358
Individuals	37	488	137
Object properties	28	367	156
Data properties	16	64	55
Superclass expressions	3	14	75
Equivalent class expressions	1	26	49
Disjoint class expressions	16	30	18
Equivalent properties	0	0	2
Disjoint properties	0	2	0
Domain exp for properties	35	386	101
Range exp for properties	25	311	77
Individual class expressions	12	179	106
Properties for individual	1	9	76
Total	**334**	**4436**	**1210**

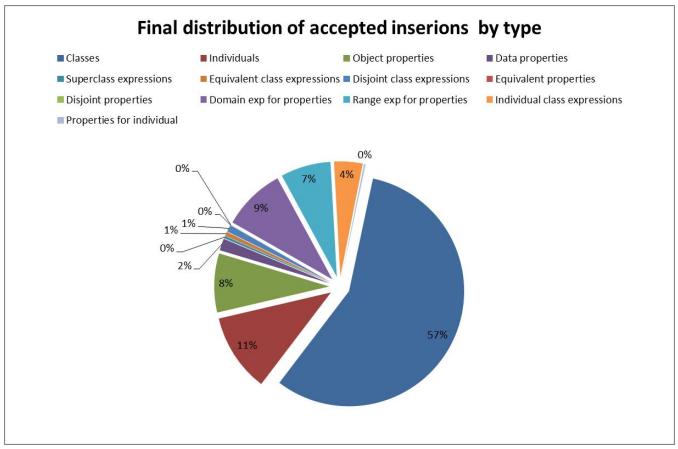


Fig. 6. Final distribution of accepted insertions by type

A total number of participants per activity is shown in Fig. 7. We can see that there is a decrease in participation but the decreased rate is much smaller compared to the decreased rate of overall interest and effort of the students as the semester is passing. Even though the number of participants slightly decreased, there were many insertions in every activity. This confirmed our belief that the tool is accepted by students, and they will continue to use it.

According to [21] there are three main approaches to ontology evaluation: Gold standard evaluation, Criteria based evaluation and Task-based evaluation.

The gold standard evaluation approach requires that the ontology is compared to another ontology that is deemed to be the benchmark.

As the process of ontology building with our tool is observed by the teachers, who have great knowledge on the domain, one can be sure that the produced final ontology will be very close to an (possibly nonexistent) ideal ontology for the domain. If we additionally consider the number of accepted insertions that are not included in the ontology (4436), we can conclude that there are even a large number of redundant

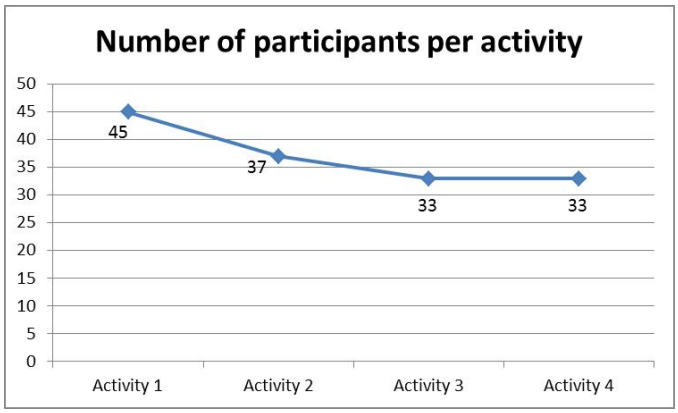


Fig. 7. Number of participants per activity

insertions, which can be a guarantee that the ontology will include almost everything that need to be included. We have to have in mind the number of the students that make insertions. A special attention has to be made to their quality. It is expected that there will be at least a few very good students. Also, the tool offers a feature that allows the teacher(s) to include some omitted (not proposed by students) insertion, and thus enrich the ontology.

The task-based evaluation approach evaluates the ontology based on the competency of the ontology in completing tasks.

In our model we tend to use our ontology firstly as 'educational material' and secondly as an input to our tool for automatic question production. The teachers are aware of that, and they direct the building accordingly.

If we look at **the criteria based evaluation approach**, we can see that this approach requires a human to make the evaluation.

As proposed in the solution, the teachers are the evaluators of the insertions, so it is expected that the resulting ontology will be graded by other human evaluator.

During the years, various criteria have been proposed for the evaluation of ontology, some of them listed in [21]. Although the resulting ontology is primarily intended for educational purposes, we can argue that it complies with a number of the proposed criteria. For example, **conciseness** (the teachers select only the necessary insertions. In the experiment only 334 included insertions out of 3770 accepted for inclusion), **correctness** (completely depending on the expertise of the teacher(s)), **completeness** (in sense of comprehensiveness, the teachers can see large number of proposed ideas and can even add their own insertions, as explained above), etc.

At the end, we present two additional observations. In the Fig. 8 are presented the numbers of insertions ("accepted and in ontology", "accepted but not in ontology", refused or penalized) by sub-activity. As we can see, the number of refused or penalized insertions is decreasing with every following sub-activity, showing that the students get more familiarized with the process. The more important result is the observation that the number of included insertions in the ontology is more or less the same in every sub-activity, and it doesn't depend much on the total number of insertions.

VI. Conclusion

In this paper we presented the COB tool which is a realization of our model of on-line e-Learning collaborative activity for building ontology. We explained that the tool is realized as Moodle activity module, we presented its structure and also explained the process of the realization with inclusion of testing volunteers. With the conducted experiment during one-semester faculty course we examined the characteristics of the tool and presented the findings and observations. They confirm that the tool is usable and well accepted by the students. The produced tool, can be used by every teacher that uses Moodle to teach a course, if she decides that there will be a benefit of the proposed activity for the students.

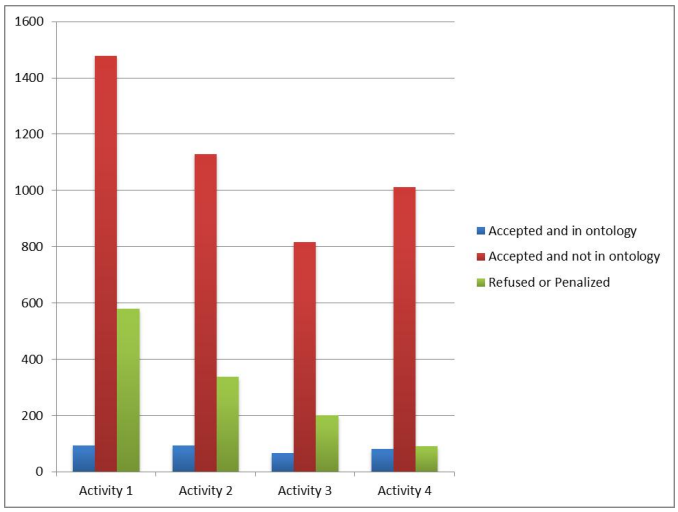


Fig. 8. Number of accepted insertions by activity

The tool can be easily added to every Moodle installation, by the Moodle administrator. At the end of the last sub-activity there is a finished rich ontology document for domain of the course. As future work, we can consider the ways of using the produced ontology in the process of students examination.

ACKNOWLEDGEMENT

The research presented in this paper is partly supported by the Faculty of Computer Science and Engineering in Skopje.

REFERENCES

[1] M. Jovanov and M. Gusev, "Performance evaluation of a new approach for automatic question production," *ICT Innovations 2009*, vol. 7, no. 2, pp. 275–283, 2010.

[2] M. Jovanov, M. Gusev, and D. Martinovikj, "A new model of on-line collaborative activity for building ontology in e-learning," University Sts Cyril and Methodius, FINKI, Tech. Rep. TR-IIT61, 2013, unpublished.

[3] M. Jovanov and M. Gusev, "Grading and rating of students in new on-line collaborative activity," in *Proc. of 4th International Conference of Education, Research and Innovations (ICERI2011)*, 2011, pp. 1976–1981.

[4] V. Nedeva, "The possibilities of e-learning, based on moodle software platform," *Trakia Journal of Science*, vol. 3, no. 7, pp. 12–19, 2005.

[5] S. E. Kameron, "A review of free online learning management systems (lms)," *TESL-EJ*, vol. 7, no. 2, 2003.

[6] (2012) Moodle. [Online]. Available: http://moodle.com/

[7] (2012) Fle3 learning environment. [Online]. Available: http://fle3.uiah.fi

[8] (2012) The manhattan virtual classroom. [Online]. Available: http://manhattan.sourceforge.net

[9] (2012) Atutor. [Online]. Available: http://www.atutor.ca

[10] (2012) Lrn. [Online]. Available: http://dotlrn.org

[11] (2013) Blackboard learning platform. [Online]. Available: http://www.blackboard.com/Platforms/Learn

[12] I. J. Marinez-Moyano, "Exploring the dynamics of collaboration in interorganizational settings," in *Creating a Culture of Collaboration*, Schuman, Ed. Jossey-Bass, 2006, ch. 4, p. 83.

[13] N. Augar, R. Raitman, and W. Zhou, "Teaching and learning online with wikis," in *Beyond the comfort zone: proceedings of the 21st ASCILITE Conference, Perth, 5-8 December*. ASCILITE, 2012, pp. 95–104.

[14] (2013) Google drive. [Online]. Available: http://googleblog.blogspot.in/2012/04/introducing-google-drive-yes-really.html

[15] Colaborative protege, system for ontology collaboration. [Online]. Available: http://protegewiki.stanford.edu/wiki/Collaborative_Protege

[16] Hozo, system for ontology collaboration. [Online]. Available: http://www.hozo.jp/

[17] Ontowiki, system for ontology collaboration. [Online]. Available: http://ontowiki.net/Projects/OntoWiki/

[18] Moodle developer documentation. [Online]. Available: http://docs.moodle.org/dev/Developer_documentation/

[19] Php scripting language. [Online]. Available: http://php.net/

[20] Manchester owl 2.0 syntax. [Online]. Available: http://www.w3.org/TR/owl2-manchester-syntax/

[21] J. Yu, J. A. Thom, and A. Tam, "Ontology evaluation using wikipedia categories for browsing," *Proc. of the Sixteenth Conference on information and knowledge management: CIKM2007*, pp. 223–232, 2007.

Object-oriented programming for graphical modeling on the example of logistics network design

Slavomir Vukmirović, Dejan Kovač and Ivana Mežnarić
Economics Faculty
University of Rijeka, Croatia
vukmirovics@gmail.com

Abstract - In this paper, we discuss methods and concepts of Object Oriented Programming (OOP) for graphical modeling. We illustrate through a set of examples how geometrical models are built, and visualized solutions to logistics network design problems are obtained. We develop a methodological framework for visual programming and linking of graphical models concepts, and introduce the basic concepts of graphical models: lines, arrows, nodes and coordinates. We use block diagrams and object-oriented programming to program and link the concepts of graphical models using Visual Basic for Excel.

I. INTRODUCTION

Graphic design has radically changed over the last 15 years. The biggest change that has happened is the change in graphical modeling techniques today working in a fully computer-aided environment. Graphical modeling can be described as an interdisciplinary activity, and graphical model as a message resulting from analysis of society, art, technology, moment, place, culture; design is an interdisciplinary activity [2]. Skills needed to apply graphical modeling techniques are changing with developments in the information and communication technology. [3]

OOP for graphical modeling encourages integration and interaction of computer programming and graphical modeling knowledge, where psychomotor skills needed in traditional hand drawing and modeling are replaced with programming skills. The basic paradigm of active operation in the information technology is programming. Specifically, computer is a universal machine that can be transformed into a tool for performing varied tasks using appropriate programs. Visual Basic for Excel has become an object-oriented visual programming language used for integrated graphical modeling and programming enabling us to better understand and visualize the process of graphical modeling.

The synergy of object-oriented programming and graphical modeling allows us to be creative and share our innovation offering us more possibilities than ever. In this paper, we illustrate object-oriented programming for graphical modeling with an example of how logistics networks are built using Visual Basic for Application (VBA) within Excel interface. We also develop a methodological framework for object-oriented programming for graphical modeling of logistics networks [2], [3].

Many studies have been conducted and presentations have been made on object-oriented programming for graphical modeling. In this paper, we illustrate object-oriented programming for graphical modeling with an example of how logistics networks are built using Visual Basic for Application (VBA) within Excel interface. We also develop a methodological framework for object-oriented programming for graphical modeling of logistics networks.

In this paper, we present some parts of the program. Complete program codes in VBA for Excel and presentations of all the programs in Excel as described in this paper, can be found on the web site www.infosustav.com/vkm_mipro2013.xls

II. DEFINITION, DESCRIPTION AND GRAPHICAL PRESENTATION OF LOGISTICS NETWORK MODEL ELEMENTS

"Logistics Network can be defined as a system of interrelated logistics centers, communications, corridors, ports, terminals, lines or services, logistics chains, logistics and distribution chains, transportation chains and transportation networks built to enable fast, safe and rational production process of logistics products" [6]. In this work, we discuss graphical modeling of logistics networks, and define the elements of logistics network. A combination of qualitative and quantitative analysis of the logistics network concept enables us to determine that a logistics network is made up of cells, nodes, and arcs. In the figure 1, we see the elements of logistics network model within Excel interface.

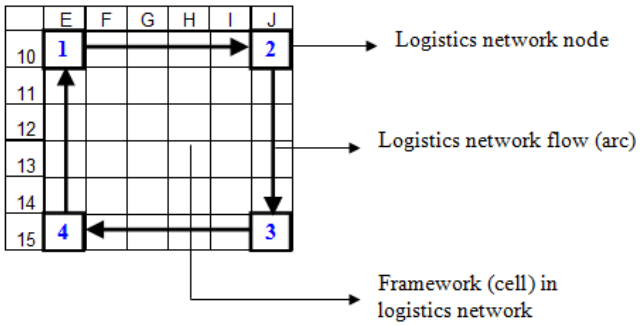


Figure 1. Logistics network model elements

Each element of logistics network has its own basic characteristics [6]:

- **Cells** are smaller or larger geographical areas having various forms, bordered by communications, corridors, channels, lines or services. Multiple such areas can refer to smaller or larger logistics gravitational zones around smaller or larger logistics and distributions centers or logistics centers within logistics networks where logistics and transportation chains are operating.
- **Nodes** are commonly smaller or larger logistics centers, where warehouses, terminals, free zones, trade centers and industrial zones are located, and are interrelated by logistics and distribution chains or logistics and transportation chains. Nodes are usually represented by circles. In this work, nodes are represented by squares, which is the most appropriate form for programming within Excel interface.

- **Arcs** are communications, railways, air corridors, ship lanes, river or lake shipping routes, pipelines etc. which link, directly or indirectly, numerous producers, suppliers, retailers and consumers with logistics centers.

III. COMPARATIVE ANALYSIS OF CARTESIAN COORDINATE SYSTEM AND EXCEL COORDINATE SYSTEM

In a coordinate system, an arbitrary point O (the origin) is chosen on a given line. The xy-coordinate plane has two coordinate axis. Horizontal axis is the x-axis (abscissa), and vertical axis is the y-axis (ordinate) [1]. In a traditional, Cartesian coordinate system as shown in the figure 2a, each point on a plane is specified by a pair of numerical coordinates (x,y). The abscissa or x-coordinate is the first coordinate. The ordinate or y-coordinate is the second coordinate. Points on the coordinate plane are specified as pairs of numerical coordinates T(x,y) given that x>0, y>0 [5].

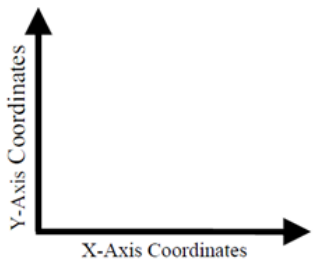


Figure 2a. Cartesian coordinate system

The coordinate system generated using Visual Basic for Applications, VBA within Excel interface (hereinafter referred to as „Excel coordinate system“) is different from the traditional, Cartesian coordinate system. In the Excel coordinate system (cf. Figure 2b), the origin is in the left, upper angle defined by the first row and first column in the table (R1,C1) [4]. In the Excel coordinate system, the order of coordinates is different so that the Y-coordinate is the first coordinate, and X-coordinate is the second coordinate. In the Excel coordinate system, each point is identified by cell address, and has a location of Cells(y,x) [4].

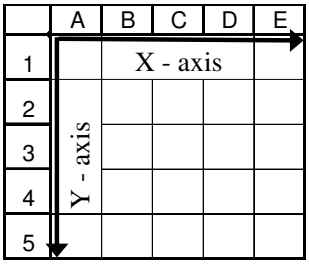
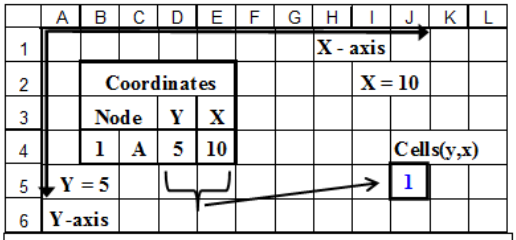


Figure 2b. Excel coordinate system (LeBlanc, 2003.)

IV. PROGRAMMING POSITIONS AND BUILDING NODES IN EXCEL COORDINATE SYSTEM

We illustrate through an example how positions are programmed and nodes are built in the Excel coordinate system using Visual Basic for Applications (cf. figure 3).




```
Private Sub Node_Click()
'Defining position
y = Cells(4, 4)
x = Cells(4, 5)
v = 1
With Cells(y, x)
      lbx = .Left
      lby = .Top
   End With
'Adding and building node
Shapes.AddShape(1, lbx, lby, 20, 20).Select
Selection.ShapeRange.Line.Weight = 2#
Selection.Characters.Text = v
Selection.HorizontalAlignment = xlCenter
Selection.VerticalAlignment = xlCenter
Selection.Characters.Font.Name = "TimesNewRoman"
Selection.Characters.Font.Size = 12
Selection.Characters.Font.Color = vbBlue
Selection.Characters.Font.Bold = True
End Sub
```

Figure 3. Programming positions and building nodes

Complete program codes in VBA for Excel and presentations of all the programs in Excel as described in this paper, can be found on the web site www.infosustav.com/vkm_mipro2013.xls

In the figure 2, node coordinates have cell references D4(y) and E4(x). B4 contains node number (1), and C4 contains node letter (A).

In the figure 3, cell B4 contains node number, and cell C4 contains node letter. Y-coordinate is fixed in cell D4, and X-coordinate is fixed in cell E4. Node letters are recorded in H column's cells H2 and H3, and node numbers are recorded in G column's cells G2 and G3. Y-coordinates are fixed in I column's cells I2 and I3, and X-coordinates are fixed in J column's cells J2 and J3.

In the figure 3, in the beginning of the program, values contained in the table cells are assigned to the variables y and x. Node number is assigned to the variable v. *With Cells(y,x)* defines node position (y=10, x=5). *Shape.AddShape* adds node symbol to its defined position. Number 1 in the brackets selects square as a node symbol. Values (20,20) in the brackets indicate dimensions of a square in points (pt) which correspond to the dimensions of a cell within Excel interface. In the second part of the program, we define parameters of a small square symbolizing the node: frame lines and color, letter font and color, and node number alignment. Square is also used as text frame (TextBox) inside which we put node denotation.

V. NODE PLACEMENT AND ARC ROUTING

To create a flow, we need more than just a single point in a logistics network. A flow can be created with at least two or more points as every flow has its starting point and destination. The beginning syntax differs from the node syntax in that it contains a "For" loop and a condition (If n >=2 Then…) to adequately establish coordinate locations. In this way, we obtain 4 coordinates (x1,y1,x2,y2) which define a location in logistics network through a "With" loop.

In the beginning of the program, we define a program loop to position and build the nodes 1 and 2. Positions of the starting point and ending point of the arc are defined by the positions of nodes 1 and 2. The starting point of the arc is specified by the coordinates (y1,x1), and its ending point is specified by the position (y2,x2). After the positioning, we add and build the arc, as illustrated in the figure 4.

VI. ADJUSTING LENGTH AND POSITION OF ARCS IN GRAPHICAL MODELS

Excel does not provide a horizontal or vertical ruler, and there is no quick method to set worksheet or cell size in centimeters. For this reason, Excel provides symbols, points and pixels as units of measure. While dragging across the row or column headings to adjust the width of a column or a row in a worksheet, in the ScreenTip we can read the width in signs and pixels in brackets (cf. Figure 5).

	E	F	G	H	I	J
1			**Nodes**		**Y**	**X**
2			**1**	**A**	**5**	**5**
3			**2**	**B**	**5**	**10**
4						
5	**1**					**2**

```
'Nodes positioning
For i = 2 To 3
y = Cells(i, 9)
x = Cells(i, 10)
v = i - 1
Next i
'Arcs (flows) positioning
i = 2
y1 = Cells(i, 9)
x1 = Cells(i, 10)
y2 = Cells(i + 1, 9)
x2 = Cells(i + 1, 10)
With Cells(y1, x1)
lbx = .Left + .Width
lby = .Top + .Height
End With
With Cells(y2, x2)
lex = .Left + .Width
ley = .Top + .Height
End With
'Adding and designing arcs
ActiveSheet.Shapes.AddLine
(lbx, lby, lex, ley).Select
```

Figure 4. Arc routing

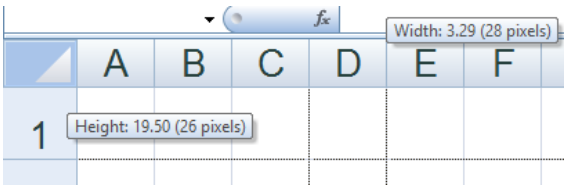


Figure 5. Units of measure in Excel

As the vertical line has 19.50 points or 26 pixels, the arc becomes positioned 10 points lower that is 9.50 points so that in the end the line (flow) is placed in the middle of the cell with respect to the vertical line. In this way, the syntax ("expression".AddLine(BeginX, BeginY, EndX, EndY)) determines the flow in logistics network. Nodes become more visible, and arcs become more accurate and do not overlap with the nodes (cf. Figure 6a and 6b).

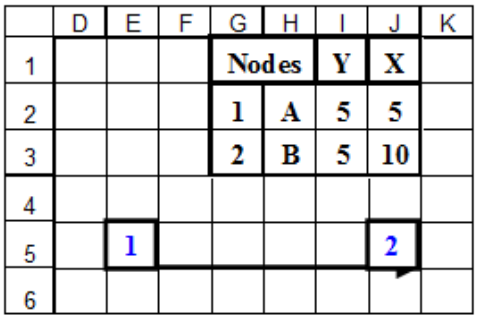


Figure 6a. Unadjusted arc parameters

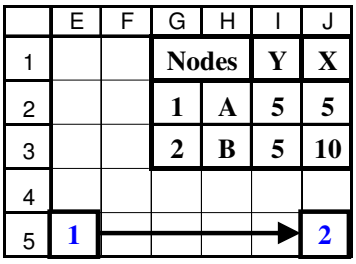


Figure 6b. Adjusted arc parameters

When building an arc to link nodes in a network, we need to adjust the arc parameters. In the figure 1a, we see the graphical model before building an arc. The program in the figure describes the adjustment of the length and position of an arc within a row. *„If y2=y1 And x2>x1"*defines the condition based on interrelationship between the coordinates of the starting and ending point of the arc as contained in the cell I2:J3. This condition is necessary for building of a logistics network, which is made up of multiple interrelated arcs and nodes. When positioning a flow, as described in the program, we are able to adjust the length of an arc using formulas for x1,y1,x2 i y2. We are also able to precisely define the location of an arc in a worksheet (cell) using formulas for lbx, lby, lex, ley and conditions. The program could look as follows:

```
'Adjusting the length
If y2 = y1 And x2 > x1 Then
x2 = x2 - 1
End If
    With Cells(y1, x1)
          lbx = .Left + .Width
           lby = .Top + .Heigh
    End With
    With Cells(y2, x2)
          lex = .Left + .Width
          ley = .Top + .Height
    End With
'Adjusting the position within a row
If y2 = y1 And x2 > x1 Then
lby = lby - 10
ley = ley - 10
End If
ActiveSheet.Shapes.AddLine(lbx, lby, lex, ley).Select
```

VII. LOGISTICS NETWORK AS GRAPHICAL MODEL

In the figure 7, we present a logistics network as a graphical model built using VBA for Excel.

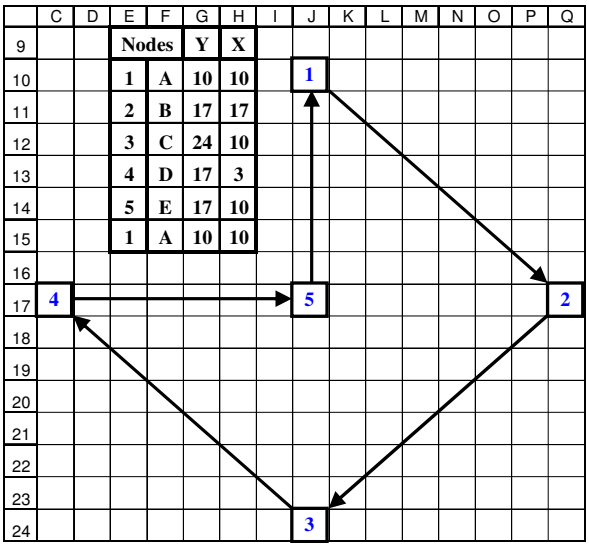


Figure 7. Logistics Network as Graphical Model

The program describes the placement of arcs and nodes. In the beginning of the program, we define a common program loop *„For and = 10 To 14"*, which links the coordinates of arc/node positions as specified in the cells G10:H15. This logistics network is a circular-flow model with the starting(source) node (1) being also the destination node. Coordinates of the cells G15:H15 that define the position of destination node are equal to the coordinates of the cells G10:H10 that define the position of starting node. Segments of the program for the placement of nodes and arcs could look as follows:

```
'Positioning arcs and nodes
For i = 10 To 14
'Positioning nodes
y = Cells(i, 7)
x = Cells(i, 8)
v = Cells(i, 5)
'Positioning arcs
y1 = Cells(i, 7)
x1 = Cells(i, 8)
y2 = Cells(i + 1, 7)
x2 = Cells(i + 1, 8)
Next i
```

Nodes and arcs in a logistics network are built by combining a code for graphical modeling and placement of nodes and flows. In the table below, we explain the lines of code in the programming formulation of the network problem.

TABLE I. PROGRAMMING WITH NODES AND FLOWS

Node design	Flow design
Shapes.AddShape(1, lbx, lby, 20, 20).Select	ActiveSheet.Shapes.AddLine(lbx, lby, lex, ley).Select
Selection.ShapeRange.Line.Weight = 2#	Selection.ShapeRange.Line.Weight = 2
Selection.HorizontalAlignment = xlCenter	Selection.ShapeRange.Line.Style = msoLineSingle
Selection.VerticalAlignment = xlCenter	Selection.ShapeRange.Line.EndArrowhead Style = msoArrowheadTriangle
Selection.Characters.Font.Name = "TimesNewRoman"	Selection.ShapeRange.Line.EndArrowhead Length = msoArrowheadLong
Selection.Characters.Font.Size = 12	Selection.ShapeRange.Line.EndArrowhead Width = msoArrowheadWide
Selection.Characters.Font.Color = vbBlue	
Selection.Characters.Font.Bold = True	

VIII. GRAPHICAL MODELING OF A RELATIONSHIP TYPE

Relationships in logistics network often have different parameters. In the figure 8, we describe how a relationship type is graphically modeled and programmed. In this example, we define a type of relationship from the point of view of max allowed speed. Relationship type (TR) is defined in the cells G9:G14. Faster relation is marked with 1, and slower relation is marked with 0. For instance, 1 in the cell G10 indicates a fast relation AB between the nodes 1 and 2, while 0 in the cell G11 indicates a slower relation BC between the nodes 2 and 3. TR values have been assigned to the variable Z. Based on the conditions set up in the program, if z=1, the arc line is bold (Line.Weight=2), and if z=0, the arc line is thinner (Line.Weught=1).

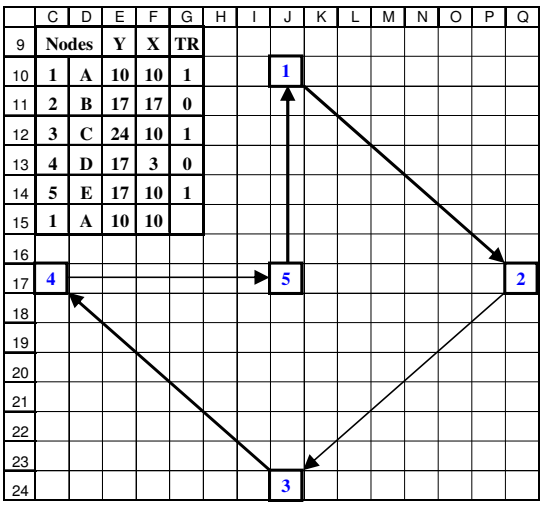


Figure 8. Graphical modeling and programming of a relationship type

Segment of the program for the relationship type could look as follows:

```
z = Cells(i, 7)
ActiveSheet.Shapes.AddLine(lbx, lby, lex, ley).Select
If z = 1 Then
Selection.ShapeRange.Line.Weight = 2
End If
If z = 0 Then
Selection.ShapeRange.Line.Weight = 1
End If
```

IX. SELECTION AND GRAPHICAL MODELING OF OPTIMAL LOGISTICS RELATION

An optimal solution to a logistics network design problem can be defined in the following way: of all combinations of transportation routes which make potential logistics relations, we select the combination which is optimal based on given criteria, where minimum distance between nodes or minimum travelling cost are the most commonly used optimization criteria. In this example, the optimal logistics relation has been specified in the column M. Cell M1:M7 defines the optimal order of transportation nodes or combination of transportation routes as an optimal logistics relation. In the figure 9, we describe how the order of the coordinates of nodes and arcs is modeled and programmed based on given optimal relation.

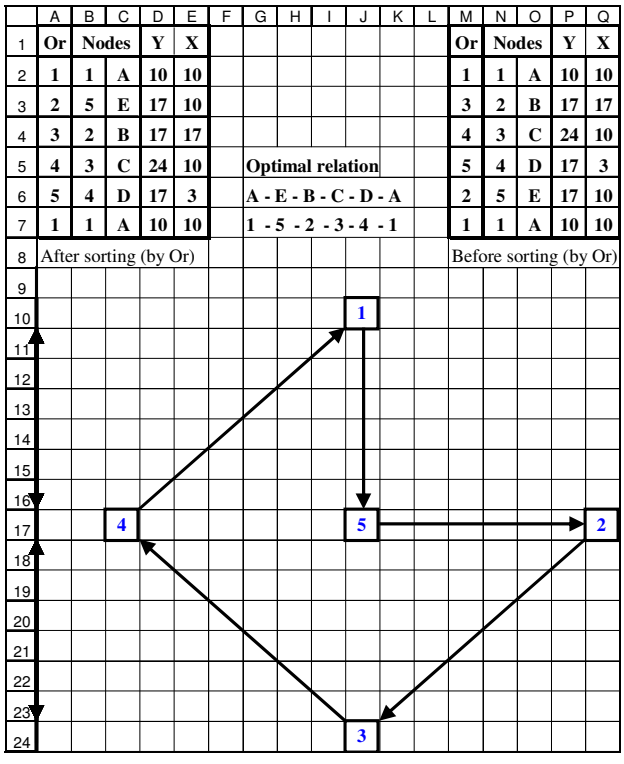


Figure 9. Modeling and programming of the order of the coordinates of nodes and arcs

The table reveals that 1-5-2-3-4-1 is the optimal path of passing through nodes.

With figure 9, we present the program for copying data on node coordinates and the order of nodes and arcs, as well as sorting based on order criteria (column Or).

```
Private Sub Sorting_Click()
'Copying
Range("M2:Q6").Copy
Range("A2:E6").PasteSpecial
'Sorting
Range("A2:E6").Select
Selection.Sort Key1:=Range("A2"), Order1:=xlAscending,
Header:=xlGuess, _OrderCustom:=1, MatchCase:=False,
Orientation:=xlTopToBottom
End Sub
```

X. CONCLUSION

Following above, we are able to propose a methodological framework for object-oriented programming for graphical modeling of logistics networks, which consists of the following few, interrelated activities:

1. Definition, Description and Graphical Presentation of Logistics Network Model Elements
2. Comparative Analysis of Cartesian Coordinate System and Excel Coordinate System
3. Programming Positions and Building Nodes in Excel Coordinate System
4. Node Placement and Arc Routing
5. Adjusting Length and Position of Arcs in Graphical Models
6. Logistics Network as Graphical Model
7. Graphical Modeling of a Relationship Type
8. Selection and Graphical Modeling of Optimal Logistics Relation

REFERENCES

[1] Element d.o.o., Zagreb, Koordinatni sustav (eng. Coordinate System), http://element.hr/artikli/file/1593

[2] Jerbić-Zorc, G., Metodika nastave informatike (eng. Methods of informatics education) www.phy.pmf.unizg.hr/~gorjana/nastava/Informatika/nastavni%20materijali/predinf.pdf

[3] Lajić, D., Grafički dizajn u poslovanju (eng. Graphic Design in Business), Diplomski rad, Ekonomski fakultet, Sveučilište u Rijeci, Rijeka, 2008.

[4] LeBlanc, L. J., Visualizing solutions in network analysis with Visual Basic for Applications, Asia Pacific Management Review (2003) 8(4). apmr.management.ncku.edu.tw/comm/updown/DW0712070554.pdf

[5] Normala, udruga za promicanje nastave matematike (eng. Association for promotion of mathematics education) www.normala.hr/interaktivna_matematika/jednadzba_pravca/sadrzaj/uvod.htm

[6] Zelenika, R., Logistički sustavi (eng. Logistics Systems), Ekonomski fakultet u Rijeci, Rijeka, 2005.

Calculating Duration and Convexity of Bonds Using Excel

Zdenko Prohaska, Ph.D.
Bojana Olgić Draženović, Ph.D.
Faculty of Economics
University of Rijeka
Ivana Filipovića 4, Rijeka, 51000, Croatia
Phone: (385) 51 355 111 Fax: (385) 51 675 750
Valter Šarić, M.Sc.
Zagrebačka banka
UniCredit Group
zdenko.prohaska@ri.t-com.hr, bolgic@inet.hr, valter.saric@unicreditgroup.zaba.hr

The presented results are part of the research project Croatian Financial Markets and Institutions in the Process of EU Accession supported by the Ministry of Science, Education and Sports of the Republic of Croatia

Abstract - To analyze interest rate risk of coupon bearing bonds and to immunize bond portfolios against this risk excel spreadsheets are developed using only plain vanilla excel, i.e. its basic functions and some additional functions from the Excel Analysis Toolpak Add-In.

The reason for not using complicated macros nor Visual Basic for Application is that especially students at the University level have already all the necessary skills to use Excel efficiently and VBA applications of different MS Excel versions are not always compatible and have to be reprogrammed to run smoothly on newer versions of Excel.

I. INTRODUCTION

Educational software is essential in finance for the students to understand better the investment process in securities, like bonds. It is also necessary to apply the methods of calculating prices, rates of return of fixed income securities in everyday practice. Because students at the Universities especially at faculties of economics are trained to use Excel, it is possible to use such a spreadsheet program and its financial functions built in to calculate and evaluate interest rate risk in bond investments. Excellent books about Using Excel in Finance and Business exist. [1], [8], [9].

II. THE BASICS OF INVESTMENTS IN BONDS

Since bonds are securities and listed on secondary capital markets they can be traded like shares and it is therefore necessary to understand beside their the nominal or face value, nominal interest rate and maturity, such categories as the current yield, yield to maturity and present value or price. After that, more sophisticated aspects of bond investments like duration and convexity can be discussed. [2]

A. Current yield

The current yield is the annual rate of return of coupon bearing bonds and is calculated by dividing the nominal interest rate, i.e. the coupon rate by the bond price in percent [6].

$$Y_c = \frac{i_n}{P_b}$$

Y_c - current yield
P_b - bond price
i_n - coupon rate

The current yield is not taking into account the price at which a bond is sold and no reinvestments of interests are considered.

B. Yield to maturity

The yield to maturity is a measure of the rate of return that will be earned on a bond if it is bought now and held until maturity. To calculate the yield to maturity it is necessary to solve the bond price equation for the interest rate given the price of the bond.

$$PV = \frac{C}{1+r} + \frac{C}{(1+r)^2} + \ldots + \frac{C}{(1+r)^n} + \frac{FV}{(1+r)^n}$$

PV - present value of the bond
C - coupon
n - number of years/periods
r - rate of return/yield to maturity
FV - future value of the bond

According to the equation presented above it is possible to prepare an Excel spreadsheet using built in financial functions to determine the yield to maturity of a bond given the current purchase price. This is the true rate

of return on a bond in the case that all coupons are reinvested at the rate of return when the bond was bought.

C. *Present value of a bond*

The present value of a bond is the sum of present values of coupons and the present value of the par or nominal value of a bond. It means that the equation presented above is to be solved calculating the present value or PV.

In Excel spreadsheets presented later the present value of a coupon bond will be determined at the rate of interest selected by the student. It is necessary to prepare input data of the interest rate and the financial information on the bond. After that the present value is calculated.

D. *Duration of a bond*

The duration is a measure of the average life of a bond, defined as the weighted average of the times until each payment is made, with weights proportional to the present value of the payment. [3]

$$D = \left[\sum_{t=1}^{N} \frac{t_C}{(1+y)^t} + \frac{NF}{(1+y)^N}\right] : \left[\sum_{t=1}^{N} \frac{C}{(1+y)^t} + \frac{F}{(1+y)^N}\right]$$

D - duration of a bond
N - maturity in years
C - coupon
F - face value
y - rate of return/yield to maturity
t - year when the payment is made

Using duration instead of maturity it is possible to minimize interest rate risk affecting investments in bond portfolios. In the Excel spreadsheet (duration.xls) duration is calculated for coupon bearing bonds both for annual and semiannual payments of interests and presented in a graph.

The main characteristics of duration are: [7]

a) the duration of a bond with coupon payments will always be less than its term to maturity because duration takes into account these interim payments.

b) an inverse relationship exists between coupon and duration. A bond with a larger coupon will have a shorter duration because more of the total cash flows are paid sooner in the form of interests.

c) there is a positive relationship between term to maturity and duration, i.e. all other factors being the same, a bond with longer term to maturity will almost always have a higher duration.

d) the relationship between the yield to maturity and duration is an inverse one. A higher yield to maturity of a bond reduces its duration.

E. *Modified duration*

An adjusted measure of duration called modified duration could also be used as an indicator of bond price volatility. Modified duration equals Macaulay duration presented above, divided by 1, plus the current yield to maturity, divided by the number of payments per year. Or written in an other form one could say that modified duration can be defined as

$$\frac{dP}{P} = -D\frac{dr}{(1+r)}$$

P - price of the bond
dP - change of the bond price
r - interest rate or yield
dr - change of interest rate or yield
D - duration

Because it is easier sometimes to deal with integrals and continuous compounding than with discrete time periods, it is possible with continuous compounding where r tends against 0 to reduce the relation stated above and get the following equation

$$\frac{dP}{P} = -Ddr$$

so it is easy to see that bond price movements will vary proportionally with modified duration for small changes in yields or as is shown specifically in last equation, an estimate of the percentage change in bond price equals the change in yield times modified duration. [4]

F. *Convexity*

The Modified duration makes it possible to estimate bond price changes for a change in interest rates. The former equation is, however, accurate only for very small changes in yields on the secondary bond markets. For larger changes of yields or nonparallel shifts of yield curves the accuracy of the estimate of the price change deteriorates because the modified duration calculation according to former equation is a linear approximation of a bond price change which in fact describes a curve or a convex function. [10]

In that case convexity is a better measure of the curvature of the price-yield relationship. To be more precise convexity is the second derivative of price with respect to yield (d^2P/di^2) divided by price. Therefore convexity is the percentage change in dP/di for a given change in yield:

$$Convexity = \frac{\frac{d^2P}{di^2}}{P}$$

Convexity is a measure how much the price-yield curve of a bond deviates from the linear approximation of that curve. When bond yields decline, bond prices rise at an increasing rate. On the other side bond prices fall at a decreasing arte when yields rise. Or to explain it in another form, bond prices do not move up and down at the same rate. Bond prices rise faster than they decline, which means that the price yield curve represents a positive convexity curve.

Similar to duration, convexity also has some main characteristics:

a) there is an inverse relationship between coupon and convexity (yield and maturity constant); the greater the coupon rate, the higher the convexity

b) there is a direct relationship between maturity and convexity (yield and coupon constant; the longer the maturity the higher the convexity

c) there is an inverse relationship between yield and convexity (coupon and maturity constant); the greater the yield, the lower the convexity

III. EXCEL SPREADSHEETS

For the calculation of duration and convexity of bonds presented above ordinary Excel spreadsheets will be used applying only standard financial functions and some additional functions from the Excel Analysis Toolpak Add-In. [5]

A. Bond duration

Taking into account EAR (Effective Annual Rate) or APR (Annual Percentage Rate) convention bond duration and modified duration will be calculated in three different ways. First duration will be as the weighted-average time to the bond's cash flows. Second a formula for duration will be used. In both cases modified duration is a simple adjustment of regular duration. Third, we will use Excel's Analysis ToolPak Add-In DURATION and MDURATION functions, which only work under the APR convention. (Fig. 1)

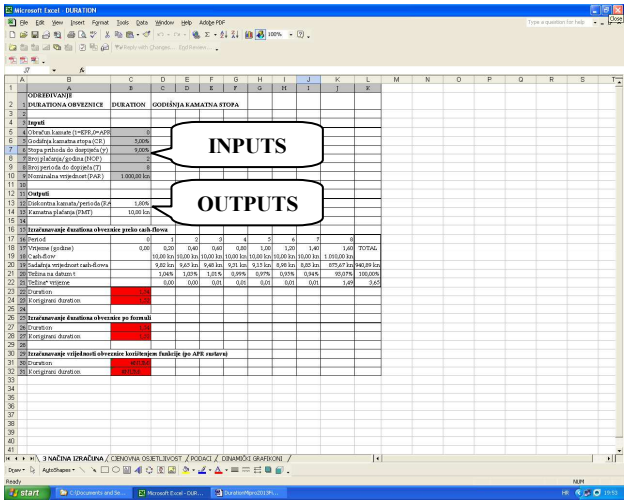


Figure 1. Bond duration – basics

After that we present in a graph the price sensitivity of a bond using duration. Bond duration is a measure of a bond's interest rate risk. It tells us approximately what percent change in bond price will result from a specific change in yield. The duration approximation can be displayed on graph and compared to the actual percent change in the bond price. (Fig. 2)

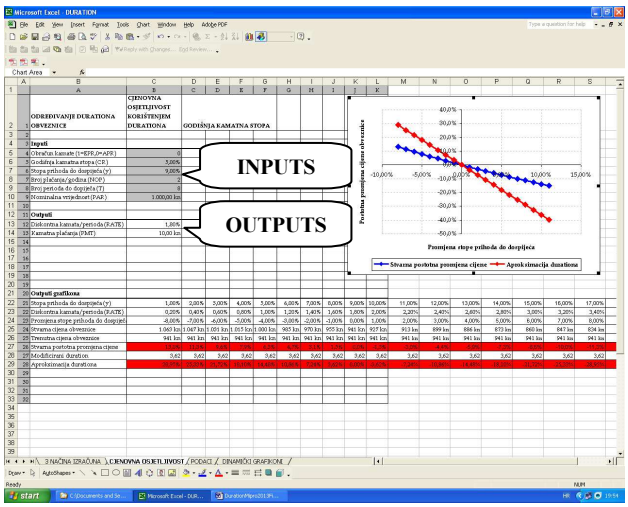


Figure 2. Bond duration – price sensitivity using duration

B. Bond convexity

Convexity will be calculated in two different ways. First, convexity will be calculated as the weighted-average to the bond's cash flows. Second, it is possible to use a special formula for the calculation of convexity. (Fig. 3)

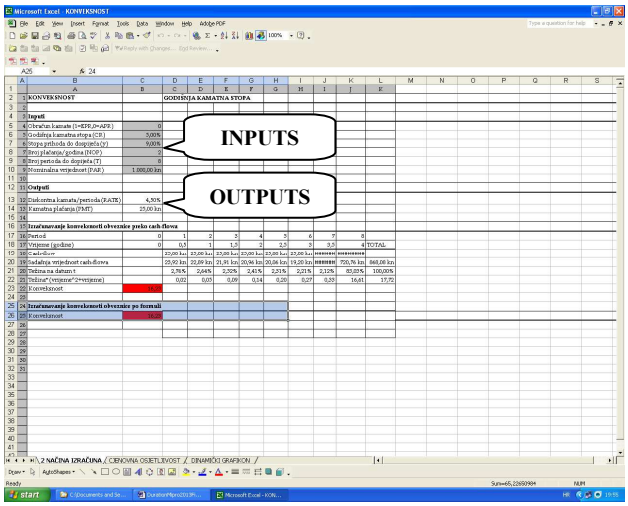


Figure 3. Bond convexity – basics using convexity

For even a more precise way to illustrate the price sensitivity of bonds to changes in interest rates we will use both duration and convexity combined in a graph to get a better measure of the bond's interest rate risk. (Fig. 4)

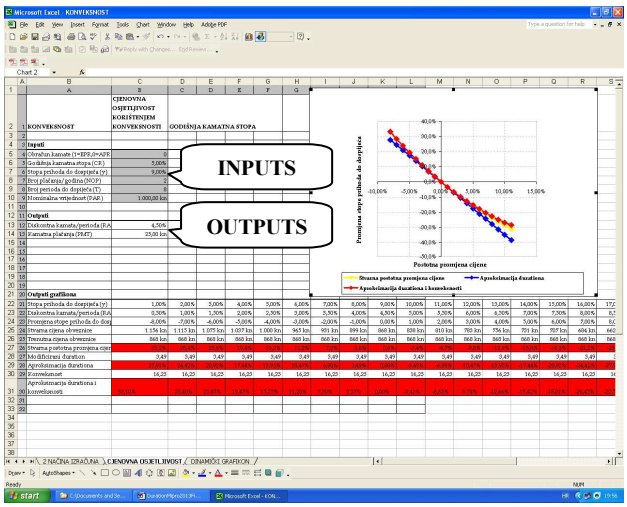


Figure 4. Bond convexity – price sensitivity including convexity

IV. DELIVERY AND INSTALLATION OF THE PROGRAM

Since all mentioned procedures and calculations are implemented in two Excel spreadsheets (duration.xls) and (convexity.xls) it is only necessary to copy these files in the working directory of the PC where Microsoft Excel program can access them. Of course Excel Analysis Toolpak Add-In has to be installed to enable some specific financial functions.

V. CONCLUSION

In this article different methods to measure interest rate risk of fixed income securities or bonds are presented, mainly duration and convexity. For this purpose Excel spreadsheets are developed using only plain vanilla Excel with only some additional financial functions from the Analysis Toolpak Add-In. The results are presented also on a graph which are included in the two spreadsheets.

REFERENCES

[1] Benninga, S., Principles of Finance With Excel, Oxford University Press, New York, 2006

[2] Bodie, Z., Kane, A., Marcus, A.J., Investments, Irwin, Chicago, 1996

[3] Fabozzi, F., Investment Management, Prentice Hall, Englewood Cliffs, New Jersey, 1995

[4] Faerber, E., All About Bonds, Bond Mutual Funds, and Bond ETFs, McGraw-Hill, New York, 2009

[5] Holden, C.W., Spreadsheet Modeling in Investments, Prentice-Hall, New Jersey, 2002

[6] Prohaska, Z., Uroda, I., Bondmaster – Computer Software for Bond Management, in Čičin-Šain, M. et al., 30 th Jubilee International Convention MIPRO 2007, Proceedings Computers in Education, Opatija, 2007

[7] Reilly, F.K., Investments, The Dryden Press, Orlando, 1996

[8] Underdahl, B., Excel Expert Solutions, Que Corporation, Indianapolis, 1996

[9] Veronesi, P., Fixed Income Securities, John Wiley & Sons, Hoboken, New Jersey, 2010

[10] Wilmott, P., Paul Wilmott Introduces Quantitative Finance, John Wiley & Sons, New York, 2001

PaintArt - An Interactive Computer Software for Education of Visual Arts

Zvjezdana Prohaska, BFA (Bachelor of Fine Arts)
Education and Teacher Training Agency
Office Branch Rijeka
Ivan Uroda, M. Intelligent E-Business
Faculty of Economics
University of Rijeka
zvjezdana.prohaska@azoo.hr, ivan.uroda@gmail.com

Abstract - In this paper a computer software consisting of a paint program and a database of specially prepared forms in bitmap format for education of visual arts on the elementary school level is presented.

By using such computer software, a paint program (in graphic format) and forms in bitmap format, where it is possible not only to write text but also to draw lines, circles etc., interactive learning gets a new dimension,. From now on it is possible not only to ask questions in the form of yes or no, multiple choice or demand answers in full-text, but really to apply the solutions.

If the lesson is about composition in painting, it is possible to show a painting of a well known artist and not only ask which form of composition is predominant in the painting, but to demand that the pupil draws the chosen composition (e.g. horizontal, diagonal, pyramidal etc.) as a line in the painting itself. By presenting this software in the class and by sharing it, pupils can use it at home and by comparing own answers with the provided solutions in the software they can verify their knowledge of the topics discussed at school.

I. INTRODUCTION

Despite constant and numerous improvements and innovations in education systems worldwide, usage of computers and software in learning and teaching Visual Arts on the elementary school level, is not common and is therefore not widespread, not only in Croatia but also in Europe. Those methods should be closer to today's pupils who, in general, belong to generations that are surrounded by and accustomed to information and communication technology (ICT) at a young age and this is a fact which should not be neglected.

Taking into account that only a minority of teachers in Croatian elementary schools implement modern teaching methods and activities to motivate their pupils' learning and even a smaller proportion of them is using computers as an aid in learning and teaching Visual Arts, especially from 5th to 8th classes, the authors explored in an earlier research recommend software that would be most appropriate for learning and teaching Visual Arts in Croatia and abroad. After having finished that research authors have come to the conclusion that beside already existing software, it would be useful to develop a simple and portable computer program in graphic mode together with an art database to be suitable for teaching Visual Arts and what at the same time will be appropriate for interactive learning at home. [10]

Authors have also taken into consideration the fact that in a lot of schools only older and/or less powerful computers are at the disposal of teachers and pupils and even at home pupils do not always have very powerful computers. Because of that simpler software although running in graphic mode would be more appropriate for both occasions. [1]

Of course there exist a lot of excellent free paint programs in the market like Microsoft Paint being part of MS Windows, Paint.Net or GIMP even for Linux operating systems, but these computer programs are so sophisticated that they very often distract pupils, meaning that they are not enough focused on the main issue, i.e. learning about visual arts, but would rather explore all the possibilities of the paint programs by "playing" with them.

Therefore authors developed a simple computer program called PaintArt with an art database included, coded in MS Visual Basic for Windows according to the recommendation of a famous PC Magazine contributor and book author Steven Holzner. [4] Other authors and books also pointed out useful aspects of Visual Basic for Windows programming. [3], [7]

With that computer software which includes a paint program (in graphic format) and forms in bitmap format, it is possible not only to write text but also to draw lines, boxes, circles etc., interactive learning gets a new dimension. From now on it is possible not only to ask questions in the form of yes/no, multiple choice or demand answers in full-text, but really to apply the solutions.

If the lesson is about composition in painting, it is possible to show a painting of a well known artist and not only ask which form of composition is predominant in the painting, but to demand that the pupil draws the chosen composition (e.g. horizontal, diagonal, pyramidal etc.) as a line in the painting itself. [2], [8] By presenting this software in the class and by sharing it, pupils can use it at home and by comparing own answers with the provided

solutions in the software they can verify their knowledge of the topics discussed at school.

II. THE STRUCTURE OF THE PROGRAM

The program itself or the main menu is divided in seven parts (submenus), i.e. the Load File submenu and the Save File, Drawing Color, Print, Paste from Clipboard, Paste to Clipboard and Exit submenus. (Fig. 1)

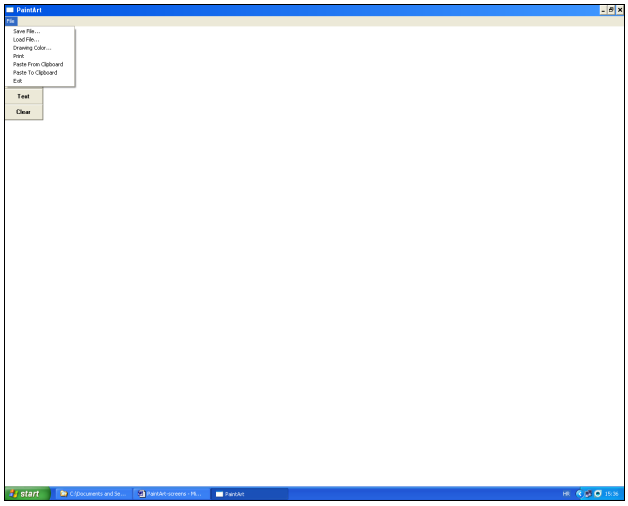


Figure 1. PaintArt main menu screenshot

In the Load File submenu it is possible to load or open a file in the current directory on the hard disk of the PC or even to change to a different directory. (Fig. 2)

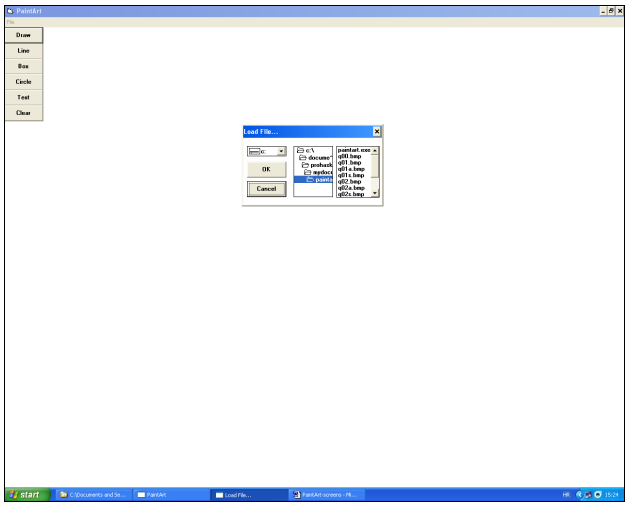


Figure 2. Load File menu screenshot

To activate scroll bars it is only necessary to click with the mouse and to drag. A Microsoft compatible mouse support is also built in the program.

The second or Save File As submenu provides the possibility to save a file under a specific name. As everywhere in the Program one has to confirm or to cancel the desired process. (Fig. 3)

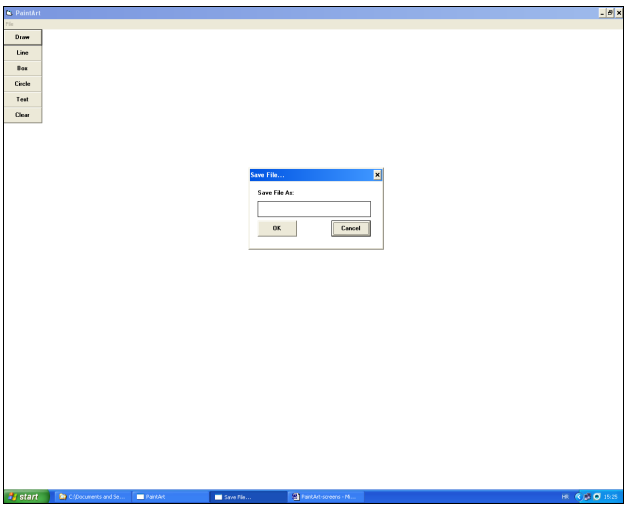


Figure 3. Save File menu screenshot

The next submenu is the Control Panel where it is possible to change the chosen color of the lines, circles, boxes etc. (Fig. 4)

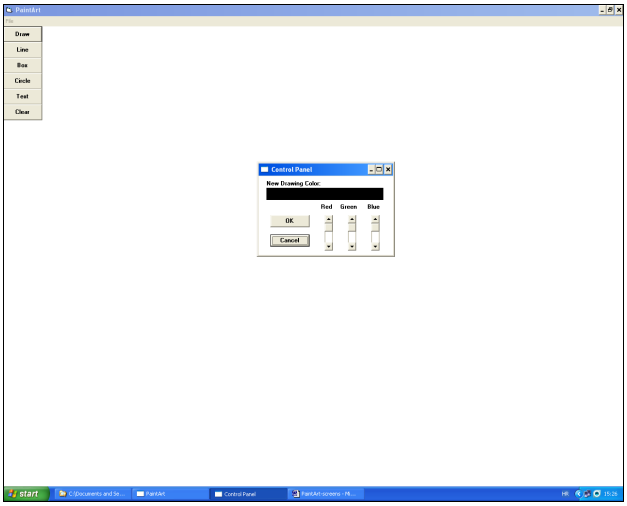


Figure 4. Control panel menu screenshot

In the next two submenus the Program takes also advantage from the Clipboard possibilities built in MS Windows operating system, i.e. it is possible to paste from and to the Clipboard any drawings, pictures and paintings in graphic or bitmap format. That is very useful for teaching Visual Arts because all paintings have to be presented as photographs in a suitable graphic format. After that there is of course an Exit submenu to leave the program immediately.

The graphic or painting capabilities of the PaintArt program are part of the main window and enable the user to draw lines freehand, straight lines, boxes, circles, text and of course to clear or erase everything. It is also possible to copy the content or the bitmap image to a file or to the Clipboard. (Fig. 5)

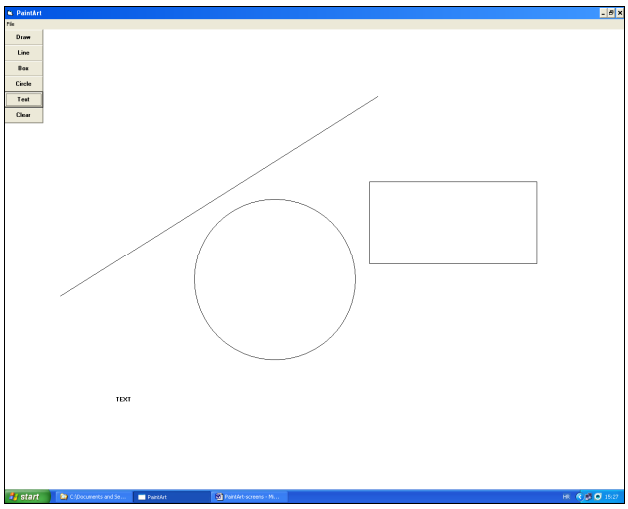


Figure 5. Drawings and text screenshot

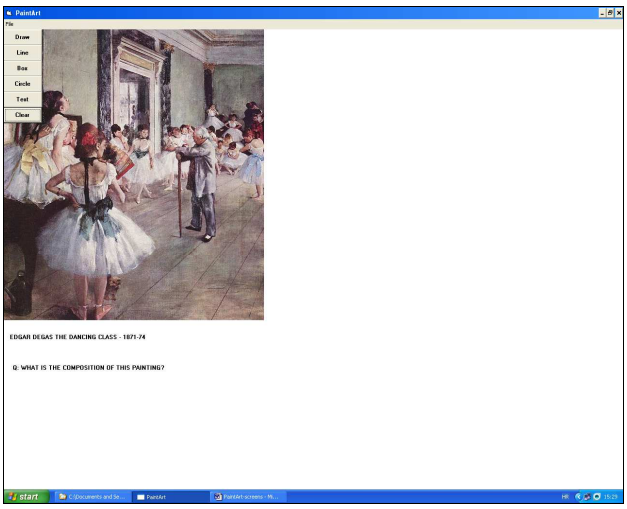


Figure 6. Painting number 1 – question screenshot

After having presented the general possibilities of the computer program PaintArt it is time to illustrate the built-in database of selected paintings and answers. The teacher of visual arts should focus on one subject at the same time, i.e. composition, perspective or contrast. It is recommended to choose paintings of well known artists which are a good example for the topic to be explored. To illustrate different forms of composition in painting the teacher has to choose appropriate paintings. To demonstrate composition authors have chosen three paintings of famous artists from different centuries. [6], [9]

As part of the computer software, the art database - composition includes the following paintings of famous artists: Edgar Degas - The Dancing Class, Paul Cezanne – The Large Bathers, Paolo Uccello - The Hunt in The Forest, etc. [5] The paintings or bitmap files are numbered in consecutive order as a question (q01.bmp), answer of the pupil (q01a.bmp) and solution provided by the teacher (q01s.bmp). After the first painting consisting of three files , the other paintings follow as well in raising order, i.e. file number 2, 3 etc.. (Fig. 2 above)

For illustration purposes authors will present in the next figure Edgar Degas' The Dancing Class. After having chosen the first question in the Composition database the picture of the painting with the adequate question appears on the screen of the computer. (Fig. 6)

In the left corner of the screen Edgar Degas painting appears with only necessary fundamental data, such as author, title and time or date when the painting was finished. Below that one can read the question about the composition of the painting. After seeing the painting and having read the question the pupil reloads the file numbered as question 1 answer (q01a.bmp) where she or he should write the appropriate answer and also draw in the picture or painting itself a line which should illustrate the sort of composition chosen. (Fig. 7)

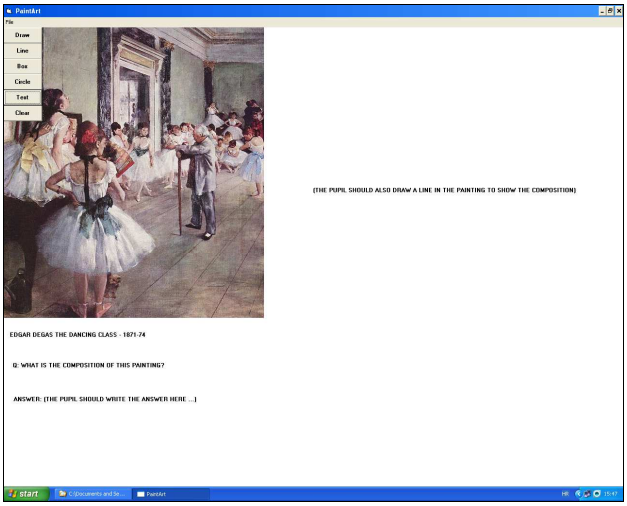


Figure 7. Painting number 1 – answer screenshot

Having done so the pupil can save her or his answer of the question in this or another file for later use and switch to the solution of the question the teacher has prepared and saved in another file. To do so she or he has to open or load the third file connected to the first question or the file containing the solution or the right answer to the imposed question. In this case it is the file called (q01s.bmp). In this file the correct answer is presented in verbal form, i.e. the composition is diagonal and the diagonal line is also drawn in the painting above. To see the line clearly it is sometimes necessary to change colors from dark (black) to light (white) like in this painting. (Fig. 8)

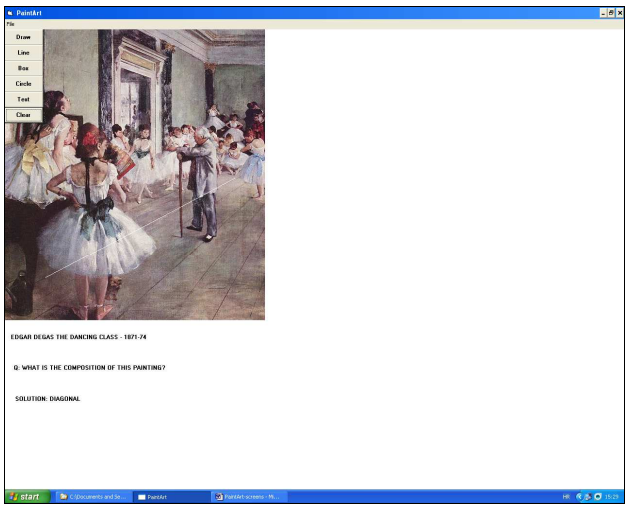


Figure 8. Painting number 1 – solution screenshot

III. INSTALLATION AND DELIVERY OF THE PROGRAM

Since PaintArt with included art database files in graphic format is an educational computer program its delivery and installation is kept simple. The program consists of only two application files, i.e. the PaintArt program file (paintart.exe) and the runtime module (vbrunxxx.dll). The Art database consists of at least 30 files in bitmap format (q01.bmp, q01a.bmp and q01s.bmp, etc.) for every single database called for example composition, perspective, contrast etc.

To install the program PaintArt it is necessary to copy all the files to a directory on the hard disk of a PC (for example C:\PaintArt) and to start the program by clicking the program file or icon (paintart.exe) with a Microsoft compatible mouse.

IV. CONCLUSION

In this paper a computer software consisting of a paint program called PaintArt and a database of specially prepared forms in bitmap format for education of visual arts on the elementary school level is presented.

By using such computer software, a paint program (in graphic format) and forms in bitmap format, where it is possible not only to write text but also to draw lines, circles etc., interactive learning gets a new dimension,. From now on it is possible not only to ask questions in the form of yes or no, multiple choice or demand answers in full-text, but really to apply the solutions.

Because of its design and computer language used (Microsoft Visual Basic for Windows) this program is running on PCs with Windows operating systems.

REFERENCES

[1] Čičin Šain, M., Čapko, Z., Vukmirović, S., Informatika za informatičko poslovanje, Ekonomski fakultet sveučilišta u Rijeci, Rijeka, 2008

[2] Damjanov, J., Likovna umjetnost, I. dio, Školska knjiga, Zagreb, 2009

[3] Halvorson, M., Microsoft Visual Basic Professional 6.0 Step by Step, Microsoft Press, Redmond, 1998

[4] Holzner, S., The Peter Norton Computing Company, Visual Basic, Brady Publishing, New York, 1991

[5] Hosack Janes, K., et al., Great Paintings, DK Dorling Kindersley Ltd., London, 2011

[6] Janson, A.F., History of Art, Prentice Hall, NJ, 6th ed., 2004

[7] Microsoft, Microsoft Visual Basic 6.0 Programmer's Guide, Microsoft Press, Redmond, 1998

[8] Peić, M., Pristup likovnom djelu, XII. Izd. Školska knjiga, Zagreb, 1991

[9] Preziosi, D., The Art of Art History, Oxford University Press, Oxford, 2009

[10] Prohaska, Z., Teaching and Learning Visual Arts by Using a Computer Program, in 33rd Internationl Convention MIPRO 2010, Conference Book, Computers in Education, Opatija, 2010

Laboratorij za emulaciju računalnih mreža i primjena u praktičnoj nastavi

Alen Jakupović*, Emanuel Pirić*
*Veleučilište u Rijeci, Poslovni odjel, Rijeka, Hrvatska
alen.jakupovic@veleri.hr
epiric@veleri.hr

Sažetak - Osnovni problem u praktičnoj nastavi iz područja računalnih mreža leži u činjenici da ju je teško organizirati za veću skupinu studenata. Ona zahtjeva veći broj različitih mrežnih uređaja, prostor za njihovo odlaganje, prostor za fizičku realizaciju topologija, organizaciju vremena kako bi svaki student mogao izvršiti vježbu itd. Rješenje ove problematike je u pretvaranju jedne radne stanice računalnoga kabineta u laboratorij za emulaciju računalnih mreža. Članak ukratko opisuje način izvedbe navedenoga laboratorija. Detaljnije se prikazuju neki nastavni sadržaji s odgovarajućim ishodima učenja, a koji se realiziraju kroz uspostavljeni laboratorij. Opisan je postupak provjere ishoda učenja primjenom laboratorija. Prikazana je primjena laboratorija u izradi studentskih projekata. U zaključku se navode moguća poboljšanja uspostavljenoga laboratorija, te se opisuje proširenje njegove primjene na neka druga područja.

I. UVOD

Osnovni cilj svake praktične nastave, pa tako i one iz domene računalnih mreža, je da se svakom studentu omogući učenje uz rješavanje stvarnih, praktičnih problema. U domeni računalnih mreža to znači da im se omogući umrežavanje različitih mrežnih uređaja i uspostava različitih mrežnih servisa. Neki od mrežnih uređaja su računala s različitim operacijskim sustavima, mrežni preklopnici (engl. switch), mrežni usmjerivači (engl. router), pristupne točke (engl. access point) itd., dok se od mrežnih servisa mogu spomenuti poslužitelji kao što su WEB, DNS, DHCP, FTP, e-mail, Proxy itd. Mrežni servisi su i uspostava P2P mreže, VPNa, sustava autentifikacije, rada na daljinu, videokonferencije, IP telefonije itd.

Za potrebe praktične nastave, na Veleučilištu u Rijeci postoje četiri računalna kabineta unutar kojih se nalazi prosječno 28 međusobno umreženih radnih stanica koje imaju pristup Internetu. Na radnim stanicama se nalaze različiti računalni programi koje studenti koriste za potrebe raznih kolegija. Studenti se svojim korisničkim računima prijavljuju na radnu stanicu na kojoj nemaju administratorske ovlasti.

Ovo razumljivo neposjedovanje administratorskih ovlasti uzrokuje veliku prepreku u praktičnoj nastavi kolegija računalne mreže, budući da za samo umrežavanje računala studenti trebaju posjedovati mogućnost promjene određenih postavki unutar operacijskoga sustava, a to bez administratorskih ovlasti nije moguće. Nadalje, administratorske ovlasti su potrebne i u slučaju uspostave mrežnih servisa budući da je tada potrebna instalacija različitih računalnih programa koji takve servise implementiraju.

Pored navedenoga problema administratorskih ovlasti koje studenti ne posjeduju, postoji i problem potrebe za različitim mrežnim uređajima nužnima za uspostavu računalne mreže. Računalna mreža se sastoji iz barem dva računala, što znači da bi se u jednome kabinetu moglo uspostaviti ukupno 14 računalnih mreža, a to znači ukupno 14 mrežnih uređaja jedne vrste. Za svih 14 mrežnih uređaja je potrebno osigurati strujni priključak, te do svakoga računala posebni mrežni kabel (njih ukupno 28). Već uspostava ovako jednostavne mreže pokazuje koliki su zahtjevi za prostorom, mrežnim kablovima, mrežnim uređajima i radnim stanicama. Uspostava imalo složenije računalne mreže (npr. umreženih pet računala preko različitih mrežnih uređaja) dovodi do toga da zbog vremenskoga ograničenja nastave, na njoj neće biti u mogućnosti efektivno sudjelovati svi studenti, a i praktična provjera njihova znanja je također upitna.

Pored svega navedenoga, postoji problem odlaganja cjelokupne mrežne opreme, ali i financijski problem njezine nabave i održavanja.

Rješenje za navedene probleme praktične nastave iz područja računalnih mreža je uspostava laboratorija na svakoj radnoj stanici unutar računalnoga kabineta. Ovime svaki student dobiva mogućnost da na svojoj radnoj stanici emulira računalnu mrežu koja se sastoji iz različitog broja računala i mrežnih uređaja, te uspostavi različite mrežne servise.

U nastavku članak donosi kratak opis načina uspostave laboratorija za emulaciju računalne mreže, prikaz nekih nastavnih sadržaja koji su realizirani kroz laboratorij, opis postupka provjere znanja i realizaciju studentskih projekata. Zaključak donosi moguća poboljšanja prikazanoga laboratorija, te moguće primjene na nekim drugim područjima.

II. LABORATORIJ ZA EMULACIJU RAČUNALNIH MREŽA

U osnovi postoje tri pristupa u dizajnu i validaciji neke strukture (topologije) računalne mreže: simulacija, emulacija i uspostava stvarne topologije. [6]

Simulacija računalne mreže predstavlja tehniku izrade i validacije računalne mreže pri kojoj se svi mrežni uređaji simuliraju, tj. njihovo ponašanje je modelirano računalnim programom.

Kod emulacije računalne mreže, pored simuliranih mrežnih uređaja, postoje i oni stvarni. Osim toga, cjelokupna emulirana računalna mreža se može dodati nekoj stvarnoj računalnoj mreži. Vanjski promatrač ne vidi razliku između emulirane i stvarne računalne mreže. [1]

Uspostava stvarne topologije predstavlja izradu računalne mreže primjenom stvarnih mrežnih uređaja.

Dvije su osnovne razlike između simulirane i emulirane računalne mreže. Simulirana računalna mreža se pokreće u simuliranom vremenu, dok se emulirana pokreće u stvarnom vremenu. U emuliranoj računalnoj mreži se ne mogu apsolutno ponoviti stanja mrežnih uređaja, dok se kod simulirane to može. [6]

Više o simulaciji i emulaciji računalnih mreža kao i o komparativnoj analizi različitih virtualizacijskih tehnika može se pronaći u [4].

Kao tehnika izrade računalnih mreža na praktičnoj nastavi izabrana je emulacija budući da ona podržava rješavanje stvarnih problemskih situacija koje studenti mogu očekivati u praksi.

Laboratorij za emulaciju računalnih mreža se sastoji iz dva glavna dijela: emulatora računala i emulatora fizičke razine računalne mreže.

Emulator računala predstavlja računalni program koji na računalu domaćinu (engl. host) omogućava pokretanje više drugih gost računala (engl. guest). Nad gost računalima studenti imaju potpune ovlasti, te na njima mogu biti instalirani različiti operacijski i aplikativni sustavi.

Emulator fizičke razine računalne mreže je računalni program koji omogućava umrežavanje emuliranih gost računala. To se postiže emulacijom samih mrežnih veza (kabela), te mrežnih preklopnika.

Kao emulator računala izabran je besplatni računalni program VirtualBox (www.virtualbox.org) koji omogućava emulaciju računala s različitim operacijskim sustavima (DOS, Windows, OS/2, FreeBSD, OpenBSD, Linux, Solaris, Mac OS X). [5] Više o emulaciji računala i analizi računalnih programa za emulaciju računala može se pronaći u [2].

Za emulator fizičke razine računalne mreže izabran je besplatni računalni program GNS3 (www.gns3.net). On omogućava grafičku izradu računalnih mreža koja se sastoji iz emuliranih računala, mrežnih preklopnika, usmjerivača itd. [3]

Pored navedenih emulatora, pripremljena su i tri emulirana računala: dva korisnička računala (jedno s Window XP operacijskim sustavom - licenca od Veleučilišta, a drugo s besplatnom TinyCore distribucijom Linuxa), te jedno računalo koje predstavlja mrežni usmjerivač (također TinyCore distribucija Linuxa). U slučaju potrebe za dodatnim korisničkim računalima ili mrežnim usmjerivačima, vrlo lako se izrade nova (metodom kloniranja unutar računalnog programa VirtualBox).

Konačno, laboratorij za emulaciju računalnih mreža koji je uspostavljen na svakoj radnoj stanici u računalnom kabinetu se sastoji iz dva besplatna računalna programa (emulator računala VirtualBox i emulator fizičke razine računalne mreže GNS3), te tri unaprijed pripremljena emulirana računala (dva korisnička računala i jedan mrežni usmjerivač).

III. PRIMJENA LABORATORIJA ZA EMULACIJU RAČUNALNIH MREŽA U PRAKTIČNOJ NASTAVI

Laboratorij za emulaciju računalnih mreža na Veleučilištu u Rijeci se primjenjuje od akademske godine 2012/2013. Trenutno su njime podržana tri osnovna nastavna sadržaja:

- izrada lokalne mreže (engl. LAN),
- izrada virtualne lokalne mreže (engl. VLAN)
- izrada više međusobno povezanih računalnih mreža (engl. WAN)

A. Izrada lokalne mreže

Nastavni sadržaj koji se odnosni na uspostavu lokalne mreže podijeljen je u četiri cjeline:

- umrežavanje dva računala koja posjeduju više mrežnih sučelja
- umrežavanje više računala primjenom više mrežnih preklopnika
- dodavanje emulirane računalne mreže stvarnoj mreži Veleučilišta u Rijeci
- omogućavanje korisničkim računalima pristup internetu

Navedene nastavne cjeline su podrška za sljedeće ishode učenja:

- postaviti ispravne IP adrese i mrežne maske na odgovarajuća mrežna sučelja unutar Windows XP i Linux operacijskog sustava
- provjeriti ispravnost uspostavljene računalne mreže
- izgraditi topologiju računalne mreže
- emuliranu topologiju računalne povezati sa stvarnom računalnom mrežom
- omogućiti korisničkom računalu pristup internetu

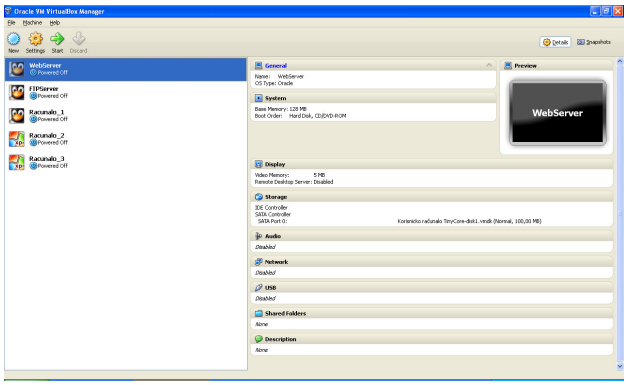
Kao primjer zadatka kojeg studenti rješavaju na praktičnoj nastavi može se uzeti sljedeća topologija računalne mreže:

"Potrebno je izraditi žičanu računalnu mrežu koja se sastoji iz tri računala s Linux operacijskim sustavom (nazivi računala su: WebServer, FTPServer i Racunalo_1) i dva računala s Windows XP operacijskim sustavom (nazivi računala su: Racunalo_2 i Racunalo_3). Adresa mreže je 199.0.0.0/8. WebServer ima dva mrežna sučelja i umrežen je preko 1. mrežne kartice. FTPServer ima dva mrežna sučelja i s računalnom mrežom je povezan preko 2. mrežne kartice. Ostala računala imaju po tri mrežna sučelja. Racunalo_1 je na mrežu povezan preko 1., Racunalo_2 preko 2. i Računalo_3 preko 3. mrežne kartice. Mrežu treba izvesti preko tri mrežna preklopnika (nazivi su im Preklopnik_1, Preklopnik_2 i Preklopnik_3). Na Preklopnik_1 se nalazi WebServer, na Preklopnik_2 se nalazi FTPServer, a na Preklopnik_3 se nalaze ostala računala."

Kao što se može primijetiti, zahtijevana računalna mreža se sastoji iz ukupno osam mrežnih uređaja (pet korisničkih računala i tri mrežna preklopnika). Ovakvu računalnu mrežu bi bilo teško izvesti sa stvarnim mrežnim uređajima i uz uvjet da je samostalno izradi svaki student. Uspostavljeni laboratorij za emulaciju računalnih mreža omogućava da svaki student na svojoj radnoj stanici izradi ovu složenu računalnu mrežu i to upravo onako kako bi je izradio i sa stvarnim mrežnim uređajima.

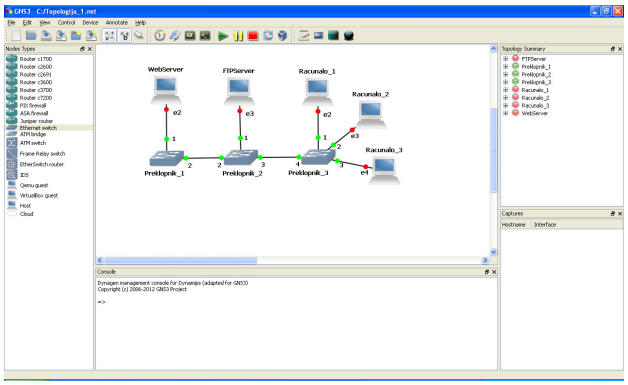
Postupak kojeg studenti provode za uspostavu opisane računalne mreže je sljedeći:

1. kroz računalni program VirtualBox studenti pripremaju pet računala (tri s Linux operacijskim sustavom i dva s operacijskim sustavom Windows XP). Slika 1. prikazuje pet emuliranih računala.



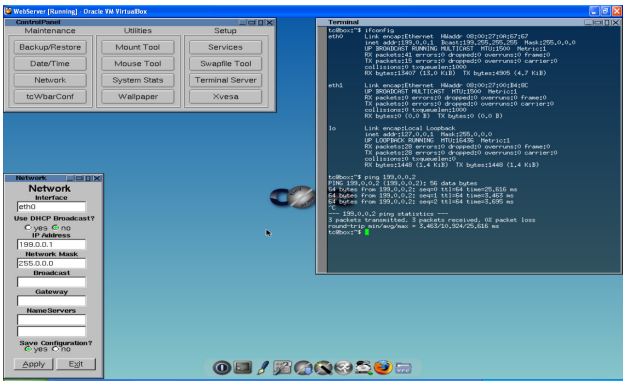
Slika 1. Pet emuliranih računala

2. unutar računalnog programa GNS3 studenti ubacuju pet emuliranih računala, te im dodjeljuju odgovarajući broj mrežnih sučelja (kako to zadatak zahtijeva). Potom emulirana računala dodaju u topologiju, u koju dodaju i tri mrežna preklopnika. Nazive mrežnih uređaja usklađuju s nazivima iskazanim u zadatku. Zatim se emulirana računala povezuju s mrežnim preklopnicima kako to zahtijeva zadatak, uz poštivanje mrežne kartice preko koje računalo treba biti umreženo. Potom se međusobno povezuju i mrežni preklopnici. Slika 2. prikazuje gotovu zahtijevanu topologiju računalne mreže



Slika 2. Zahtijevana topologija računalne mreže

3. pokreću se emulirana računala iz računalnog programa GNS3. Nakon pokretanja, potrebno je u svako emulirano računalo (od njih pet), na odgovarajuću mrežnu karticu postaviti ispravnu IP adresu i mrežnu masku (kako to zahtijeva zadatak). Potom treba provjeriti odgovarajućim naredbama ispravnost postavki i povezanost računala.



Slika 3. Emulirano računalo "WebServer"

Slika 3. prikazuje jedno od pet pokrenutih emuliranih računala (WebServer) s odgovarajućom IP adresom i mrežnom maskom, te provjerom povezanosti s računalom u mreži koje posjeduje IP adresu 199.0.0.2.

B. Izrada virtualne lokalne mreže

Nastavni sadržaj koji se odnosi na uspostavu virtualne lokalne mreže sastoji se u izradi više virtualnih lokalnih mreža primjenom više mrežnih preklopnika.

Za ovaj nastavni sadržaj postavljeni su sljedeći ishodi učenja:

- postaviti ispravne identifikatore virtualnih lokalnih mreža i odgovarajuće tipove priključaka na mrežnim preklopnicima
- izgraditi topologiju računalne mreže koja uključuje više virtualnih lokalnih mreža i mrežnih preklopnika

Jedan od praktičnih zadataka koje studenti rješavaju unutar nastavnoga sadržaja koji se odnosi na izgradnju virtualnih mreža je sljedeći zadatak:

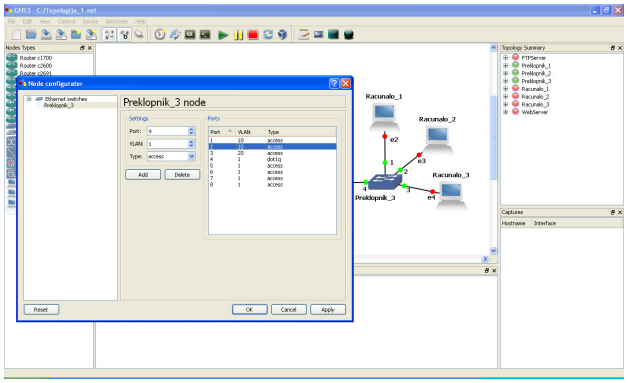
"Potrebno je izraditi žičanu računalnu mrežu VLAN10 koja se sastoji iz tri računala s Linux operacijskim sustavom (nazivi računala su: WebServer, Racunalo_1 i Racunalo_2) i VLAN20 s dva računala pod Windows XP operacijskim sustavom (nazivi računala su: FTPServer i Racunalo_3). WebServer ima jedno mrežno sučelje i njegova IP adresa je 11.11.11.1/24, dok računala u VLAN20 pripadaju mreži 192.0.0.0/16. FTPServer ima dva mrežna sučelja i s računalnom mrežom je povezan preko 2. mrežne kartice. Ostala računala imaju po tri mrežna sučelja. Racunalo_1 je na mrežu povezan preko 1., Racunalo_2 preko 2. i Računalo_3 preko 3. mrežne kartice. Mrežu treba izvesti preko tri mrežna preklopnika (nazivi su im Preklopnik_1, Preklopnik_2 i Preklopnik_3). Na Preklopnik_1 se nalazi WebServer, na Preklopnik_2 se nalazi FTPServer, a na Preklopnik_3 se nalaze ostala računala."

I ovaj zadatak zahtijeva ukupno osam mrežnih uređaja (pet računala i tri mrežna preklopnika). Prilikom rješavanja zadatka, studenti provode isti postupak koji je

već opisan za zadatak uspostave lokalne mreže, uz sljedeći dodatak:

1. priključcima na mrežnim preklopnicima na koja su spojena računala, treba postaviti odgovarajući tip (u ovom slučaju to je tip *access*), i odgovarajuću identifikaciju (za VLAN10 je to broj 10, a za VLAN20 broj 20).

2. priključcima na mrežnim preklopnicima na koja su spojeni drugi mrežni preklopnici treba postaviti odgovarajući tip (u ovom slučaju to je *dot1q*). Slika 4. prikazuje postavke na jednom mrežnom preklopniku.



Slika 4. Postavke mrežnog preklopnika "Preklopnik_3"

3. po završetku postavljanja mrežnih preklopnika potrebno je provjeriti povezanost računala u odgovarajućim virtualnim lokalnim mrežama.

C. Izrada više međusobno povezanih računalnih mreža

Nastavni sadržaj koji se odnosi na međusobno povezivanje računalnih mreža sastoji se iz dvije cjeline:

- povezivanje više računalnih mreža
- povezivanje više virtualnih računalnih mreža

Ishodi učenja koji su postavljeni za ovaj nastavni sadržaj su:

- postaviti ispravne tablice usmjeravanja na mrežnim usmjerivačima
- izraditi topologiju računalne mreže koja uključuje povezivanje više virtualnih lokalnih mreža
- izraditi topologiju računalne mreže koja uključuje povezivanje više računalnih mreža

Primjer praktičnog zadataka povezanoga s nastavnim sadržajem međusobnog povezivanja više računalnih mreža je:

"Potrebno je izraditi tri žičane računalne mreže. U 1. računalnoj mreži se nalazi jedno računalo s Windows XP operacijskim sustavom (Racunalo_1). Njegova IP adresa je 100.100.100.100/24. U 2. i 3. računalnoj mreži se nalaze po dva računala s Linux operacijskim sustavom (nazivi računala su: Racunalo_2, Racunalo_3, Racunalo_4 i Racunalo_5). Adresa 2. mreže je 192.168.0.0/16, a 3. mreže je 10.0.0.0/24. Potrebno je međusobno povezati ove tri mreže kako bi računalo iz jedne mreže bilo dostupno iz ostalih mreža. Potrebno je koristiti dva mrežna usmjerivača (između 1. i 2. mreže postaviti Usmjerivac_12, a između 2. i 3. mreže Usmjerivac_23). Broj mrežnih sučelja računala je proizvoljan."

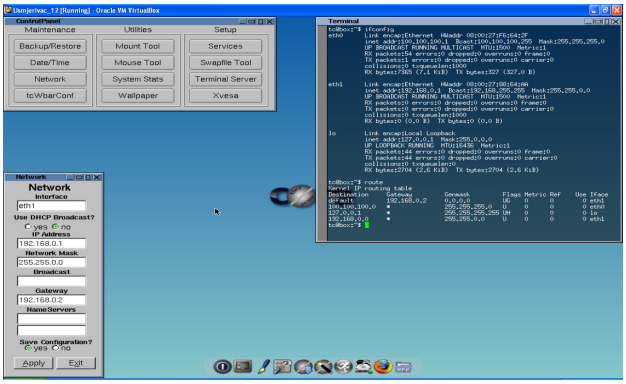
Rješenje prikazanoga zadatka zahtijeva ukupno sedam računala (pet korisničkih računala i dva mrežna usmjerivača), te tri mrežna preklopnika. Za definiranje pet korisničkih računala studenti koriste već opisani postupak. Posebno za ovakvu vrstu zadataka studenti trebaju učiniti sljedeće:

1. kroz računalni program VirtualBox studenti pripremaju dva mrežna usmjerivača (posebna vrsta emuliranoga računala s Linux operacijskim sustavom).

2. unutar računalnoga programa GNS3 studenti definiraju koliko mrežnih sučelja trebaju imati mrežni usmjerivači. U našem primjeru mrežni usmjerivači povezuju dvije računalne mreže pa prema tome trebaju imati dva mrežna sučelja.

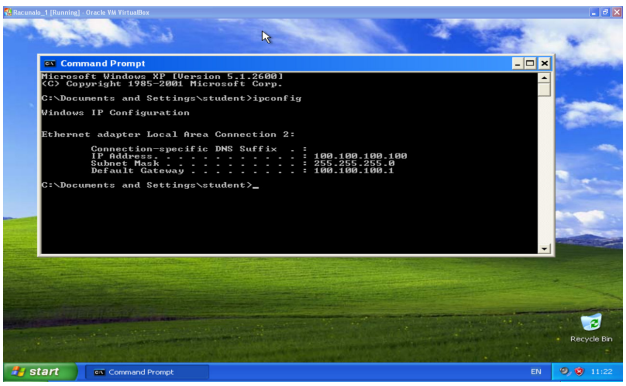
3. povezati definirana korisnička računala i mrežne usmjerivače preko odgovarajućih mrežnih sučelja primjenom mrežnih preklopnika.

4. definirati odgovarajuće IP adrese na mrežnim usmjerivačima, odnosno odgovarajuće tablice usmjeravanja. Slika 5. prikazuje definiranu tablicu usmjeravanja za mrežni usmjerivač Usmjerivac_12.



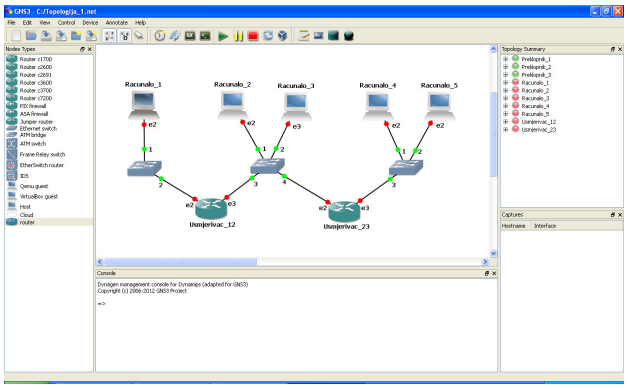
Slika 5. Tablica usmjeravanja mrežnog usmjerivača Usmjerivac_12

5. definirati odgovarajuće IP adrese na korisničkim računalima vodeći brigu o IP adresi mrežnog usmjerivača preko kojega je korisničko računalo povezano s ostatkom računalne mreže. Slika 6. prikazuje definirane IP adrese za korisničko računalo Racunalo_1



Slika 6. IP adrese korisničkog računala Racunalo_1

6. provjeriti uspostavljenu računalnu mrežu primjenom naredbi *ping* i *tracert* ili *traceroute*.



Slika 7. Konačno rješenje zadatka

Slika 7. prikazuje konačno rješenje međusobnoga povezivanja tri računalne mreže.

IV. PRIMJENA LABORATORIJA ZA EMULACIJU RAČUNALNIH MREŽA U PROVJERI ISHODA UČENJA

Ishodi učenja navedeni u prethodnim poglavljima se lako provjeravaju primjenom laboratorija za emulaciju računalnih mreža. Naime, tijekom provjere, svaki student dobiva deskriptivan opis zahtjeva koje računalna mreža treba zadovoljiti. Potom studenti definiraju emulirana računala unutar računalnoga programa VirtualBox, uz korištenje gotovih predložaka (po jedno za korisnička računala s Windows XP i Linux operacijskim sustavom, te jedan mrežni usmjerivač). Nakon toga, studenti izrađuju računalnu mrežu unutar računalnog programa GNS3. Po završetku izrade ili isteku propisanoga vremena za rješavanje zadatka odmah se provjerava valjanost računalne mreže iz perspektive postavljenih ishoda učenja.

U slučaju da se provjera sastoji iz više zadataka, studenti prelaze na naredni zadatak tek nakon pregleda prethodnoga. Razlog je taj što se u narednom zadatku mogu iskoristiti ista emulirana računala koja su se koristila u prethodnom zadatku (ne moraju se stvarati nova čime se smanjuje vrijeme pripreme emuliranih računala).

Prilikom definiranja zadataka, svakako treba voditi brigu o broju emuliranih računala potrebnih u njegovu rješenju. Naime, svako emulirano računalo zahvaća resurse fizičkoga računala domaćina, te se može dogoditi da količina resursa ne bude dovoljna za količinu potrebnih emuliranih računala.

Zbog dijeljenja resursa fizičkoga računala domaćina, ponekad se znaju dogoditi ispadi operacijskoga sustava, iako rijetko. Kako bi se ublažile posljedice na rješavanje zadatka, prije pokretanja emuliranih računala poželjno je spremiti računalnu mrežu izrađenu u računalnom programu GNS3 jer je ispad operacijskog sustava ponekad takav da zahtijeva njegovo ponovno pokretanje.

Ovakva jednostavna i zorna provjera ishoda učenja bez laboratorija za emulaciju računalnih mreža u nekim slučajevima uopće ne bi bila izvediva.

V. PRIMJENA LABORATORIJA ZA EMULACIJU RAČUNALNIH MREŽA U IZRADI STUDENTSKIH PROJEKATA

Jedna od nastavnih aktivnosti koje provode studenti je i samostalna izrada projektnih zadataka. Izrada projektnoga zadatka se sastoji iz izrade projektne dokumentacije, praktične realizacije projekata i njegove obrane. Projektni zadaci su iz domene uspostave nekog mrežnog servisa (npr. različiti mrežni poslužitelji - Web, DNS, DHCP, Proxy, e-mail; P2P mreže; VPN; sustav autentifikacije; rada na daljinu; videokonferencije; IP telefonije itd.).

Temeljni cilj projektne dokumentacije je uspostaviti osnovni terminološki sustav potreban za razumijevanje projektnog zadatka, opisati zahtjeve koje treba zadovoljiti mrežni servis, detaljno prikazati njegovu uspostavu, te topologiju računalne mreže unutar koje će biti implementiran.

Praktična realizacija projektnog zadatka, odnosno mrežnog servisa, zahtijeva instalaciju odgovarajućih računalnih programa. U tu svrhu se koriste emulirana računala i računalni program VirtualBox. U svrhu uspostave topologije računalne mreže unutar koje će djelovati mrežni servis koriste se druga emulirana računala, mrežni uređaji i računalni program GNS3. Cjelokupnu implementaciju mrežnoga servisa studenti mogu izraditi na svojim osobnim računalima izvan računalnih kabineta i termina nastave.

Za potrebe obrane projektnih zadataka studenti svoje projekte trebaju implementirati i na radnim stanicama u računalnom kabinetu. Ovo je jednostavno izvedivo budući da studenti sa sobom mogu donijeti sva emulirana računala koja su prethodno pripremili, te samu topologiju računalne mreže. Sve ovo se ubacuje u radnu stanicu kroz računalne programe VirtualBox i GNS3. Obrana projekta se sastoji iz kratke prezentacije sadržaja projektne dokumentacije, te prezentacije mrežnog servisa u radu.

Kao što je vidljivo, laboratorij za emulaciju računalnih mreža je omogućio praktičnu realizaciju projektnih zadataka neovisno izvodi li se ona na nastavi u računalnom kabinetu ili na osobnom računalu studenta na kojemu također treba biti pripremljen laboratorij. Izrazito važna je činjenica što je laboratorij omogućio laku prenosivost realiziranih projekata čime je omogućena njihova javna prezentacija.

VI. ZAKLJUČAK

U članku je prikazan laboratorij za emulaciju računalnih mreža iz perspektive njegove uspostave, primjene u nastavi, provjere ishoda učenja i primjene u praktičnim projektnim zadacima. Laboratorij se u nastavi na Veleučilištu u Rijeci primjenjuje od akademske godine 2012/1013. Primjenjuje se unutar svih segmenata nastave koji su u članku opisani. Dosadašnja iskustva pokazuju da laboratorij zadovoljava postavljene zahtjeve te je ukazao i na mogućnost njegove primjene u nekim drugim nastavnim sadržajima iz domene računalnih mreža. To se prije svega odnosi na neke mrežne servise koji će se u budućnosti svakako nastojati pokriti nastavnim sadržajem (npr. DHCP poslužitelj, DNS poslužitelj, WEB poslužitelj, NAT, vatrozid, VPN itd.).

Laboratorij za emulaciju računalnih mreža se može primijeniti i u sadržajima nekih drugih kolegija. Primjera radi, on se može koristiti za zorni prikaz malicioznih napada na neke mrežne servise i njihovu obranu. Može se primijeniti i za praktičnu realizaciju sustava obrane. Sve su ovo sadržaji iz domene sigurnosti informacijskih sustava.

Laboratorij se može primijeniti i u prikazu implementacije izrađenih programskih rješenja. Naime, poznata je problematika instalacije distribuiranih programskih rješenja, tj. programskih rješenja čije se komponente ne nalaze samo na jednom računalu, već se nalaze na više njih. Laboratorij bi omogućio zorni prikaz problematike i rješenja ovoga problema. Ovdje izneseni sadržaji spadaju u domenu područja razvoja informacijskih sustava.

Iz perspektive domene računalnih mreža laboratorij će se dalje nastaviti razvijati u tri osnovna smjera. Budući da je računalni program GNS3 otvorenoga koda, namjera je izmijeniti ga kako bi njegovo sučelje bilo bolje prilagođeno potrebama nastave (npr. trenutno se na sučelju nalazi veliki broj mrežnih uređaja čija uporaba zahtijeva posjedovanje licence odgovarajućeg operacijskog sustava). Također bi se pokušalo pronaći rješenje za prikaz bežičnih računalnih mreža koje trenutno nisu podržane. Posljednji pravac razvoja laboratorija je priprema emuliranih računala koja bi podržala neke mrežne servise bitne za nastavni sadržaj (već su spomenuti DHCP poslužitelj, DNS poslužitelj itd., ali tu je i bolja emulacija VLAN-ova).

LITERATURA

[1] A. Jakupović, M. Knežević, "Emulacija računalne mreže primjenom računalnih programa GNS3 i VirtualBox", Zbornik radova konferencije Elektroničke komunikacijske tehnologije i norme u informatici - KOM 2012, Zagreb, Hrvatska, pp. 105-112.

[2] CERT, LS&S, Virtualizacija računala, http://www.cert.hr/sites/default/files/NCERT-PUBDOC-2009-12-285.pdf, 15.12.2012.

[3] Introduction to GNS3, http://www.gns3.net/gns3-introduction/, 15.12.2012.

[4] M. Zec, "Simulacija i emulacija računalnih mreža i raspodijeljenih sustava primjenom virtualizacijskih mehanizama", http://www.fer.unizg.hr/_download/repository/Zec,_KDI.pdf, 15.12.2012.

[5] Oracle VM VirtualBox User Manual, http://download.virtualbox.org/virtualbox/UserManual.pdf, 15.12.2012.

[6] S. Guruprasad, R. Ricci, J. Lepreau, "Integrated Network Experimentation using Simulation and Emulation", In Proceedings of the First International Conference on Testbeds and Research Infrastructures for the DEvelopment of NeTworks and COMmunities - TRIDENTCOM '05, Washington, DC, USA, pp. 204-212.

Otvoreni edukacijski *online* program u funkciji kvalitetnije dostupnosti službenih informacija

Renata Pekorari, Tamara Horvat
Hrvatska informacijsko-dokumentacijska referalna agencija - Hidra, Zagreb, Hrvatska
pekorari@hidra.hr, thorvat@hidra.hr

Sažetak – Jedna od osnovnih zadaća Hrvatske informacijsko-dokumentacijske referalne agencije - Hidre, stručne službe Vlade Republike Hrvatske, je osiguranje dostupnosti javnih službenih informacija svim zainteresiranim korisnicima pod jednakim uvjetima, nepristrano, besplatno i u otvorenom pristupu. Kako bi ostvarila tu zadaću, Hidra u svoj rad uključuje i različite edukacijske programe. Procjena potreba i analiza korisnika dovela je do zaključka da je nužan novi pristup korisnicima Hidrinih usluga te se krenulo u izradu programa za e-učenje. Osim toga, budući da je većina Hidrinih proizvoda i usluga dostupna *online*, nametnuo se zaključak da su primjerene i upute za njihovo korištenje u e-obliku. U radu će se opisati projektni plan prvog Hidrinog programa za e-učenje pod nazivom «Naučite kako, brzo i lako, do pravnih propisa!». Prikazat će se opis programa i nastavnog problema, karakteristike mogućih korisnika, ishodi učenja i nastavni materijali te će se evaluirati tehnologija koja će se koristiti za realizaciju programa kao i potrebni resursi. Posebno će se razmotriti mogući problemi i rizici tijekom rada na programu.

I. UVOD

Hrvatska informacijsko-dokumentacijska referalna agencija (u daljnjem tekstu Hidra) bavi se prikupljanjem, obradom, pohranom i diseminacijom službenih informacija Republike Hrvatske. Jedna od osnovnih zadaća joj je osigurati dostupnost, pod jednakim uvjetima i nepristrano, besplatno i u otvorenom pristupu, javnim službenim informacijama Republike Hrvatske svim zainteresiranim korisnicima [1]. Pružajući informacije (i) putem portala, (ii) tiskanim publikacijama te (iii) odgovorima na pojedinačne upite korisnika, Hidra omogućuje građanima ostvarivanje prava na pristup informacijama koje je jedno od osnovnih ljudskih prava, a u Republici Hrvatskoj i Ustavno i zakonsko pravo [2]. Kako bi što bolje ostvarila tu zadaću, unapređujući stalno kvalitetu dostupnosti službenih informacija i svoje usluge, Hidra u svoj rad uključuje i različite edukacijske i promotivne programe pomoću kojih svojim polaznicima približava svijet službenih informacija te načine na koje najlakše i najjednostavnije mogu do njih doći. Najveći dio tih programa posvećen je predstavljanju Hidrinog portala, odnosno, baza podataka s prezentacijom svih mogućnosti pretraživanja, odnosno, dolaženja do podataka. Te su programe anketirani polaznici uvijek ocjenjivali izvrsnim ocjenama odnosno vrlo korisnima za njihov rad.

Do listopada 2012. godine provodili su se isključivo klasični edukacijski programi. Međutim, malobrojni kadar uposlen na izvedbama Hidrinih edukacijskih i promotivnih programa, nije bio u mogućnosti klasičnim izvedbama pokriti sve potrebe što je, uz česte poticaje Hidrinih korisnika, bio osnovni razlog odluke uvođenja potpunih edukacijskih asinkronih *online* programa [3], otvorenih i lako dostupnih putem Hidrinog portala svima zainteresiranima. Dodatni smisao uvođenju e-učenja pridonijela je i činjenica da se predstavljaju sadržaji koji su dostupni isključivo putem Interneta.

Predstavljanje tražilice pravnih propisa Republike Hrvatske Cadial (cadial.hidra.hr) jedan je od najizvođenijih Hidrinih klasičnih programa. Radi se o programu namijenjenom svima zainteresiranima za pretraživanje pravnih propisa Republike Hrvatske. S obzirom da se tražilica Cadial, odnosno, baza pravnih propisa, od svih Hidrinih informacijskih resursa najviše koristi, izabrana je upravo ta tražilica, mogućnosti koje pruža u pretraživanju te prezentacija najjednostavnijeg dolaženja do željenih reuzultata pretraživanja kao tema za prvi *online* program. Na temelju procjene potreba za e-učenjem i analize potencijalnih korisnika te analizom sadržaja, utvrdivši ciljeve i ishode učenja tim, sastavljen od dva predmetna stručnjaka iz Hidre, izradio je potpuni asinkroni edukacijski *online* program pod nazivom «Naučite kako, brzo i lako, do pravnih propisa!» (blog.hidra.hr). Program je dostupan na početnoj stranici tražilice, naziva *Video upute za pretraživanje*.

II. OPIS PROGRAMA

A. Opis nastavnog problema

Građanima, odnosno svima zainteresiranima, u skladu s pravom na pristup informacijama, treba omogućiti jednostavan, brz, besplatan i trajan pristup službenim informacijama. Dio tih službenih informacija su i pravni propisi, odnosno zakoni i ostali pravni akti po kojima su uređena pitanja u jednoj državi. Građanima, bilo u poslovne, bilo u privatne svrhe, mogu svakodnevno zatrebati pojedini propisi, a tražilica Cadial im omogućuje jednostavan *online* pristup. S obzirom na dosadašnje iskustvo i praksu, građani u mnoštvu informacija teško prepoznaju onu relevantnu i najkvalitetniju, a pravni su propisi, osim toga, jedan iznimno zahtjevan skup podataka koji treba prezentirati na razumljiv i jednostavan način.

S obzirom na sve navedeno, tim za izradu programa postavio si je za cilj prezentirati tražilicu Cadial i sve njezine prednosti te načine korištenja tražilice putem svima otvorenog i dostupnog asinkronog *online* programa čiji je cilj naučiti korisnike jednostavnom i lakom pronalaženju propisa.

B. Opis korisnika

Korisnici mogu biti upravo svi kojima su potrebni pravni propisi, oni koji su pretraživali Internet i došli do tražilice Cadial više, manje slučajno, ili oni koji su znali za tražilicu i da putem nje mogu doći do kvalitetne i cjelovite informacije. Očekuje se da će ih većina posegnuti za ponuđenim edukacijskim programom kako bi naučili pretraživati ili, ako već koriste tražilicu, usavršili svoja znanja i vještine.

S obzirom da je program otvorenog tipa i namijenjen svima zainteresiranima te dostupan s početne stranice tražilice, lepeza potencijalnih korisnika je vrlo široka s obzirom na dob, obrazovni status, vještine korištenja računala i snalaženja u Internetu i virtualnom svijetu te komunikacije putem diskusijskih foruma ili sličnih kanala. Sve nabrojeno činilo je izradu programa posebno zahtjevnom jer je morao biti prilagođen na mnogo načina.

Da bi mogli pristupiti programu za e-učenje, korisnici moraju imati samo osnovna znanja u radu na računalu i korištenju Interneta.

C. Ishodi učenja

Kao ishod učenja očekuje se da će korisnici, nakon što prođu edukacijski program, moći suvereno koristiti tražilicu Cadial i vrlo lako pronalaziti pravne propise koji im trebaju.

Korisnici mogu sami provjeriti jesu li savladali uporabu tražilice kroz samo njezino korištenje. Ako ne dobiju optimalne rezultate, uvijek se ponovo mogu vratiti u program za e-učenje i ponoviti određene korake kako bi kod stvarnog pretraživanja dolazili do svih sadržaja koji su im potrebni.

D. Opis modula

Program sadržava 12 modula, kratkih, jednostavnih za usvajanje i jasnih naziva koji opisuju njihov sadržaj. Pročitavši naziv, korisnici znaju što će u kojem modulu pronaći, a mogu se uvijek ponovo vratiti u modul koji im zatreba. Svaki modul sadrži posebne ishode vezane za savladavanje određenih vještina vidljivih iz naziva modula.

Svi su moduli koncipirani tako da ih korisnik može pregledavati kao samostalne cjeline ovisno o svojim potrebama, a posloženi su od općih ka posebnim temama. Izrađeni su pomoću alata screencast (na engleskom jeziku izraz za alat za snimanje događanja na monitoru računala). Uvodni modul je statičan, odnosno prikazuje sadržaj PowerPoint prezentacije, dok su ostalih 11 modula snimke sučelja, rezultata pretrage te radnji koje je potrebno obaviti pomoću tražilice Cadial za dobivanje optimalnih rezultata pretrage. Svaki modul prikazuje konkretne primjere pretraživanja i upute kako pretraživati. Uz sliku, svaki snimak praćen je i zvučnim uputama, pa se tako spojem slike, zvuka i teksta pokušalo što više približiti univerzalnom dizajnu [4]. Postoji i dodatni, tzv. nulti modul u kojem se kroz kratki video prilog, snimljen za televiziju, daje širi kontekst tražilice i njezina uloga u ostvarivanju prava na pristup informacijama. Svi moduli i popratni sadržaji postavljeni su na blog.

E. Resursi

Za tehničku izvedbu programa, kao što je i planirano, nisu potrošena dodatna novčana sredstva.

Koristili su se besplatni screencast alat *Screencast-O-Matic* (www.screencast-o-matic.com) te blog *WordPress* (hr.wordpress.org). U izradi programa, osmišljavanju sadržaja, izboru alata i implementaciji sadržaja sudjelovala su dva predmetna stručnjaka, koja su u program unijela svoje iskustvo iz Hidrinih klasičnih edukacijskih programa te znanja koja su stekli u sklopu programa *E-learning Course Design* u E-learning akademiji (jednogodišnji obrazovni program o e-učenju koji od 2004. godine provodi CARNet - Hrvatska akademska istraživačka mreža). Povremeno, pri tehničkoj izvedbi, konzultirani su informatički stručnjaci.

S druge strane, korisnici za korištenje programa, osim poznavanja osnova rada na računalu i korištenja Interneta, ne trebaju nikakvo posebno tehničko znanje da bi pratili postavljene module. Od opreme trebat će računalo, priključak na Internet i zvučnike. Takvi, minimalni tehnički zahtjevi su prednost i stoga što je krug potencijalnih korisnika vrlo širok, pa se ne može pretpostaviti da bi svi mogli pratiti program kada bi on bio tehnički zahtjevniji.

F. Nastavni materijali

Posebno pažljivo odabrani su sadržaji programa pod geslom „više je previše“. Uključili su se samo oni sadržaji koji su neophodni za realizaciju cilja nastave kao i postizanje definiranih ishoda učenja. Sadržaji su organizirani po strukturnoj logici. Prema Shepherdu (2010) izrađen je potpuni asinkroni *online* program slijedeći strategiju izlaganja *(exposition)* [5]. Obrazovna poruka upućena je korisnicima uvodnim, jasno napisanim i u cjeline odgovarajuće veličine organiziranim tekstom, koji ih uvodi u gradivo i u kojem su navedeni očekivani ishodi učenja te pregled sadržaja. Koristila su se maksimalno moguće pravila univerzalnog dizajna kako bi program bio pristupačan što većem broju potencijalnih korisnika. U njemu prevladavaju video zapisi – simulacije pretraživanja koje prate audio zapisi.

G. Procjena tehnologije

Za izradu modula, kao što smo već napomenuli, koristio se screencast. Screencast podrazumijeva snimanje radnji koje se odvijaju na zaslonu računala, najčešće popraćene zvučnim zapisom: govorom snimatelja, zvukom koji proizvodi program koji se izvodi i snima ili zvukom iz neke druge datoteke. Ova definicija screencasta u potpunosti je odgovarala potrebama edukacijskog programa o kojem govorimo jer je osnovni njegov cilj naučiti potencijalne korisnike kako se koristi jedna internetska tražilica. Snimanjem radnji koje se odvijaju na zaslonu računala na najjednostavniji i realističan način omogućuje se korisnicima vidjeti koje je sve radnje potrebno učiniti da bi se došlo do rezultata pretraživanja. Za procjenu tehnologije koristio se model SECTIONS [6]:

S (Students = polaznici, korisnici). Potencijalni korisnici mogu biti svi koji se nađu u potrazi za pravnim propisom.

Za pretpostaviti je da je to populacija odraslih koja poznaje osnove korištenja računala i Interneta. Za pregledavanje screencasta (snimljenih modula) korisnicima nisu potrebna posebna predznanja. Nisu potrebna nikakva posebna tehnička znanja, niti poznavanje stranog jezika.

E (Ease of use = lakoća korištenja). Snimanje se izvodi jednostavno: autor na računalu izvodi radnje koje želi snimiti uz popratno objašnjavanje ukoliko alat omogućava snimanje zvuka. Softverski zahtjevi kod korisnika najčešće su minimalni, za pregled videa potreban je internetski preglednik i program kao što je *QuickTime, Flash* ili *Windows Media Player*. Screencast je moguće gledati izravno putem Interneta ili preuzeti na svoje računalo i pogledati kasnije. Alat se često koristi, provjeren je, lak za upotrebu i od strane dizajnera programa i korisnika. U upotrebi je niz godina i izrađeni su praktični vodiči (tutorijali) namijenjeni osobama koje žele izrađivati sadržaje putem screencastinga.

C (Costs = troškovi). Na tržištu se mogu pronaći brojni screencast alati (besplatni ili komercijalni). Dizajner programa se može odlučiti i za neki komercijalni alat, cijene se kreću od nekoliko dolara pa do nekoliko stotina dolara, ali troška za krajnjeg korisnika nema, osim, naravno, troška korištenja Interneta. Besplatni alati se također mogu koristiti i dati izvrsne rezultate, a u tom slučaju niti dizajner nema dodatnih troškova. Nema potrebe za tehničarom koji bi održavao alat.

T (Teaching and learning = poučavanje i učenje). Za pregledavanje screencasta korisnici ne moraju imati posebna predznanja i pogodan je za različite stilove učenja. Integrira sliku i zvuk. Također, omogućava fleksibilnost učenja u vremenu i prostoru.

I (Interactivity = interaktivnost). Screencast ne daje veliku mogućnost interakcije, radi se više o demonstraciji rada s programskom podrškom kako bi korisnici lakše savladali određenu radnju. Ipak, korisnici imaju kontrolu nad učenjem odnosno mogućnost ponovnog pregledavanja modula.

O (Organisational issues = organizacijska pitanja). Izrada programa u screencast alatu zahtjeva računalo, odabrani screencast alat, mikrofon i Internet. Organizacijski je bitno dobro proučiti više screencast alata i odabrati onaj koji najbolje odgovara potrebama i financijskim mogućnostima dizajnera programa.

N (Novelty = novine). Tehnologija se vrlo uspješno koristi već godinama, u prilog tome govori i da je koristi velik broj *YouTube* korisnika. Alat je stabilan, pouzdan i jednostavan za korištenje, a radi opstanka na tržištu pojedinim alatima stalno se dodaju nove opcije kako bi ostali konkurentni.

S (Speed = brzina). Alati se mogu instalirati u nekoliko minuta, a neki su *web-based* pa nije potrebna instalacija na računalo. Dizajneru je potrebno malo vježbe da bi uskladio snimanje zvuka i slike, pa je s pripremljenim sadržajem moguće brzo pokretanje programa. Određeni screencast alati omogućuju naknadno uređivanje snimke, dodavanje naslova ili efekata što znači da je materijale moguće jednostavno mijenjati i dorađivati.

S obzirom da je bila nužna platforma za module, a znanje u izradi internetskih stranica bilo je nedostatno, moduli snimljeni pomoću screencast alata postavljeni su na blog koji je zadovoljio potrebe.

III. PROBLEMI/IZAZOVI

Tim čine dva predmetna stručnjaka koja su trebala obaviti sve poslove na izradi i implementaciji programa. Nedostajao je i nedostaje ekspert na području alata za izradu programa e-učenja koji bi za zamisli predmetnih stručnjaka pronalazio adekvatna tehnološka rješenja.

Screencast kao osnovni izabrani alat ne nudi mogućnost interakcije s korisnicima, kao niti korisnika međusobno. Taj problem riješen je uvođenjem prostora na blogu gdje korisnici mogu postavljati pitanja i davati komentare na koje će administratori programa naknadno odgovarati.

Poseban izazov predstavlja promoviranje programa i poticanje korisnika na njegovo korištenje za što bi idealna platforma bile društvene mreže.

IV. ZAKLJUČAK

U opisu programa «Naučite kako, brzo i lako, do pravnih propisa!» navedeni su nastavni problem i potencijalni korisnici iz čega su proizašli ciljevi i ishodi programa e-učenja. Također, definiran je tim za izradu koji čine dva predmetna stručnjaka, potrebni resursi te mogući problemi na koje je trebalo obratiti pažnju pri izradi i implementaciji cjelokupnog programa.

Još jedanput utvrđujući cilj (prezentacija, prednosti i načini korištenja tražilice pravnih propisa Cadial) tim za izradu programa ima dobro razrađenu kontrolnu točku [7] na koju se uvijek može vraćati da bi potvrdio jesu li izrađeni materijali, sadržaj i izvedba primjereni i u skladu s ciljevima. U istu svrhu služi i stalno osvrtanje na zadane ishode učenja. Tim se neprestano morao i mora podsjećati da sva promišljanja i aktivnosti trebaju biti usmjerene prema korisniku i na to da korisnik, kada prođe edukacijski program, može suvereno koristiti tražilicu i lako pronalaziti pravne propise.

Budući da je odabiru sadržaja najbolje pristupati iz perspektive korisnika [8], kod odabira i organizacije nastavnih materijala bilo je potrebno držati se jednostavnih i jasnih formulacija i univerzalnog dizajna.

Modelom SECTIONS procijenjeno je da bi screencast alat bio najbolji odabir s obzirom na potencijalne korisnike i male tehničke zahtjeve za korisnika. Također, alat nije financijski zahtjevan. Nije interaktivan, ali komunikacija je riješena otvaranjem prostora za pitanja i komentare na blogu. Iako alat nije previše zahtjevan, dobro bi došla pomoć i uključivanje barem jednoga informatičara u tim koji bi bio zadužen za nova tehnološka rješenja.

Očekuje se da će program, implementiran i dostupan javnosti, biti u funkciji povećanja kvalitete dostupnosti službenih informacija u dijelu pravnih propisa. Povratne informacije korisnika pridonijet će evaluaciji programa te će se on, ukoliko se uoče nedostaci, po potrebi revidirati. Ako reakcije korisnika budu pozitivne i ako se procijeni da je program koristan i ispunjava svoju zadaću, isti tim radit će dalje na novim programima e-učenja i približavanju korisnicima i drugih Hidrinih usluga i baza podataka s pripadajućim tražilicama.

LITERATURA

[1] www.hidra.hr, pristupljeno 31.01.2013.

[2] Zakon o pravu na pristup informacijama, NN 172/2003, NN 37/2011, NN 77/2011

[3] Hrastinski, S. (2008). Asynchronous and Synchronous E-Learning *EDUCAUSE Quarterly*, vol. 31, no. 4 (October–December 2008). Preuzeto 31.01.2012. s http://www.educause.edu/ero/article/asynchronous-and-synchronous-e-learning

[4] Istrate, O. (2009). Visual and pedagogical design of eLearning content. eLearning Papers. Preuzeto 31.01.2013. s http://www.elearningeuropa.info/files/media/media21215.pdf

[5] Shepherd, C. (2010). Strategies for learning and performance support 1: Exposition. Onlignment. Preuzeto 31.01.2012. s http://onlignment.com/2010/07/strategies-for-learning-and-performance-support-1-exposition

[6] Bates, A.W. i Poole, G. (2003). Effective Teaching with Technology in Higher Education: Foundations for Success. San Francisco: Jossey-Bass.

[7] Carnet, ELA Course Design 2011. 2. cjelina: Evaluacija projekata e-učenja.

[8] Carnet, ELA Course Design 2011. 1. cjelina: Modeli i proces instrukcijskog dizajna

Primjena računalnih igara u obrazovanju

Marina Mirković, dipl. inženjer elektrotehnike
Tehnička škola, Požega, Republika Hrvatska
marina.mirkovic@po.t-com.hr

Sažetak

Ono što će motivirati i aktivirati današnje generacije učenika su edukativne računalne igre koje bi mogle postati most između sadašnjeg načina učenja i poučavanja i budućih pristupa učenju.

Edukativne igre koriste se za poučavanje, vježbanje, učenje, mjerenje znanja te razvijanje vještina. Njihovom uporabom razvija se u djeteta interaktivan rad (pojačava znatiželju, razvija međusobnu pomoć, kontrolu ispravnosti i pohvale za točnost) te navika individualnog ali i grupnog rada.

Veliki broj internetskih portala omogućuje pristup računalnim igrama prema različitim nastavnim područjima, čime je učenicima pružena mogućnost razvijanja vještina i znanja iz većine nastavnih predmeta. Kvalitetne računalne igre moraju biti koncipirane tako da simuliraju školske situacije (vježbanje, ispitivanje, rješavanje zadataka). One kod učenika moraju poticati aktivnost, razmišljanje, pamćenje i brzinu.

I. UVOD

Brojna istraživanja igrača video i računalnih igara kod djece otkrila su da su oni koji ne igraju izuzetna manjina. Pritom gotovo polovica igra svaki dan, a većina igra više različitih tipova igara. Dio roditelja, uglavnom mladi, igra igre sa svojom djecom. Igre doprinose razvoju komunikacijskih vještina te većoj zainteresiranosti za društvena događanja.

Škotske su škole prihvatile korištenje računalnih igara u obrazovne svrhe. Utemeljen je centar za učenje kroz igru pod nazivom The Consolarium: The Scotish Centre for Games based learning. Njegova glavna svrha je istraživanje različitih tehnologija koje podržavaju računalne igre, osiguranje mjesta na kojem nastavnici i druge osobe uključene u obrazovanje imaju pristup resursima kao i poticanje nastavnika na raspravu o takvoj vrsti tehnologije u njihovom razredu, školi i lokalnoj zajednici [1].

Računalo u obrazovanju zadržava djetetovu koncentraciju učenja, omogućuje interaktivno učenje, pruža mogućnost otkrivanja i potiče aktivno učenje. Unaprijeđen je proces učenja i poučavanja uz uporabu kognitivnih alata za izgradnju znanja, pri čemu učenici postaju kreatori, kroz vlastito interpretiranje i organiziranje znanja.

II. PRIMJENA RAČUNALNIH IGARA U NASTAVI

Učenici današnjih generacija kao digitalni urođenici navikli su na veliku brzinu, istodobno izvođenje više zadataka, slučajan pristup, aktivnost, povezanost, zabavu i maštovitost. Učenje putem digitalnih igara jedan je od dobrih načina da se dopre do digitalnih urođenika korištenjem njihova »urođenog jezika«. Igre su medij s kojim su dobro upoznati i u kojem stvarno uživaju [2].

Računalne igre raspolažu s bogato sofisticiranim scenarijem odvijanja i od učenika zahtijevaju dosta promišljanja za donošenje odluka tijekom igranja. Učenje kroz računalne igre otkriva nove načine dizajna nastave pomoću informatičke i komunikacijske tehnologije i istovremeno pruža učenicima mogućnost stjecanja vještina i sposobnosti potrebnih za budućnost. Učenje kroz računalne igre može biti nadopuna predavanju u učionici.

Učenicima korištenje igara za učenje može predstavljati i učenje i zabavu, suočavanje s izazovom i priliku za postizanje boljih rezultata, isprobavanje različitih uloga, razvijanje sposobnosti eksperimentiranja i promatranja rezultata istog, sposobnost izražavanja osjećaja te razmišljanje o određenim konfliktnim situacijama.

Kroz igru nastavnici mogu dati uvod u novu temu učenja, kojom se povećava zanimanje učenika ili kao dodatnu aktivnost za mnoge druge svrhe, npr. za stvaranje složenih mogućnosti učenja, povećanje motivacije učenika, ili kao drugi način interakcije i komunikacije. Igre možemo koristiti pri ponavljanju, usvajanju različitih novih sadržaja te uvježbavanju. Mogu se koristiti za individualan rad, rad u parovima ili grupni rad [3].

Kvalitetne računalne igre moraju biti koncipirane tako da simuliraju školske situacije (vježbanje, ispitivanje, rješavanje zadataka). One kod učenika moraju poticati aktivnost, razmišljanje, pamćenje i brzinu. Međutim, igra na računalu mora biti dozirana i mora pozitivno utjecati na sposobnost mišljenja, zaključivanja, rješavanja problema, te na sposobnost snalaženja u novim situacijama. Učenje i rad na računalu korisni su za djecu, stoga ih treba usmjeravati na igre obrazovnog tipa, u kojima se na zabavan način vježba opažanje, brzina zaključivanja, čitanje i slično.

Korištenje računalnih igara u nastavi pokazalo se vrlo motivirajućim, većina učenika aktivno sudjeluje na satu, a među učenicima vlada pozitivno natjecateljsko ozračje. Iskustva pokazuju da s računalnim igrama učenici mnogo lakše i angažiranije pristupaju učenju, duže vremena aktivno uče, teže boljem rezultatu i rjeđe odustaju ako ne

uspiju od prvog pokušaja. Rezultati učenja su mjerljivi i boljim uspjehom u školi. Igre također mogu biti primamljive i roditeljima da ponove sitnice iz svojih školskih dana te nauče nešto novo, a istovremeno se druže i natječu sa svojom djecom za koju su igra i pobjeđivanje roditelja izrazito poticajna.

Kroz edukativne programe, djeca ostvaruju niz dobrobiti. Oni uče na jednostavan i njima zanimljiv način, razvijaju kognitivne vještine, inteligenciju, vještine rješavanja problema. Kroz igre stječu osjećaj kompetencije i samopouzdanja. Igre imaju pozitivan učinak i na motoriku.

Igre se mogu koristiti i za osobni razvoj te za poboljšanje samopouzdanja igrača odnosno učenika. Osobama s posebnim potrebama digitalne igre mogu pružiti priliku da iskuse svijet na način koji većina nas uzima zdravo za gotovo. Model učenja kroz igru koristi se u nekim područjima redovnog školovanja, a posebice u vojnoj, medicinskoj, poslovnoj, fizičkoj i drugim obukama. U mnogim slučajevima primjena ozbiljnih igara i simulacija za učenje predstavlja priliku da učenici primjene stečeno znanje i eksperimentiraju, dobiju povratnu informaciju u obliku rezultata te na taj način stječu iskustvo u „sigurnom virtualnom svijetu“ [4].

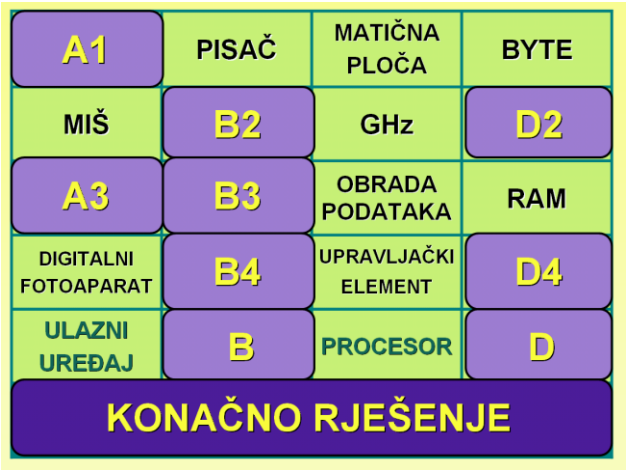
Igranje video igara nudi sjajnu priliku za razvijanje pismenosti. Budući da uspostavljanje veza između govorene i pisane riječi razvija pismenost, mehanika igrica treba nuditi i zvučne i tekstualne mogućnosti. To se može opravdati u kontekstu priče pomoću govora u igrici ili pomoću predmeta koji nude tekstualna rješenja ili potrebne informacije. Komercijalna obrazovna igra treba sadržavati i igračke i obrazovne rezultate. Igrači će znati da uče, ali glavna svrha će biti dovršiti igru. Rezultati igranja igrica su uspjeh i pohvale kolega, ali postupak igranja će učiniti mnogo više. [5].

III. PRIMJERI EDUKATIVNIH RAČUNALNIH IGARA I PORTALA

Unutar Udruge hrvatskih učitelja razredne nastave «Zvono», oformljena je radna skupina Informatički tečajevi čiji članovi sudjeluju u izradi i vođenju online tečajeva u LMS sustavu Moodle. Tečaje «Upotreba MS PowerPointa» osmišljen je s ciljem da pomogne učiteljima u radu, te ih motivira da upotrebljavaju MS PowerPoint ne samo kao sredstvo za prezentiranje nastavnih sadržaja, već da otkrivajući njegove skrivene mogućnosti izrade kvizove i igrice zanimljive učenicima (igra asocijacija, otkrivanje slike, utrka znanja).

U online tečaju «Izrada nastavnih materijala» opisuje se izrada križaljki i osmosmjerki koje učenici vole rješavati pri ponavljanju nastavnih sadržaja. Na internetu se mogu pronaći alati koji na vrlo jednostavan način omogućuju njihovu izradu.

Britanci Wayne Holmes i Ben Barton su autori programa Zondle. To je online platforma za učenje kroz računalne igre koja pomaže učiteljima, roditeljima i učenicima u stvaranju i igranju obrazovnih igri koje su prilagođene njihovim obrazovnim potrebama. Pokrenut je 2010. godine i već ima više od 50.000 registriranih korisnika iz više od 100 zemalja, koji su zajedno pripremili više od 100.000 primjeraka obrazovnih sadržaja za dijeljenje i igranje kroz više od 40 interaktivnih multimedijskih igri.

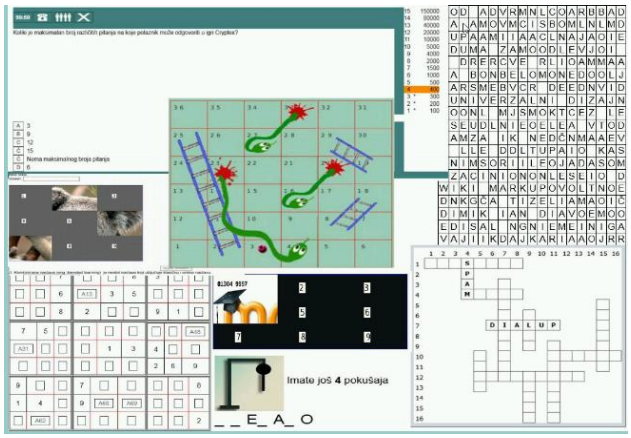


Slika 1. Igre asocijacija u Power Pointu

Portal «Sjedi 5» pokrenula je Udruga za razvoj suvremenih obrazovnih metoda u Republici Hrvatskoj koja je osnovana 2010. godine u Osijeku. Na portalu «Sjedi 5» u izbornicima su raspoređene igre iz područja hrvatskog jezika, matematike i geografije. Moguć je izbor glazbenih i logičkih igara, prezentacija iz matematike i fizike, te video instrukcije iz matematike [6].

Slika 2. Igre u nastavi hrvatskog jezika na portalu Sjedi 5

Sve je veći broj online tečajeva izrađenih u sustavu Moodle koje koriste nastavnici i učenici u osnovnim i srednjim školama. CARNetov online tečaj «Igre u Moodle-u» namijenjen je nastavnom osoblju osnovnih i srednjih škola kao i visokoškolskih ustanova članica CARNeta koji online tečajeve u Moodle-u žele učiniti zanimljivijim za svoje učenike i studente, te pomoću igara povećati njihovu motivaciju za učenje. Svrha tečaja «Igre u Moodle-u» je upoznati nastavno osoblje s mogućnostima korištenja igara u sustavu Moodle, te kako prikladno odabrati igru i podesiti njene postavke u okruženju vlastitog tečaja.



Slika 3. Igre u Moodleu

Igre su svakom nastavniku na tečaju jednostavno dostupne u padajućem izborniku aktivnosti. Modul igara trenutno nudi osam igara: Vješalo, Cryptex, Križaljka, Milijunaš, Sudoku, Zmije i ljestve, Skrivena slika, Knjiga s pitanjima. Igre su koncipirane na način da sadržaj uzimaju iz tri moguća izvora kreirana u tečaju: Baze pitanja, Rječnika ili Testa [7].

Stručnjaci koji rade na multimedijskoj produkciji u nas nimalo ne zaostaju za europskom i svjetskom prilagodbom računalne tehnike u školskoj nastavi. Iskustva pokazuju da se primjenom računala već u predškolsko doba stvaraju dobri temelji za primjenu računalne tehnike u školi. Hrvatski stručnjaci izradili su u sprezi s didaktičarima, odgojiteljima, psiholozima i nastavnicima interaktivan ciklus „Sunčica“ koji je namijenjen djeci predškolske dobi i učenicima 1. razreda.

«Sunčica» je odgojno-obrazovni program usklađen s temama i sadržajima nastave, te pomaže i motivira usvajanje, vježbanje i ponavljanje planom i programom predviđenih sadržaja. Postoji 8 inačica «Sunčice» čiji su sadržaji vezani za slova, brojeve, prirodu, promet, svemir, glazbu, Hrvatsku i Europu. Zabavni i poučni CD-ovi s igrama «Sunčice» namijenjeni su djeci od četiri do osam godina. Na njima se nalaze igre kao što su memo, slagalice, bojanke, spremalice, biralice. Tu su i igre putovanja, priče i pjesme.

Računalna igra "Školski sraz" sadržajno pokriva gradivo od 3. do 8. razreda osnovne škole. «Školski sraz» je primjeren za korištenje u nastavi, omogućuje rad u grupama do četiri učenika po jednom računalu, rad preko mreže, ostvarena je korelacija nastavnih predmeta, učenici imaju posebnu motivaciju za rad. Ukupno šest naslova pokriva gradivo od 3. do 8. razreda, sa ukupno 25.066 pitanja. Sraz predstavlja svojevrsnu simulaciju razredne situacije te je kao takav svrhovit u pripremi za kontrolno ispitivanje znanja učenika, potiče natjecateljski duh, ali i upotpunjuje znanja učenika, otkrivajući točan odgovor nakon netočnog što ga je dao učenik.

«Učilica» je multimedijski sustav poučnih i odgojnih igrica, hrvatske proizvodnje, a datira od 2002. godine. Sustav obuhvaća bitno gradivo nastavnih predmeta od 1. do 8. razreda osnovne škole, a koncipiran je tako da se njime usvaja sadržaj i informacije kroz faze ponavljanja, provjere i evaluacije znanja.

Za učenje engleskog jezika može se koristiti program «Tell me More Kids». To je zabavan, interaktivan način učenja jezika za djecu od 4 do 12 godina starosti. Aktivnosti i igre s prepoznavanjem govora pomažu djeci u procesu učenja. Djeca pritom razvijaju vještine pamćenja, logičkog zaključivanja i koncentracije. Djeca uče potpuno samostalno za računalom dok roditelji mogu pratiti njihov napredak preko upravitelja u tečaju.

Glazbene igre razvijaju dječje glazbene sposobnosti, obogaćuju njihova glazbena znanja i u djeci razvijaju potrebu za slušanjem kvalitetne glazbe. Na taj način djeca postupno kultiviraju svoj glazbeni ukus i razvijaju kriterije za vrjednovanje glazbe, i to ne samo umjetničke, nego i glazbe ostalih stilova. To je posebno važno u današnje vrijeme, u kojemu su djeca, ukoliko izostane djelovanje odgojitelja, učitelja i roditelja, prepuštena utjecaju masovnih medija i glazbe upitne kvalitete koju takvi mediji emitiraju.

Korištenjem interaktivne glazbene igre «Orašar», učitelju je omogućeno da na vrlo pristupačan način uvede djecu u svijet glazbe. Ova igra ima veliki potencijal za razvijanje glazbenih sposobnosti djeteta, a kvaliteta njezinog dizajna izuzetno je motivirajuća. Ima utjecaj na razvoj glazbenih sposobnosti djece, poput glazbenog pamćenja te osjetljivosti za boju i visinu tona. Njen nastavak je glazbena igra "ALISA i Vivaldijeva Četiri godišnja doba" prema Vivaldijevom koncertnom ciklusu.

Slika 4. Orkestar u glazbenoj igri “Orašar”

Na portalu Zdravlje.hr u Igraonici zdravlja, učenici mogu naučiti kako se svakodnevno ponašati kako bi očuvali svoje zdravlje odgovarajući na pitanja u ponuđenim kvizovima i sudjeljujući u interaktivnim igrama s temama o zdravlju. Ponuđene edukativne računalne igre su: Put znanja, Toranj znanja, Pobijedi HIV i druge spolne infekcije, Stop HIV/AIDS–znanje pobjeđuje! Ove igre doprinose promoviranju širenja znanja i komunikacije o zdravlju među mladima, posebice u području vršnjačke edukacije o spolnom zdravlju i odgovornom spolnom ponašanju [8].

Na portalu BBC Schools edukativne igre su raspoređene prema nastavnim područjima i uzrastu djece. Moguće je naći igre iz područja biologije, kemije, znanosti, umjetnosti, povijesti. [9]. Posebno je zanimljiva mogućnost upoznavanja građe ljudskog tijela pomoću igre Interactive Body [10].

Slika 5. Igra "Interactive Body" upoznaje učenike s ljudskim tijelom

Igre su korisne i u mršavljenju, posebno one koje se igraju pokretom i omogućuju vježbanje. Pažnju izaziva Kinect, kontroler za igraču konzolu koji omogućuje igranje uporabom ruku, nogu, gesti i glasa. Kinect prati pokrete lika u igri i prezentira ih na velikom ekranu.

Obrazovne igre imaju potencijal za poboljšanje učenja i motivaciju među igračima, osobito u inženjerskom obrazovanju i istraživanju. Na internetu je dostupan sve veći izbor igara u području tehnike, namijenjenih učenicima ali i studentima. Zajednica učenja na bazi igre odličan je način za unapređenje učinkovitosti nastave i poticajni element. Omogućeno je natjecanje u malom, ne zagađujući okolinu. Omogućena je diskusija učenika. Učenje kroz rad u igri stvara uspješnu radnu sredinu i dobar način za poboljšanje odnosa nastavnik-učenik. Kvalitetan je prikaz fizikalnih pojava. Učenike upoznaju s vještinama koje razvijaju inženjeri u radu s uređajima, strojevima i pogonima [11]. Video igre pomažu u pripremi inženjera strojarstva. Igre u području tehnologije održiv su i plodan način kako motivirati studente u jačanju svog znanja i razumijevanju inženjerskih sustava.

Udruga "Suradnici u učenju" je uz financijsku potporu Ministarstva uprave pripremila seriju igara Radoznalci u akciji "Sigurniji internet za djecu i mlade". Igre obrađuju teme: zaštita osobnih podataka, sigurnost na društvenim mrežama i zamke elektroničke pošte. Osmišljene su kroz niz kratkih priča koje djecu stavljaju u problemske situacije kakve susreću na internetu, društvenim mrežama i prilikom svakodnevnog rada s elektroničkom poštom. Djeca igru mogu rješavati samostalno jer je uz svako pitanje pripremljen i savjet koji im objašnjava moguće probleme i posljedice raznih postupaka i oblika ponašanja na internetu. Igra može biti upotrjebljena kao motivacijski dio nastavnog sata o sigurnijem, primjerenijem i odgovornijem korištenju interneta, ali i kao podloga za razgovor roditelja s djecom o važnosti čuvanja i zaštite osobnih podataka kako u stvarnom tako i u virtualnom okruženju [12].

Slika 6. Igra "Zaštita osobnih podataka" iz serije igara Radoznalci u akciji „Sigurniji internet za djecu i mlade"

Tim Centra za nestalu i zlostavljanu djecu pripremio je igru "Robi i Neta u svijetu Interneta" u kojoj na zabavan i edukativan način djeca mogu naučiti mnogo o sigurnom korištenju interneta [13].

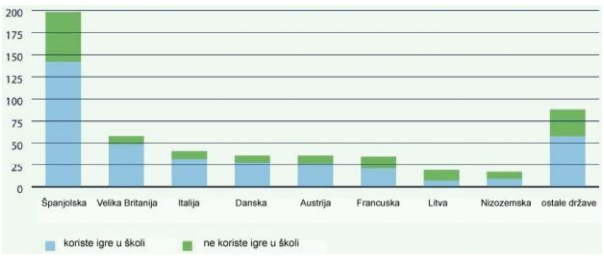
Igra «Through Wild Web Woods» kreirana je od strane Vijeća Europe kao dio programa «Gradimo Europu za djecu i s djecom». Ona potiče djecu na razmišljanje o njihovim pravima i kako ih primijeniti u radu na internetu. Putem ove igre djeca uče kako se zaštiti od različitih oblika nasilja.

Pravi izazov za učenike i nastavnike je izrada vlastite igre. Potrebna je dobra i kreativna ideja, malo vremena i alat za izradu igara. Jedan od njih je online alat Sploder koji vam omogućuje da kreirate svoje vlastite igre i da ih objavite na webu. Igre objavljene na webu dostupne su svima na pregledavanje, igranje, glasanje za najbolju igru i komentiranje samih igara.

Još je bolja mogućnost programiranja igre što omogućuje Pygame. To je modul koji olakšava izradu 2D igara u sve popularnijem Python programskom jeziku. Vizualni programski jezik Kodu osmišljen je za programiranje računalnih igara pri čemu se zapravo igrate stvarajući igre. Namijenjen je djeci i svima kojima je vizualno programiranje igara privlačno područje.

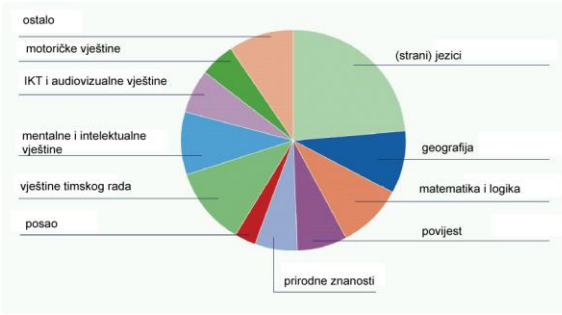
IV. ISTRAŽIVANJA O PRIMJENI IGARA U OBRAZOVANJU

Istraživanje o uporabi elektroničkih igara od strane nastavnika u njihovim nastavnim procesima provedeno je putem online upitnika na devet različitih jezika u razdoblju od listopada 2008. godine do veljače 2009. godine. Rezultati istraživanja objavljeni su putem biltena i mrežne stranice European Schoolnet, te ministarstava obrazovanja i agencija nadležnih za obrazovanje na nacionalnoj i regionalnoj razini [14].



Slika 7. Rezultati europske ankete o uporabi igara u nastavi

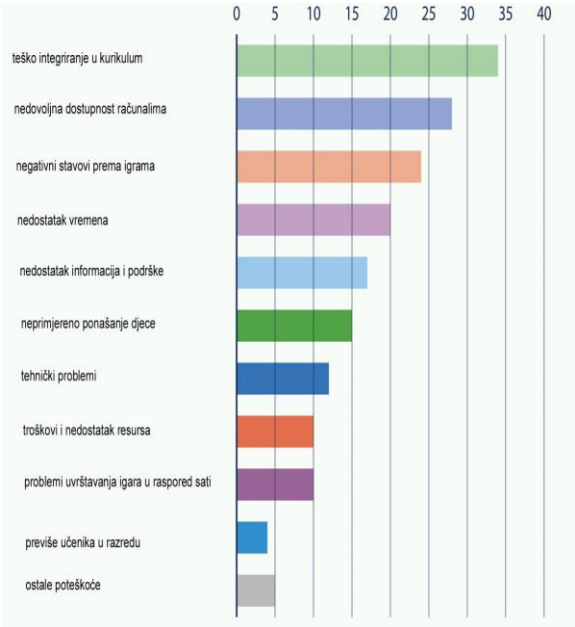
U anketi su sudjelovali nastavnici 27 različitih europskih zemalja. Od 528 ispitanika, 373 (ili 70,6%), potvrdilo je uporabu igara u nastavi.



Slika 8. Raspodjela primjene igara u nastavi prema predmetima i vještinama

Na temelju odgovora na pitanja u upitniku izveden je zaključak u kojim nastavnim predmetima se igre najviše koriste (slika 8.). Vidljiva je pobjeda primjene jezičnih lekcija i igara - kako u materinjem tako i stranim jezicima. Procjenjuje se da je oko četvrtinu igara koje se koriste u nastavi namijenjeno poboljšanju učenja jezika. Ostala predmetna područja su zemljopis, matematika i povijest. Igre se također koriste za razvoj određenih vještina (timski rad, mentalne vještine, komunikacijske vještine i motoričke sposobnosti).

Posebna pozornost posvećena je identifikaciji raznih prepreka s kojima se nastavnici susreću prilikom korištenja igara u učionicama. Odgovori su grupirani u 10 kategorija, a raspodjela broja odgovora je vidljiva na slici 9.



Slika 9. Raspodjela poteškoća u primjeni igara u nastavi

Kroz analizu odgovora zašto i u koju svrhu učenja nastavnici koriste igre, navedeno je da su neke igre dobri obrazovni alati za olakšavanje učenja pojedinih predmeta i razvoja određenih vještina. Često spomenuti argumenti su: motivacija učenika, povećanje njihova interesa za učenje, bolja okolina učenja (učenici su naviknuti na uporabu računala i igranje igara), učenje ugodnije za učenike, uporaba igara kao način kako privući pozornost učenika za pojedine predmete i motivirati ih da uče. Dodatni argumenti primjene igara su: testovi, ponavljanje, revizija, postizanje različite brzine u učenju, nagrada za učenike, aktivno sudjelovanje učenika u razredu.

V. ZAKLJUČAK

Digitalne igre donose niz obrazovnih prednosti. One razvijaju kognitivne i motoričke sposobnosti te pomažu unaprijediti primjenu informacijsko komunikacijske tehnologije. Mogu se koristi za učenje činjenica, načela, rješavanje složenih problema. Povećavaju kreativnost ili pružaju praktične primjere koncepata i pravila koja se teško ilustriraju u stvarnom svijetu. Posebno su korisne za izvođenje pokusa koji mogu biti opasni u stvarnom životu, kao što je uporaba opasnih kemikalija. Digitalne igre pomažu raspravljati o temama koje mogu biti tabu, kao što su zagađenje okoliša, prijetnje, seksualno zdravlje ili nasilničko ponašanje.

Igranje je učenje i jedna od glavnih prednosti digitalnih igara je njihova sposobnost da igrači uče u izazovnom okruženju, gdje mogu učiniti pogreške i učiti radeći. Digitalne igre promiću suradnju između korisnika, te su, u određenoj mjeri, slične kolaborativnom okruženju za učenje gdje sudionici dijele informacije i uče jedni od drugih. Igranje može imati smirujući učinak na sudionike koji, kao i u drugim zabavnim aktivnostima (npr. gledanje filmova), osjećaju širok raspon emocija, ali u sigurnim i kontroliranim uvjetima. Emocije koje doživljavaju tijekom igre mogu varirati kroz radost, suosjećanje, ljutnju frustraciju ili trijumf [15].

Krajnje je vrijeme da prilagodimo obrazovni sustav i odgojne pristupe djeci novog doba koja su navikla na veliku brzinu, istodobno izvođenje više zadataka, slučajan pristup, aktivnosti, povezanost, zabavu i maštovitost.

Igre pozitivno utječu na sposobnost mišljenja, zaključivanja, rješavanja problema te kreativnost i inteligenciju, odnosno sposobnost snalaženja u novim situacijama. Igre omogućuju istraživanje učenika i bude radoznalost u igračima koji se nalaze u različitim ulogama. Dijete se uči nositi sa stresom te posebno vlastitim neuspjehom i tako postaje emocionalno inteligentnije i kompetentnije.

LITERATURA

[1] Maravić, J. Škotski primjer korištenja računalnih igara u nastavi, Skole.hr - portal za škole, 2008. Preuzeto 4. 05. 2012. s http://www.skole.hr/nastavnici/ucionica?news_id=920

[2] Prensky, M. Digitalni urođenici, digitalne pridošlice: Razmišljaju li doista drugačije? Edupoint časopis, 2005. Preuzeto 5.05.2012. s http://www.carnet.hr/casopis/32/clanci/2

[3] Maravić, J. Igra u nastavi, Skole.hr - portal za škole, 2007. Preuzeto 4.05.2012. s http://www.skole.hr/nastavnici/strucni-suradnici?news_id=405

[4] Pivec, M.. Igra i učenje: Potencijali učenja kroz igru, Edupoint časopis, 2006. Preuzeto 5. 5.2012. s http://edupoint.carnet.hr/casopis/49/clanci/1

[5] Royle, K. Učenje kroz igrice – druga perspektiva, Edupoint časopis, 2008. Preuzeto 4.05.2012 s http://edupoint.carnet.hr/casopis/64/novosti/2.html

[6] Portal Udruge za razvoj suvremenih obrazovnih metoda u Republici Hrvatskoj «Sjedi 5», http://sjedi5.hr/

[7] Online tečaj za samostalno učenje: Igre u Moodle-u, CARNet, 2012., https://moodle.carnet.hr/mod/book/view.php?id=72496&chapterid=16843

[8] Nemeth Blažić, T. Put znanja – edukativna računalna igra, 2012. Preuzeto 6. 05. 2012. s http://zdravlje.hzjz.hr/clanak.php?id=13507

[9] BBC – Schools: Games, http://www.bbc.co.uk/schools/games/

[10] BBC - Science: Human Body & Mind, http://www.bbc.co.uk/science/humanbody/body/index_interactivebody.shtml

[11] Try Engineering: Play Games, http://www.tryengineering.org/play.php

[12] Suradnici u učenju - Sigurniji internet za djecu i mlade, http://ucitelji.hr/Sigurnijiinternet/Sigurnijiinternet2012.aspx

[13] Centar za nestalu i zlostavljanu djecu, zlhttp://www.cnzd.org/site2/

[14] How are digital games used in schools?, rezultati istraživanja European Schoolnet 2009.

[15] Digital games in schools – Handbook for teachers, European Schoolnet 2010.

Istraživanje o učestalosti korištenja računala i Interneta kod učenika Osnovne škole Centar u Puli

Branka Kontošić, dipl.učitelj
Osnovna škola Centar
Danteov trg 2, Pula, Hrvatska
E-mail: b_kontosic@net.hr

Mr.sc. Alen Kontošić, mag.oec.
Sveučilište Jurja dobrile u Puli
Zagrebačka 30, Pula, Hrvatska
E-mail: akontosi@net.hr

Sažetak - Koliko vremena učenici provedu za računalom? Koliko vremena provedu koristeći usluge Interneta? Koriste li računalo za zabavu ili stjecanje znanja? Na provedenom istraživanju među učenicima Osnovne škole Centar u Puli utvrdit će se tko više koristi računalo i usluge Interneta, učenici razredne ili učenici predmetne nastave, koriste li ga više učenici ili učenice, odnosno utječu li dob i spol na učestalost korištenja računala i u koje svrhe.

Do podataka se došlo anketnim upitnikom kojim su anonimno bili ispitani učenici nižih i viših razreda. Istraživanjem se, također, htjelo uvidjeti postoji li poveznica između općeg uspjeha učenika i učestalosti korištenja računala i Interneta. Analizom prikupljenih podataka došlo se do rezultata koji pokazuju da učenici uglavnom računalo koriste u svrhu zabave i virtualnog druženja, učenici viših razreda služe se uslugama Interneta za rješavanje školskih zadataka više od učenika razredne nastave koji računalo najčešće upotrebljavaju za zabavu kroz igrice i videozapise.

Ključne riječi: korištenje računala, korištenje Interneta, učenici razredne nastave, učenici predmetne nastave.

I. UVODNA RAZMATRANJA

A. Uvod

Računalo i Internet danas su sastavni dio svakodnevice i prisutni su u svim sferama života. Korištenje računala i usluga Interneta ne postavlja nikakve granice, ni dobne, ni spolne kao ni granice u stupnju obrazovanja. Svrha korištenja računala individualne je prirode, dakle zabava, opuštanje, informiranje, posao. Danas se čovjek susreće s računalom već pri samom rođenju i ono postaje sastavni dio odrastanja djeteta.

Budući da djeca odrastaju i stječu prve osnove obrazovanja kroz rad na računalu koristeći edukativne sadržaje i sadržaje zabavnog karaktera prilagođene njihovoj dobi ovim će se istraživanjem uvidjeti koliko često učenici koriste računalo, koliko vremena dnevno provedu ispred računala i u koje svrhe. Brzo usvajaju potrebna znanja i vještine za rad na računalu te su nerijetko mentori svojim roditeljima, a osobito bakama i djedovima.

Informacijsko-komunikacijska tehnologija (ICT) danas je dio života svih generacija dok ih mlađi naraštaji prihvaćaju u svim domenama na putu svoga odrastanja.

B. Svrha i ciljevi istraživanja

Cilj rada *Istraživanje o učestalosti korištenja računala i Interneta kod učenika Osnovne škole Centar u Puli* jest istražiti koliko vremena učenici provedu za računalom, koliko vremena provedu koristeći usluge Interneta te koriste li računalo za zabavu ili stjecanje znanja. Između ostalog, želi se istražiti tko među učenicima Osnovne škole Centar u Puli više koristi računalo i usluge Interneta, učenici razredne ili učenici predmetne nastave, koriste li ga više učenici ili učenice, odnosno utječu li dob i spol na učestalost korištenja računala i u koje svrhe. Istraživanjem se, također, htjelo uvidjeti postoji li poveznica između općeg uspjeha učenika i učestalosti korištenja računala i Interneta.

II. METODOLOGIJA PROVEDENOG ISTRAŽIVANJA

Istraživanje je izvršeno anketnim upitnikom kojim su bili anonimno ispitani učenici nižih i viših razreda. U istraživanju se primijenio postupak anketiranja s mjernim instrumentom anketnog lista s ukupno 16 pitanja (od čega su na 14 pitanja unaprijed ponuđeni DA-NE odgovori, dok su na dva pitanja učenici zaokruživali unaprijed ponuđene odgovore: kakav uspjeh postižu u školi (odličan, vrlo dobar, dobar, dovoljan, nedovoljan) te za koje predmete najčešće koriste računalo i Internet (hrvatski jezik, strani jezik; matematika, fizika; biologija, kemija, priroda; povijest, geografija; likovna kultura, glazbena kultura, tjelesna i zdravstvena kultura). U istraživanju su se osim navedene metode anketiranja primijenile i druge znanstvene metode poput statističke metode, metode deskripcije, analize i sinteze. Podaci su obrađeni primjenom programa za tablično računanje MS Excel, a primijenjene su statističke metode mjere centralne tendencije (frekvencija i postotak).

III. REZULTATI I ANALIZA DOBIVENIH REZULTATA

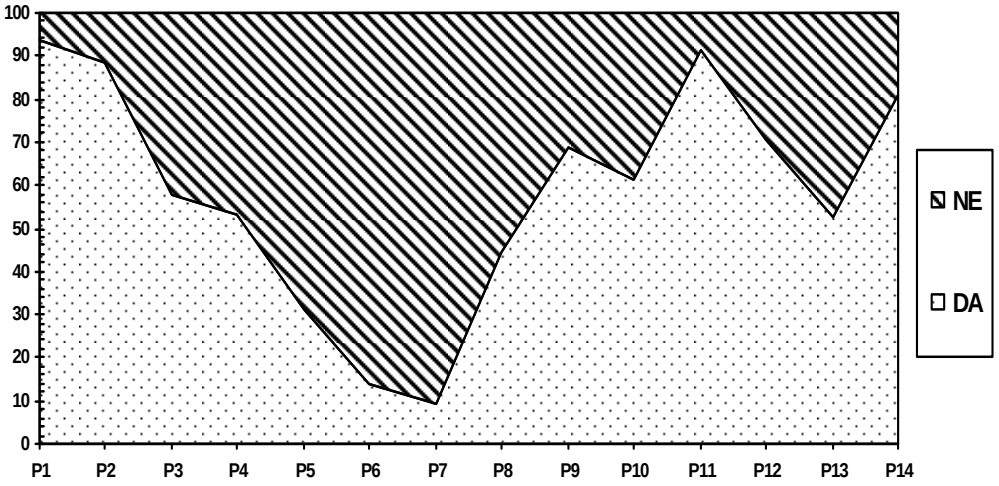
Istraživanje je provedeno u Osnovnoj školi Centar u Puli na uzorku od 221 učenika od ukupno 282 učenika (od kojih 43,98% dječaka i 56,02% djevojčica) te je u analizi rezultata vrijedno spomenuti velik postotak ($p = 0,784$) anketiranih učenika škole: 91 učenika nižih razreda (72,8% od 125 učenika nižih razreda škole od kojih je 42,4% dječaka i 57,6% djevojčica) te 130 učenika viših razreda (82,8% od 157 učenika viših razreda škole od kojih je 45,2% dječaka i 54,8% djevojčica). Obzirom da je ispunjene upitnike vratilo toliko velik broj učenika, stopa povrata može se ocijeniti vrlo zadovoljavajućom, uzimajući u obzir da se u sličnim istraživanjima ona kreće u puno manjem postotku. Stoga, standardna pogreška proporcije je mala obzirom da je n velik ($n > 30$). Veličina aritmetičke sredine približno je normalno distribuirana te stoga, prema pravilu *dvije sigme*, postiže s vjerojatnošću većom od 0,95 sve vrijednosti u intervalu ± *dvije sigme oko sredine*. Obzirom da je gotovo četiri od pet učenika ispunilo anketni upitnik i da su prikupljeni upitnici u velikom postotku zastupani u razrednoj i u predmetnoj nastavi uzorak se može ocijeniti reprezentativnim pa je na temelju njega moguće donositi odgovarajuće zaključke koji se odnose na populaciju.

Analizom prikupljenih podataka došlo se do rezultata koji pokazuju da 93,7% učenika od ukupnog broja ispitanih učenika kod kuće posjeduje računalo, od kojih 92,2% čine dječaci te 95% djevojčice u odnosu na ukupan broj zasebno ispitanih dječaka i djevojčica. Pristup Internetu ima 88,7% učenika od ukupnog broja ispitanih učenika, od kojih 87,3% čine dječaci te 89,9% djevojčice u odnosu na ukupan broj zasebno ispitanih dječaka i djevojčica. Rezultati istraživanja ukazuju da gotovo 58% učenika svakodnevno koristi računalo, od toga 65,7% učenika te nešto manje učenice, 51,3%. Koliko vremena dnevno provedu ispred računala te koristeći usluge Interneta, rezultati pokazuju da više od polovice ispitanih učenika, njih 53,4%, dnevno provedu do 2 sata na računalu i to 52,94% djevojčica i 53,9% dječaka. Od 2 do 4 sata dnevno računalo koristi oko 31,2%, i to učenici u nešto većem postotku (33,3%) od djevojčica. Manji broj učenika (14%) za računalom provede dnevno od 4 do 6 sati i to uglavnom učenici (22,5%) te u mnogo manjem broju učenice (6,7%). Veći dio svog slobodnog vremena, više od 6 sati, svakodnevno provodi 9,5% učenika za računalom, 13,7% učenici i 5,9% učenice. Uslugu elektroničke pošte koristi manje od polovice ispitanih učenika (44,3%), dok *Facebook* koristi više od dvije trećine ispitanih učenika (68,8%) i to učenici 72,5% i učenice 65,5%. U svrhu usvajanja novih sadržaja Internet koristi 61,1% učenika, dok za zabavu i igranje igrica 91,1% ispitanih učenika, u nešto većem broju učenice u odnosu na dječake. Na računalu, putem Interneta, igrane i animirane filmove gleda 70,8% ispitanih učenika, od kojih učenice u nešto većem broju (72,2%), dok za rješavanje školskih zadataka računalo i Internet koristi nešto više od polovine ispitanih učenika (52,4%) od toga 58,3% učenica. Nastavu informatike polazi 81,1% ispitanih učenika. Podaci anketnog istraživanja od prvog do četrnaestog pitanja prikazani su u tabličnom prikazu 1. te u grafičkim prikazima na slikama 1., 2. i 3. Uspoređujući razrednu i predmetnu nastavu u posjedovanju računala kod kuće, nema razlika, oko 93% učenika ima računalo dok pristup Internetu ima veći broj učenika predmetne nastave (90,8%) u odnosu na razrednu nastavu (85,7%). Računalo svakodnevno koristi 66,9% učenika predmetne nastave te 45,1% razredne nastave, učenici u mnogo većem broju koriste računalo od učenica, osobito u razrednoj nastavi. Do 2 sata dnevno računalo koristi 57,1% učenika razredne te 50,8% učenika predmetne nastave, učenice u nešto manjem broju od učenika. Od 2 do 4 sata dnevno računalom se koristi 17,6% ispitanih učenika razredne nastave te u većem broju učenici predmetne nastave (40,8%). Manji broj učenika, njih 8,8%, ispitanih u razrednoj nastavi, koristi računalo od 4 do 6 sati, te njih 17,7% u predmetnoj nastavi, uglavnom su to dječaci (17,5% u razrednoj i 25,8% u predmetnoj nastavi). Na pitanje koriste li računalo više od 6 sati

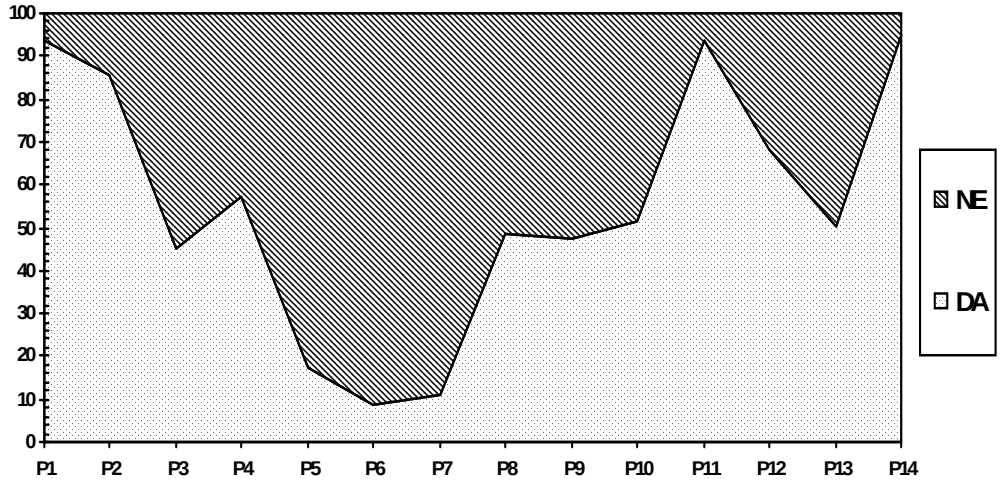
dnevno, u razrednoj nastavi potvrdno je odgovorilo 11% ispitanih učenika (15% učenici, 7,8% učenice), a u predmetnoj nastavi 8,5% (12,9% učenici, 4,4% učenice).

	Pitanje	Razredna nastava (RN):				Predmetna nastava (PN):				Ukupno (RN+PN):			
		DA [%]		NE [%]		DA [%]		NE [%]		DA [%]		NE [%]	
		M [%]	Ž [%]	M [%]	Ž [%]	M [%]	Ž [%]	M [%]	Ž [%]	M [%]	Ž [%]	M [%]	Ž [%]
P1	Imaš li kod kuće računalo?	93,4		6,6		93,8		6,2		93,7		6,3	
		92,5	94,1	7,5	5,9	91,9	95,6	8,1	4,4	92,2	95,0	7,8	5,0
P2	Imaš li pristup Internetu?	85,7		14,3		90,8		9,2		88,7		11,3	
		85,0	86,3	15,0	13,7	88,7	92,6	11,3	7,4	87,3	89,9	12,7	10,1
P3	Koristiš li računalo svakodnevno?	45,1		54,9		66,9		33,1		57,9		42,1	
		55,0	37,3	45,0	62,7	72,6	61,8	27,4	38,2	65,7	51,3	34,3	48,7
P4	Koristiš li računalo do 2 sata	57,1		42,9		50,8		49,2		53,4		46,6	
	dnevno?	57,5	56,9	42,5	43,1	51,6	50,0	48,4	50,0	53,9	52,9	46,1	47,1
P5	Koristiš li računalo od 2 do	17,6		82,4		40,8		59,2		31,2		68,8	
	4 sata dnevno?	22,5	13,7	77,5	86,3	40,3	41,2	59,7	58,8	33,3	29,4	66,7	70,6
P6	Koristiš li računalo od 4 do	8,8		91,2		17,7		82,3		14,0		86,0	
	6 sati dnevno?	17,5	2,0	82,5	98,0	25,8	10,3	74,2	89,7	22,5	6,7	77,5	93,3
P7	Koristiš li računalo više od	11,0		89,0		8,5		91,5		9,5		90,5	
	6 sati dnevno?	15,0	7,8	85,0	92,2	12,9	4,4	87,1	95,6	13,7	5,9	86,3	94,1
P8	Koristiš li se elektroničkom	48,4		51,6		41,5		58,5		44,3		55,7	
	poštom (e-mailom)?	42,5	52,9	57,5	47,1	45,2	38,2	54,8	61,8	44,1	44,5	55,9	55,5
P9	Koristiš li *Facebook*?	47,3		52,7		83,8		16,2		68,8		31,2	
		55,0	41,2	45,0	58,8	83,9	83,8	16,1	16,2	72,5	65,5	27,5	34,5
P10	Pretražuješ li Internet u svrhu	51,6		48,4		67,7		32,3		61,1		38,9	
	usvajanja novih znanja?	50,0	52,9	50,0	47,1	62,9	72,1	37,1	27,9	57,8	63,9	42,2	36,1
P11	Koristiš li računalo i Internet	93,4		6,6		89,3		10,7		91,1		8,9	
	za zabavu i igranje igrica?	92,5	94,1	7,5	5,9	87,9	90,6	12,1	9,4	89,8	92,2	10,2	7,8
P12	Koristiš li računalo i Internet za	68,1		31,9		72,7		27,3		70,8		29,2	
	gledanje igranih i animiranih filmova?	60,0	74,5	40,0	25,5	75,4	70,3	24,6	29,7	69,1	72,2	30,9	27,8
P13	Koristiš li računalo i Internet za učenje	50,5		49,5		53,7		46,3		52,4		47,6	
	i rješavanje školskih zadataka?	50,0	51,0	50,0	49,0	42,1	64,1	57,9	35,9	45,4	58,3	54,6	41,7
P14	Polaziš li nastavu informatike?	94,5		5,5		71,1		28,9		81,1		18,9	
		95,0	94,1	5,0	5,9	70,2	71,9	29,8	28,1	80,4	81,7	19,6	18,3

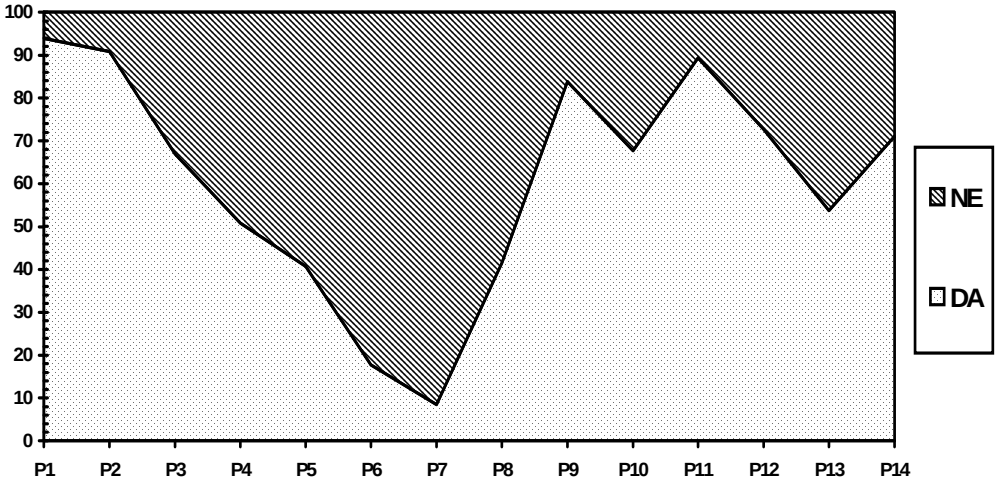
Tablica 1. Tablični prikaz rezultata na pitanja od 1. - 14.



Slika 1. Ukupni pozitivni i negativni (u %) odgovori učenika na postavljena pitanja



Slika 2. Pozitivni i negativni (u %) odgovori učenika razredne nastave (1. - 4. razreda) na postavljena pitanja



Slika 3. Pozitivni i negativni (u %) odgovori učenika predmetne nastave (5. - 8. razreda) na postavljena pitanja

Elektroničkom poštom koriste se, u većem broju, ispitani učenici razredne nastave (48,4%) u odnosu na ispitane učenike predmetne nastave (41,5%), u razrednoj nastavi učenice više od učenika, njih 52,9%, dok u predmetnoj nastavi učenici više od učenica, njih 45,2%. U mnogo većem broju *Facebook* koriste učenici predmetne nastave, 83,8% ispitanih učenika te 47,3% učenika u razrednoj nastavi te u gotovo podjednakom omjeru učenici i učenice u predmetnoj nastavi, dok u razrednoj nastavi nešto više dječaci (55%) od djevojčica (41,2%). U svrhu usvajanja novih znanja Internet koristi 51,6% ispitanih učenika razredne nastave, te 67,7% predmetne nastave, dok za zabavu i igrice računalo i Internet nešto više koriste učenici nižih razreda (93,4%) u odnosu na učenike u višim razredima (89,3%) i to učenice više od učenika. Igrane i animirane filmove na računalu i Internetu više gledaju ispitani učenici viših razreda (72,7%) u odnosu na učenike nižih razreda (68,1%), učenice u nešto većem postotku u odnosu na dječake u nižim razredima dok je u predmetnoj nastavi obrnuto. Oko polovine ispitanih učenika, njih 50,5% u razrednoj te 53,7% u predmetnoj nastavi, koristi Internet za učenje i rješavanje školskih zadataka, i to učenice viših razreda u većem broju (64,1%) u odnosu na dječake (42,1%), dok je u nižim razredima omjer gotovo podjednak. Nastavu informatike polazi 94,5% ispitanih učenika u razrednoj te 71,1% u predmetnoj nastavi, u gotovo podjednakim omjerima.

Osim 14 pitanja s unaprijed ponuđenim DA-NE odgovorima u istraživanju se na dva pitanja tražilo da učenici zaokruže unaprijed ponuđene odgovore na pitanja. U odgovoru na pitanje broj 15. „*Kakav uspjeh postižeš u školi?*“ ponuđeni odgovori bili su: odličan, vrlo dobar, dobar, dovoljan, nedovoljan. Redoslijed odgovora određen je prema intenzitetu frekvencije prikazanih u tabličnom prikazu br. 2.

	Pitanje	Redoslijed odgovora prema intenzitetu frekvencije ponuđenih odgovora:	Razredna nastava (RN): Ukupni %		Predmetna nastava (PN): Ukupni %		Ukupno (RN+PN): Ukupni %	
			M [%]	Ž [%]	M [%]	Ž [%]	M [%]	Ž [%]
P15	Kakav uspjeh postižeš u školi?	1. Odličan [%]	71,43		38,02		49,46	
			73,33	69,70	29,82	45,31	44,83	53,61
		2. Vrlo dobar [%]	15,87		39,67		31,52	
			20,00	12,12	29,82	48,44	26,44	36,08
		3. Dobar [%]	7,94		16,53		13,59	
			3,33	12,12	29,82	4,69	20,69	7,22
		4. Dovoljan [%]	3,17		4,13		3,80	
			0,00	6,06	8,77	0,00	5,75	2,06
		5. Nedovoljan [%]	1,59		1,65		1,63	
			3,33	0,00	1,75	1,56	2,30	1,03

Tablica 2. Tablični prikaz rezultata na 15. pitanje u razrednoj nastavi, predmetnoj nastavi i njihovom zbroju

Prema intenzitetu frekvencije prikazanih u tabličnom prikazu br.2. od ukupnog broja ispitanika 49,46% učenika postiže odličan uspjeh, osobito u razrednoj nastavi (71,43%) gdje je odnos učenika i učenica gotovo podjednak, dok je u predmetnoj nastavi veći broj učenica s odličnim uspjehom. U predmetnoj nastavi najviše je zastupljeno učenica s vrlo dobrim uspjehom. U odgovoru na pitanje broj 16. „*Za koje predmete najčešće koristiš računalo i Internet?*“ ponuđeni odgovori bili su: hrvatski jezik, strani jezik; matematika, fizika; biologija, kemija, priroda; povijest, geografija; likovna kultura, glazbena kultura, tjelesna i zdravstvena kultura. Redoslijed odgovora određen je prema intenzitetu frekvencije prikazanih u tabličnom prikazu br. 3.

	Pitanje	Redoslijed odgovora prema intenzitetu frekvencije ponuđenih odgovora:	Razredna nastava (RN): Ukupni %		Predmetna nastava (PN): Ukupni %		Ukupno (RN+PN): Ukupni %	
			M [%]	Ž [%]	M [%]	Ž [%]	M [%]	Ž [%]
P16	Za koje predmete najčešće koristiš računalo i Internet?	1. hrvatski jezik, strani jezik	50,85		31,43		36,32	
			50,00	51,61	38,89	26,21	42,00	32,09
		2. matematika, fizika	18,64		8,57		11,11	
			17,86	19,35	9,72	7,77	12,00	10,45
		3. biologija, kemija, priroda	18,64		16,00		16,67	
			14,29	22,58	15,28	16,50	15,00	17,91
		4. povijest, geografija	3,39		38,29		29,49	
			7,14	0,00	31,94	42,72	25,00	32,84
		5. likovna i glazb. kultura, tjelesna i zdrav. kultura	8,47		5,71		6,41	
			10,71	6,45	4,17	6,80	6,00	6,72

Tablica 3. Tablični prikaz rezultata na 16. pitanje u razrednoj nastavi, predmetnoj nastavi i njihovom zbroju

Prema intenzitetu frekvencije prikazanih u tabličnom prikazu br.3. od ukupnog broja ispitanika najviše ih koristi računalo i Internet za hrvatski jezik i strani jezik s naglaskom na strani jezik s većom zastupljenošću dječaka, posebice u predmetnoj nastavi. Kod korištenja računala i Interneta u predmetnoj nastavi najviše se koriste za predmete povijest i geografija. Ženska populacija više koristi računalo i Internet za predmete povijest i geografiju.

IV. ZAKLJUČAK

Obradom podataka dobiveni su rezultati koji su dali odgovore na postavljena pitanja o učestalosti korištenja računala i Interneta. Stavljajući u odnos varijable vremena provedenog na radu za računalom i broja učenika rezultati su obrnuto proporcionalni. Ispitana populacija najviše koristi računalo i Internet za zabavu te gledanje filmova i videozapisa i to djevojčice u većem omjeru od dječaka u nižim razredima dok je u višim razredima obrnuto. Veća odstupanja ustanovljena su kod korištenja elektroničke pošte, u razrednoj nastavi više koriste djevojčice, a u predmetnoj dječaci. Za učenje i rješavanje školskih zadataka računalo i Internet koriste u znatno većem postotku djevojčice u predmetnoj nastavi.

Usporedimo li učenike po dobi i spolu, kao nezavisne varijable, o učestalosti korištenja računala i Interneta te u koje svrhe, dolazi se do rezultata koji pokazuju da učenici nižih razreda dnevno provode manje vremena za računalom. S povećanjem vremena rada na računalu utvrđen je veći postotak muške u odnosu na žensku populaciju.

V. ZAHVALE

Autori su zahvalni Marleni Lakić, prof. što je pročitala rad i dala im korisne povratne informacije i sugestije.

LITERATURA

[1] P. Aftab, „Dangers of the Internet guide for schools and parents“ Zagreb: Neretva 2003.
[2] N. Laniado, G. Pietra, “Naše dijete, videoigre, Internet i televizija”, [studio TIM], Rijeka, 2005.
[3] LJ. Baban, S. Jelinić, M. Lamza-Maronić, K. Ivić, “Primjena metodologije znanstvenog istraživanja”, Ekonomski fakultet Osijek, Osijek, 1993.
[4] D. Tapscott, “*Growing up digital: The rise of the Net generation*”, New York: McGraw-Hill, 1998.

Elektronski udžbenik u osnovnim i srednjim školama

Jurica Đurić, Vanja Jelkić, Anja Nikitović
Visoka škola za informacijske tehnologije Zagreb
Klaićeva 7, 10000 Zagreb, Croatia
jdjuric@vsite.hr; vjelkic@vsite.hr; anikitovic@vsite.hr

Sažetak - Učenje putem elektronskih udžbenika. Anketiranjem nastavnika iz osnovnih i srednjih škola želimo proći kroz par pitanja na koje su odgovorili nastavnici. Postavlja se par pitanja uvođenjem u školstvo elektroničkih udžbenika, olakšavanje učenja ili otežavanje. Koliko je to jednostavno za korištenje, jednostavno za nabaviti, kako pratiti nastavu. Koliko je školstvu jednostavnije predavati putem elektroničkih udžbenika? Laganim uvođenjem elektronskih udžbenika u školstvo, trebalo bi se olakšati i pojednostaviti nabavljanje udžbenika za školarce i omogućiti im jednostavniji rad. Sve što je potrebno za učenje je dobra volja i računalo koje danas posjeduje svako kućanstvo.

I. UVOD

Knjigu svi već poznajemo od naše mladosti kada su nam roditelji čitali priče. Svakim danom se sve više uvode razne nove tehnologije koje mijenjaju svaki dio naših života pa tako klasične knjige dobivaju „konkurenciju“ elektronske knjige. Elektronske knjige ili udžbenici mogu se prikazivati na bilo kojoj vrsti računala, stolnom računalu, prijenosnom računalu, tabletu ili mobilnom telefonu. Sami elektronski udžbenici se polako uvode na svih sedam kontinenata, počevši sa obrazovnim ustanovama. Pogotovo u nerazvijenim dijelovima svijeta zbog svojih beneficija su sve učestaliji, kao što su manja cijena nabave, manja cijena izrade i jednostavnije rukovanje. Diljem svijeta otvaraju se i e-knjižnice, koje u Sjedinjenim Američkim Državama ili zapadnoj Europi ostavljaju pozitivne komentare. Uvođenje elektronskih udžbenika te njihovo prihvaćanje uskoro bi se trebalo proširiti i na ostatak svijeta.

II. NAKLADNICI

U 21. stoljeću, tehnologija se uvelike razvila. To potvrđuje i proizvodnja praktičnih uređaja čija je namjena čitanje elektronskih udžbenika, kao npr. Appleov iPad ili Samsungovi tableti. Općenito, elektronska knjiga je tekst u digitalnom obliku te podsjeća na klasičnu knjigu s riječima tiskanim na papiru. Niti jedan veći izdavač u Hrvatskoj se ne bavi izdavanjem elektronskih knjiga pa tako ni elektronskih udžbenika. Ako se pogledaju brojke na portalu Društva za promicanje književnosti na novim medijima, može se zaključiti da definitivno postoji interes za elektronskim knjigama [1].

Ipak, nastavni proces primjenom elektronskih udžbenika iziskuje više od samog čitanja. Podcrtavanje teksta ili pisanje bilješki u području margina je moguće raditi na papiru pa je stoga iste te radnje potrebno omogućiti i na elektronskom udžbeniku. U Hrvatskoj su 2012. godine u 30-ak škola uvedeni elektronski imenici. Sljedeći korak bi mogli biti elektronski udžbenici. Prema članku objavljenom u devetom mjesecu prošle godine, CARNet bi sada već za mjesec dana trebao dovršiti platformu za elektronske udžbenike. Prethodno navedene radnje će biti moguće na elektronskim udžbenicima, a također će uključivati i brojne multimedijalne i interaktivne sadržaje te linkove na internet [2].

U popisu udžbenika za 2011. godinu se nalazi 68 nakladnika. Svima od 68 nakladnika poslan je e-mail s pitanjima izdaju li elektronske udžbenike. Ako ih izdaju, kada su počeli, a ako ne izdaju da li ih planiraju izdavati i kada. Nakladnici još uvijek u Hrvatskoj ne izdaju elektronske udžbenike. Neki nakladnici ih za sada ne planiraju izdavati dok neki planiraju tijekom izrade novih udžbenika, za dvije do tri godine. Mali i specijalizirani nakladnici školskih udžbenika upozoravaju na određenu problematiku. Ukazuju na tržište školskih udžbenika koji prema njima nije kvalitetno regulirano. Veliki nakladnici se prema postojećim propisima favoriziraju. Smatraju da se ne omogućava tehnološki napredak te uvođenje elektronskih udžbenika [3].

Cijena elektronskih udžbenika bi trebala biti jeftinija od klasičnih, i to čak 50 %. Pri izradi udžbenika 50 % troškova otpada na grafičku pripremu, troškove tiska, skladištenje i distribuciju. Bez obzira na manju cijenu elektronskih udžbenika, prihod nakladničkim kućama i autorima ne bi bio smanjen [4]. Međutim, još je uvijek potrebno dosta toga definirati, poput formata u kojima će se izrađivati elektronski udžbenici, platforme i alata koji će se koristiti ili zaštita od kopiranja i zlouporabe. Autori i nakladnici su uvelike zahtjevniji kad je riječ o elektronskim udžbenicima i to zbog pedagoške dimenzije koja uključuje tekst i vizualne elemente kao sastavni dio udžbenika kao nastavnog sredstva [5]. Status udžbenika u učionici prije svega treba biti jasno definiran zakonom i standardima kako bi se mogli koristiti u nastavi. U svakom slučaju prilagodba će biti neophodna kako autorima, nastavnicima i učenicima, tako i samim nakladnicima.

III. PREDNOSTI I NEDOSTACI

Uvođenje elektronskih udžbenika u osnovne i srednje škole ima dosta prednosti, ali i nedostatka. Način i cijena nabavljanja tableta će zasigurno najviše zanimati širu javnost i roditelje. Ukupni iznosi za sada još nisu definirani jer će dosta čimbenika utjecati na njihovu veličinu. Cijena tableta je od jedne do četiri tisuće kuna. Nakon što se tablet nabavi, svake školske godine je samo potrebno kupiti nove digitalne udžbenike.

Naravno, postojat će i mogućnost posudbe udžbenika npr. na samo mjesec dana kada recimo učenici uče za maturu [6]. Troškovi kompleta papirnatih udžbenika ovise o razredima. Cijene se kreću od oko petsto do oko dvije tisuće kuna. Tako se npr. cijena udžbenika za prvi razred kreće od 500 do 700 kuna, a za osmi razred oko 1 500 kuna [7].

Cijena elektronskih udžbenika bi trebala biti duplo manja u odnosu na papirnate udžbenike jer kod izdavanja nema troškova tiskanja, distribucije i skladištenja, a neprodani primjerci se neće gomilati. Kupnja tableta i digitalnih udžbenika bi se trebala financijski isplatiti kroz četiri do pet školskih godina. Postupak nabavljanja i cijena tableta još uvijek nije definirana. Na nižu cijenu tableta bi mogle utjecati velike količine narudžbi tableta, participacija proizvođača ili možda pomoć države u obliku investiranja u tablete kao što je prije par godina bilo investiranje države u besplatne udžbenike.

Prednost elektronski udžbenika je i veličina. Školska torba s papirnatim udžbenicima bi postala znatno lakša. Elektronski udžbenici bi omogućili učenicima da umjesto teške torbe nose svega nekoliko stotina grama težak tablet u kojem bi se nalazile stranice tisuće „papirnatih udžbenika". Gužve i redovi u knjižarama bi bili prošlost. Drugim riječima, dostava elektronskih udžbenika je puno brža. Ne zauzimaju baš mnogo prostora i dostupni su 24 sata u danu [8].

Elektronski udžbenici bi mogli djeci olakšati učenje zbog lakšeg pretraživanja, zvukova i animacija. Upisivanjem ključnih pojmova, traženje određene informacije je brzo i jednostavno. Kod papirnatih udžbenika, pretraživanje iziskuje vrijeme potrebno za prelistavanje stranica ili kazala pojmova. Također, olakšano je označavanje bitnih dijelova pomoću opcija kopiranja, lijepljenja ili isticanja teksta raznim bojama. Udžbenici mogu biti pretvoreni i u audio datoteke, što će olakšati učenje učenicima koji najbolje procesuiraju informacije u mozgu na auditoran način. Usmeno izražavanje, animacije, glazba i zvučni efekti stvaraju interaktivnost između učenika i knjige.

Elektronski udžbenici nisu fizički predmeti, osim ako se otiskaju na papir pa ih je zbog toga nemoguće fizički izgubiti ili fizički oštetiti. Jednostavni su za preuzimanje te im u isto vrijeme može pristupiti više korisnika. Također, veličina teksta za čitanje je podesiva što olakšava čitanje djeci koja nose naočale. Po potrebi mogu smanjiti ili povećati tekst [9]. Uvođenje elektronskih udžbenika bi posebno bilo korisno u izoliranim sredinama kao što su brdsko-planinske i otočne škole. Korisno bi bilo i za bolesne učenike kod kuće ili u bolnicama jer bi tako mogli izravno sudjelovati u nastavi [10].

Uz sve navedene prednosti, elektronski udžbenici i njihovo uvođenje imaju i neke nedostatke. Vijek trajanja je nepouzdan u smislu dugoročne digitalne pohrane. Neki od problema su dostupnosti i pouzdanosti interneta te kompatibilnost softvera. Škole imaju internet, ali ne i razredi, što ukazuje na slabu infrastrukturu. Svi nastavnici bi trebali proći edukaciju. Pojavljuje se i problem autorskih prava na udžbenike. Istraživanja pokazuju da djeca koja su potisnula knjigu kao glavno didaktično sredstvo, odnosno informatički pismena djeca mogu imati poremećaj koncentracije i hiperaktivnosti. Mnogi se boje da bi djeca mogla zaboraviti pisati zbog konstantnog korištenja tableta i elektronskih udžbenika.

Ako je tablet priključen na punjač, na struju, učenje nije vremenski ograničeno. Međutim, ako tablet nije priključen na punjač, onda radi pomoću svoje baterije čime je učenje vremenski ograničeno. Nakon što se baterija isprazni, učenik ne može pristupiti udžbenicima sve dok ne napuni bateriju. Neki modeli tableta ne dolaze s promjenjivim baterijama pa kad bateriji istekne radni vijek morate zamijeniti cijeli tablet. Papirnati udžbenik je „zdraviji" od fluorescentnog zaslona jer višesatno gledanje u ekran može izazvati naprezanje očiju. Teže je pročitati rečenicu na zaslonu nego na papiru [11].

IV. ERGONOMIJA

Ergonomija je znanstvena disciplina koja se bavi razumijevanjem interakcija između ljudi i drugih elemenata sustava te struke koja primjenjuje teoriju, načela, podatke i metode za dizajn u cilju optimiziranja ljudske dobrobiti i ukupni sustav performanse.

Ljudski čimbenici i ergonomija su multidisciplinarno područje koje uključuje doprinose iz psihologije, inženjeringa, industrijskog dizajna, grafičkog dizajna, statistike, operacijskih istraživanja i antropometrije. U biti, to je studij projektiranja opreme i uređaja koji utječu na ljudsko tijelo i njegove kognitivne sposobnosti [12].

Poput odraslih, djeca koja provode mnoge sate pred računalom imaju veći rizik od razvijanja problema sa vidom zbog računalne ergonomije. Specijalisti vjeruju da dugotrajno korištenje računala među djecom prouzrokuje rizik progresivne kratkovidnosti.

Djeca provode sate i sate neprekidnog igranja računalnih igara i radom na raznoraznim školskim projektima. Postoji nesklad između veličine dječjeg tijela i veličina namještaja za odrasle i računalne opreme. Raspon veličine tijela kod djece, općenito djevojaka, je znatno različit. One rastu brže nego dječaci u dobi od 8 do 12 godine. Djeca postaju fizički neaktivna i sjede sve više i više. Podesivi namještaj je poželjan, ali zbog različitih učenika često je nemoguć zbog toga što djeca imaju najviše promjenjivo tijelo. 83% osnovnoškolaca sjedi na stolica-stol kombinaciji koje nisu primjerene za njihove tjelesne visine. Djeca u školi sjede otprilike 7 sati po danu. Zbog toga oprema mora biti ergonomska, a što se sve na našem tržištu nudi pod tim nazivom najčešće ne odgovara standardima i normama. Laboratorij za ispitivanje namještaja Šumarskog fakulteta u Zagrebu je utvrdio da se od stolaca koji su se našli u našim trgovinama u ergonomskoj konfekciji, našli kao odgovarajući samo njih 20 % [13]. Većina škola ima namještaj koji je zastario i nije u skladu s minimalnim ortopedskim pravilima. Stolice imaju tvrde sjedalice koje uzrokuju nedostatak cirkulacije, pritisak u ramenima, vrata i leđnih mišića, leđna moždina se pritisne na jednu stranu, sužavanje probavnih organa. Takva je stvar i u kućanstvima. Potrebno je postaviti smjernice za djecu kada je riječ o vremenu koje provode pred računalom.

Pravilni ergonomski dizajn je potreban kako bi se spriječile ponavljajuće deformacije i druge ozljede mišićno-koštanog sustava koji se može razviti tijekom vremena i može dovesti do dugotrajnog invaliditeta.

Djecu treba učiti kako sjediti pred računalom i zato postoji par pravila kojih bi se djeca trebala držati:

- Ne sjediti preblizu monitora,

- Uzeti česte pauze i odmoriti oči,

- Ustati se i protegnuti se i

- Izbjegavanje korištenja miša za duže vrijeme [14].

Najbolja mjera sigurnosti za djecu za rad na računalu su odrasli, njihov nadzor i vođenje.

V. CILJ ISTRAŽIVANJA

Cilj istraživanja je prikazati koliko nastavnici koriste računala u provođenja nastave, koliko su informirani i spremni prihvatiti elektronske udžbenike. Istraživanje je provedeno u srednjim i osnovnim školama u Zagrebu, Sisku i Bistri.

VI. REZULTATI ISTRAŽIVANJA

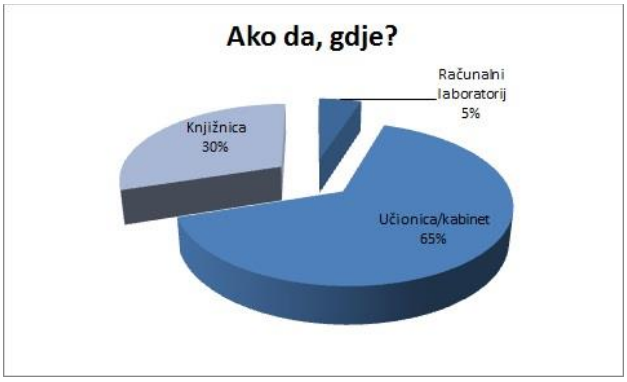
A. *Osnovne škole*

Uzorak u osnovnim školama je bio 30 nastavnika u dvije škole. Od toga je bilo 7 muških nastavnika i 23 ženske nastavnice. Anketa se sastojala od niza pitanja na koja su nastavnici odgovarali na ponuđene odgovore.

Na pitanje: „ Koristite li računalo tijekom provođenja nastave? “ 67% nastavnika je odgovorilo „Da“ dok je 33% nastavnika odgovorilo „Ne“ (Slika 1).

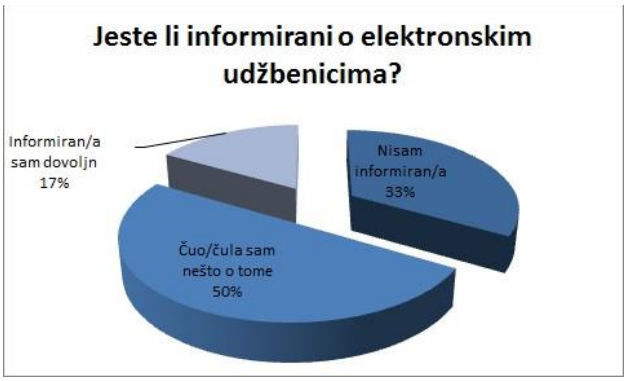
Slika 1. Korištenje računala u nastavi u osnovnoj školi

Na pitanje: „Ako da, gdje?” 5% nastavnika je odgovorilo „Računalni laboratorij”, 30% nastavnika je odgovorilo „Knjižnica” dok je 65% nastavnika odgovorilo „Učionica/kabinet” (Slika 2).



Slika 2. Mjesto korištenja računala

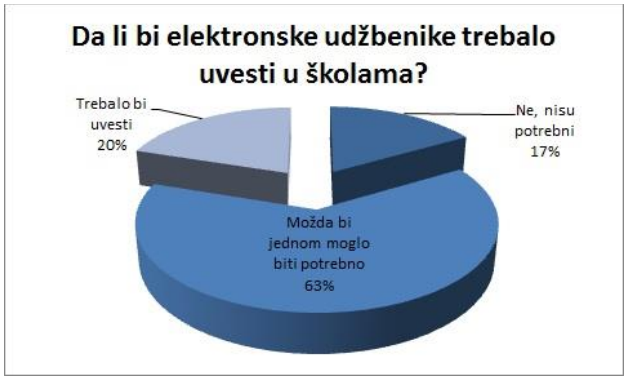
Na postavljeno pitanje: „ Imate li e-mail adresu?” svi nastavnici su potvrdno odgovorili da imaju e-mail adresu. Na pitanje: „Jeste li informirani o elektronskim udžbenicima?“ 17% nastavnika je odgovorilo „Informiran/a sam dovoljno“, 33% ih je odgovorilo „Nisam informiran/a“ dok je 50% nastavnika osnovnih škola odgovorilo sa „Čuo/čula sam nešto o tome“ (Slika 3).



Slika 3. Informiranost nastavnika o elektronskim udžbenicima

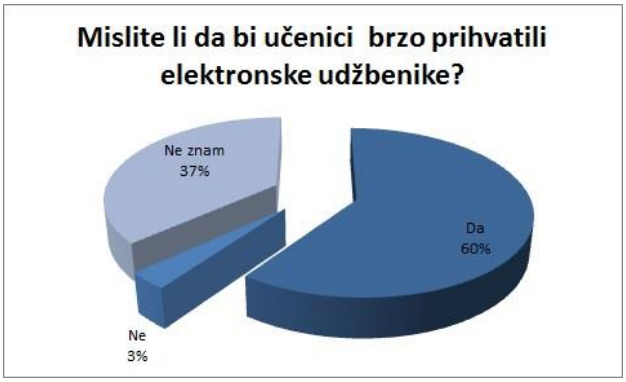
Sljedeće pitanje je bilo: „Ako da, navedite barem jednu pozitivnu i negativnu stranu elektronskog udžbenika.“ Najčešći odgovori za pozitivne strane uvođenja elektronskog udžbenika su bili: manja opterećenost učeničkih torbi, brzi način komuniciranja, multimedijski pristup učenju, razvijanje informatičke pismenosti. Neki od najčešćih odgovora za negativne strane elektronskih udžbenika su bili: nedostupnost tehnologije svima, nedovoljna razvijenost škola, zabrinutost oko neposjećivanja knjižnica, vrijeme prilagodbe za rad s takvim udžbenikom.

Na pitanje: „Da li bi elektronske udžbenike trebalo uvesti u školama?“ 17% nastavnika je odgovorilo „Ne, nisu potrebni“, 20% ih je odgovorilo „Trebalo bi uvesti“ te 63% nastavnika je odgovorilo „Možda bi jednom moglo biti potrebno“ (Slika 4).



Slika 4. Potreba uvođenja elektronskog udžbenika

Na pitanje: „Mislite li da bi učenici brzo prihvatili elektroničke udžbenike?“ 3% nastavnika misli da učenici ne bi prihvatili elektronske udžbenike, njih 37% nije znalo kako bi učenici prihvatili dok je njih 60% odgovorilo kako bi učenici brzo prihvatili elektronske udžbenike (Slika 5) .



Slika 5. Prihvaćanje elektronskog udžbenika od strane učenika

Na zadnje pitanje u anketi: „Mislite li da bi nastavnici brzo prihvatili elektroničke udžbenike?“ 17% nastavnika je odgovorilo kako nastavnici ne bi brzo prihvatili elektronske udžbenike, 33% nastavnika misli kako bi nastavnici brzo prihvatili elektronske udžbenike dok 50% nastavnika ne zna kako bi nastavnici prihvatili elektronske udžbenike (Slika 6).

Slika 6. Prihvaćanje elektronskog udžbenika od strane nastavnika

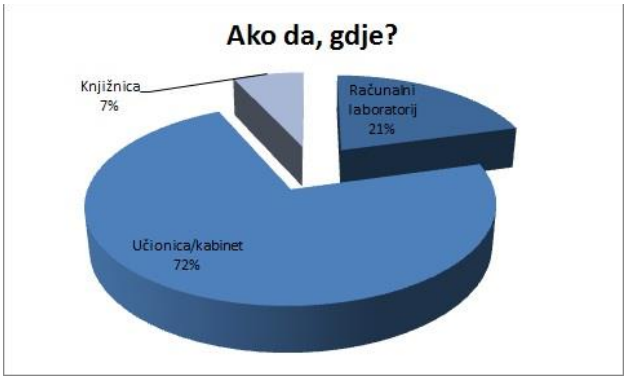
B. Srednje škole

Uzorak u srednjoj školi je bio 40 nastavnika u dvije škole. Od toga je bilo 16 muških nastavnika i 23 ženske nastavnice.

Na pitanje: „ Koristite li računalo tijekom provođenja nastave? “ 28% nastavnika je odgovorilo „Ne“ dok je 72% nastavnika odgovorilo „Da“ (Slika 7).

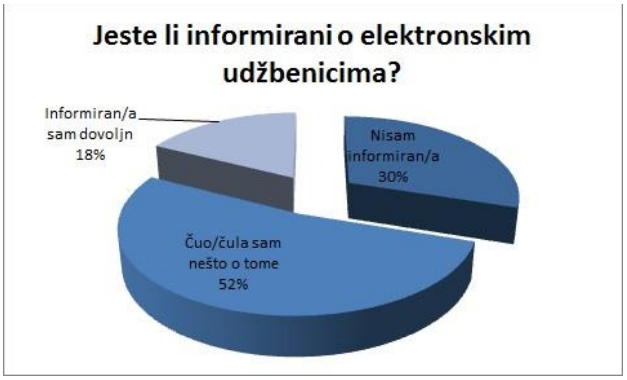
Slika 7. Korištenje računala u nastavi u srednjoj školi

Na pitanje: „Ako da, gdje?” 7% nastavnika je odgovorilo „Knjižnica”, 21% nastavnika je odgovorilo „Računalni laboratorij” dok je 72% nastavnika odgovorilo „Učionica/kabinet” (Slika 8).



Slika 8. Mjesto korištenja računala

Na postavljeno pitanje: „ Imate li e-mail adresu?” svi nastavnici su potvrdno odgovorili da imaju e-mail adresu. Na pitanje: „Jeste li informirani o elektronskim udžbenicima?“ 18% nastavnika je odgovorilo „Informiran/a sam dovoljno“, 30% ih je odgovorilo „Nisam informiran/a“ dok je 52% nastavnika osnovnih škola odgovorilo sa „Čuo/čula sam nešto o tome“ (Slika 9).

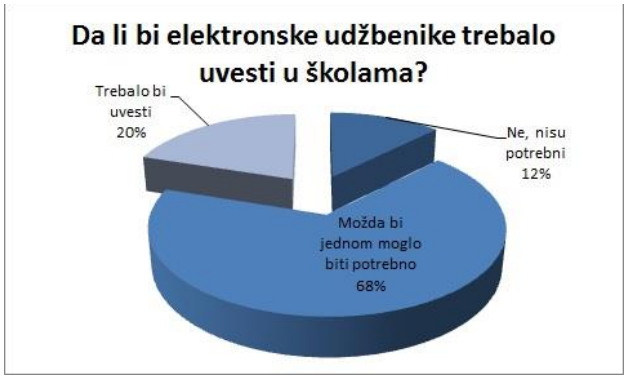


Slika 9. Informiranost nastavnika o elektronskim udžbenicima

Sljedeće pitanje je bilo: „Ako da, navedite barem jednu pozitivnu i negativnu stranu elektronskog udžbenika.“

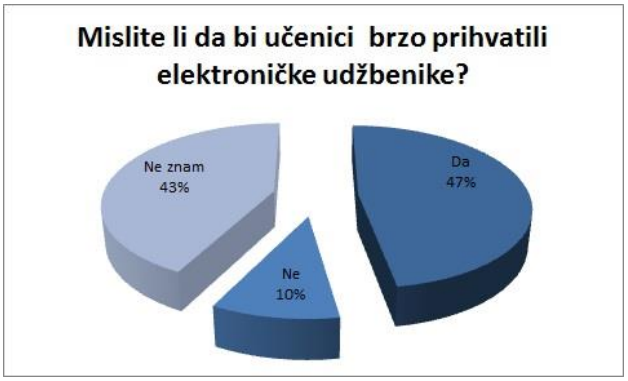
Najčešći odgovori za pozitivne strane uvođenja elektronskog udžbenika su bili: manja opterećenost učeničkih torbi, brzi način komuniciranja, multimedijski pristup učenju, razvijanje informatičke pismenosti. Neki od najčešćih odgovora za negativne strane elektronskih udžbenika su bili: nedostupnost tehnologije svima, nedovoljna razvijenost škola, zabrinutost oko neposjećivanja knjižnica, vrijeme prilagodbe za rad s takvim udžbenikom.

Na pitanje: „Da li bi elektronske udžbenike trebalo uvesti u školama?“ 12% nastavnika je odgovorilo „Ne, nisu potrebni“, 20% ih je odgovorilo „Trebalo bi uvesti“ te 68% nastavnika je odgovorilo „Možda bi jednom moglo biti potrebno“ (Slika 10).



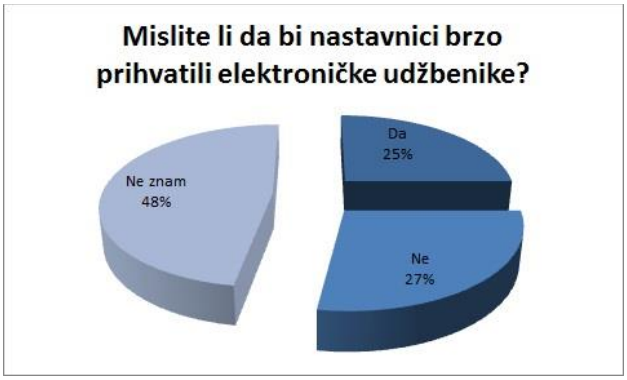
Slika 10. Potreba uvođenja elektronskog udžbenika

Na pitanje: „Mislite li da bi učenici brzo prihvatili elektroničke udžbenike?“ 10% nastavnika misli da učenici ne bi prihvatili elektronske udžbenike, njih 43% nije znalo kako bi učenici prihvatili dok je njih 47% odgovorilo kako bi učenici prihvatili elektronske udžbenike (Slika 11).



Slika 11. Prihvaćanje elektronskog udžbenika od strane učenika

Na zadnje pitanje u anketi: „Mislite li da bi nastavnici brzo prihvatili elektroničke udžbenike?“ 27% nastavnika je odgovorilo kako nastavnici ne bi brzo prihvatili elektronske udžbenike, 25% nastavnika misli kako bi nastavnici brzo prihvatili elektronske udžbenike dok 48% nastavnika ne zna kako bi nastavnici prihvatili elektronske udžbenike (Slika 12).



Slika 12. Prihvaćanje elektronskog udžbenika od strane nastavnika

VII. INTERPRETACIJA REZULTATA ISTRAŽIVANJA

Gledajući rezultate anketa te pogledom na grafikone, vidljivo je kako se odgovori na prvo pitanje previše ne razlikuju u osnovnim i srednjim školama. Nastavnici koriste računalo u provođenju nastave oko 70% što je veliki postotak. Veća razlika korištenja računala u osnovnoj i srednjoj školi je u prostorijama gdje se računala koriste što je donekle i razumljivo jer u srednjim školama sve više nastavnika svoju nastavu bazira na računalima te tako više koriste računalne laboratorije. Postotak informiranosti o elektronskim udžbenicima je vrlo sličan i u srednjim i osnovnim školama te iznosi oko 18%. Na pitanje treba li uvesti elektronske udžbenike, postotak potvrdnih odgovora je kako u osnovnim tako i u srednjim školama i iznosi 20%. Najveće razlike u odgovorima između osnovnih i srednjih škola su na pitanjima da li bi učenici i nastavnici brzo prihvatili elektronske udžbenike. U osnovnim školama 60% nastavnika misli kako bi učenici brzo prihvatili elektronske udžbenike dok u srednjim školama taj postotak iznosi 47%. Na pitanje da li bi nastavnici brzo prihvatili elektronske udžbenike, rezultati su vrlo slični kao i kod prethodnog pitanja te iznose 33% za osnovne škole te 25% za srednje škole. Tu je najveća razlika kod negativnog odgovora gdje su nastavnici u osnovnim školama rekli kako nastavnici ne bi prihvatili elektronske udžbenike i to njih 17% dok su za isto pitanje 27% nastavnika u srednjim školama rekli kako ne bi prihvatili elektronske udžbenike.

VIII. ZAKLJUČAK

Digitalno doba sve više kuca na vrata hrvatskih osnovnih i srednjih škola. Uvođenje elektronskih udžbenika ne bih trebalo biti tabu tema. Kao i za sve, postoje pozitivne i negativne strane, ali sve u svemu, bitno je da se učenicima olakša njihov učenički život.

Provođenjem anketa za nastavnike, u hrvatskim školama se pokazao glavni problem, a to je neinformiranost nastavnika i proces za prilagodbu na novitete.

Hrvatski izdavači još su uvijek slijepi za elektronske udžbenike ali se nadamo da će se to kroz brži vremenski period pokrenuti u smjeru uvođenja elektronskih udžbenika.

LITERATURA

[1] http://www.zadarskilist.hr/clanci/10022011/velika-lomaca-za-tiskane-knjige Pristupljeno 01.02.2013.

[2] http://www.novilist.hr/Vijesti/Hrvatska/E-udzbenik-pojeftinjuje-udzbenike-i-olaksava-torbu Pristupljeno 01.02.2013.

[3] http://danas.net.hr/hrvatska/robert-sipek-neven-anticevic-skolski-udzbenici-djecu-uce-da-je-sanader-premijer-a-mesic-predsjednik Pristupljeno 01.02.2013.

[4] http://www.novilist.hr/Vijesti/Hrvatska/E-udzbenik-pojeftinjuje-udzbenike-i-olaksava-torbu Pristupljeno 01.02.2013.

[5] http://www.carnet.hr/info2012/opisi_predavanja Pristupljeno 01.02.2013.

[6] http://www.slobodnadalmacija.hr/Hrvatska/tabid/66/articleType/ArticleView/articleId/188170/Default.aspx Pristupljeno 01.02.2013.

[7] http://www.zadarskilist.hr/clanci/21082012/visi-razred---veci-udar-na-kucni-budzet Pristupljeno 01.02.2013.

[8]http://textbooks-online-review.toptenreviews.com/advantages-and-disadvantages-of-e-textbooks-online.html Pristupljeno 02.02.2013.

[9] http://blookict.blogspot.com/2012/10/advantages-and-disadvantages-of-e-books.html Pristupljeno 02.02.2013.

[10] http://www.dubrovnikportal.com/djecji_kantun/index.php?box_id=111&clanak_id=50171 Pristupljeno 02.02.2013.

[11] http://www.ehow.com/list_6157579_advantages-disadvantages-electronic-books.html Pristupljeno 02.02.2013.

[12] http://en.wikipedia.org/wiki/Human_factors_and_ergonomics Pristupljeno 03.02.2013.

[13] http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=48423 Pristupljeno 03.02.2013.

[14] http://www.umanitoba.ca/faculties/kinrec/bsal/miniu/summer/backpacks.pdf

Istraživanje utjecaja rada za računalom na zdravlje učenika Osnovne škole Centar u Puli

Branka Kontošić, dipl.učitelj
Osnovna škola Centar
Danteov trg 2, Pula, Hrvatska
E-mail: b_kontosic@net.hr

Mr.sc. Alen Kontošić, mag.oec.
Sveučilište Jurja dobrile u Puli
Zagrebačka 30, Pula, Hrvatska
E-mail: akontosi@net.hr

Sažetak – Prilikom rada za računalom potrebno je voditi računa i o zdravlju korisnika odnosno učenika. U tu je svrhu u Osnovnoj školi Centar u Puli provedeno istraživanje anonimnim ispitivanjem učenika. Analizom prikupljenih podataka došlo se do rezultata u kojoj mjeri rad za računalom utječe na zdravstveno stanje učenika. Rad za računalom, osobito dugotrajni, izaziva probleme s vidom, glavobolju, bolove u leđima te je važno voditi brigu o ergonomskim uvjetima za rad.

Ključne riječi: rad za računalom, utjecaj dugotrajnog rada na računalu, zdravlje učenika.

I. UVODNA RAZMATRANJA

A. Uvod

Blagodati modernog doba i ubrzanog razvoja tehnologije donose niz pogodnosti u obavljanju svakodnevnih obveza koje znaju prepoznati sve generacije, od najmlađih do najstarijih članova obitelji. Međutim, rad za računalom, osobito dulji i češći rad, može izazvati i poneke zdravstvene probleme kod korisnika. Važno je voditi računa o ergonomskim uvjetima koji osiguravaju ugodan rad za računalom i utječu na zdravstveno stanje korisnika. Veći broj zadovoljenih ergonomskih uvjeta omogućava ugodniji rad, a time i mogućnost dužeg rada za računalom.

Mogućnosti informacijsko-komunikacijske tehnologije treba koristiti radi lakšeg i bržeg obavljanja obveza koje nam nameće brzi tempo života, ali pritom uvažavati signale koje nam šalje tijelo s ciljem očuvanja što boljega zdravstvenog stanja.

B. Svrha i ciljevi istraživanja

Cilj rada *Istraživanje utjecaja rada za računalom na zdravlje učenika Osnovne škole Centar u Puli* jest istražiti koliko vremena učenici provedu kod kuće za računalom te usporediti dužinu rada za računalom sa tegobama koji učenici osjećaju nakon takvog rada. Između ostalog, želi se istražiti tko među učenicima Osnovne škole Centar u Puli više koristi računalo i koje su posljedice toga za učenike nižih razreda (razredne nastave) i učenike viših razreda (predmetne nastave) te osjećaju li tegobe više dječaci ili djevojčice nakon rada za računalom.

II. METODOLOGIJA PROVEDENOG ISTRAŽIVANJA

Istraživanje je izvršeno anketnim upitnikom kojim su bili anonimno ispitani učenici nižih i viših razreda. U istraživanju se primijenio postupak anketiranja s mjernim instrumentom anketnog lista s ukupno 11 pitanja s unaprijed ponuđenim DA-NE odgovorima. U istraživanju su se osim navedene metode anketiranja primijenile i druge znanstvene metode poput statističke metode, metode deskripcije, analize i sinteze. Podaci su obrađeni primjenom programa za tablično računanje MS Excel, a primijenjene su statističke metode mjere centralne tendencije (frekvencija i postotak).

III. REZULTATI I ANALIZA DOBIVENIH REZULTATA

Istraživanje je provedeno u Osnovnoj školi Centar u Puli na uzorku od 200 učenika od ukupno 282 učenika (od kojih 43,98% dječaka i 56,02% djevojčica) te je u analizi rezultata vrijedno spomenuti velik postotak (p = 0,709) anketiranih učenika škole: 91 učenika nižih razreda (72,8% od 125 učenika nižih razreda škole od kojih je 42,4% dječaka i 57,6% djevojčica) te 109 učenika viših razreda (69,43% od 157 učenika viših razreda škole od kojih je 45,2% dječaka i 54,8% djevojčica). Obzirom da je ispunjene upitnike vratilo toliko velik broj učenika stopa povrata može se ocijeniti vrlo zadovoljavajućom, uzimajući u obzir da se u sličnim istraživanjima ona kreće u puno manjem postotku. Stoga, standardna pogreška proporcije je mala obzirom da je *n* velik (n > 30). Veličina aritmetičke sredine približno je normalno distribuirana te stoga, prema pravilu *dvije sigme*, postiže s vjerojatnošću većom od 0,95 sve vrijednosti u intervalu ± *dvije sigme oko sredine*. Obzirom da je gotovo 71% učenika ispunilo anketni upitnik i da su prikupljeni upitnici u velikom postotku zastupani u razrednoj i u predmetnoj nastavi uzorak se može ocijeniti reprezentativnim pa je na temelju njega moguće donositi odgovarajuće zaključke koji se odnose na testiranu populaciju.

Analizom prikupljenih podataka došlo se do rezultata koji pokazuju da 94,5% učenika od ukupnog broja ispitanih učenika posjeduje računalo kod kuće, od kojih 93,7% čine dječaci te 95,2% djevojčice u odnosu na ukupan broj zasebno ispitanih dječaka i djevojčica. Rezultati istraživanja ukazuju da 61,5% učenika svakodnevno koristi računalo, od toga 65,3% učenika te nešto manje učenice, 58,1%. Koliko vremena dnevno provedu ispred računala rezultati pokazuju da više od polovice ispitanih učenika, njih 61,5%, dnevno provedu do 1 sata na računalu i to 58,1% djevojčica i 65,3% dječaka. Od 1 do 3 sata dnevno računalo koristi oko 37%, i to učenici u nešto većem postotku (37,9%) od djevojčica (36,2%). Manji broj učenika (22%) za računalom provede dnevno od 3 do 5 sati, i to uglavnom učenici (30,5%) te u mnogo manjem broju učenice (14,3%). Veći dio svog slobodnog vremena, više od 5 sati, svakodnevno provodi 15,5% učenika za računalom, 18,9% učenici i 12,4% učenice. Kod rada za računalom, pogotovo dužeg rada za računalom, probleme s očima primjećuje 26% ispitanih učenika, od kojih učenice u većem broju (35,2%) od učenika (15,8%). Kod ovog su pitanja učenici u velikoj većini zaokruživali ili podvlačili redoslijed ponuđenih odgovora (svrbež, peckanje, suzenje očiju) te je temeljem toga određen odgovor, prema intenzitetu frekvencije prikazanih odgovora, bio na prvom mjestu suzenje očiju, peckanje, a tek onda svrbež, prema spolu više djevojčice nego dječaci. Na pitanje rade li očne vježbe (gledanje u daljinu), 39% učenika ih radi, iako o važnosti navedenog uče od 1. razreda na nastavi informatike. Na pitanje osjećaju li glavobolju nakon rada za računalom, odgovorilo je potvrdno 16% učenika (dječaci 10,5%, djevojčice 21% od ukupnog broja ispitanih učenika). Prema intenzitetu frekvencije najviše učenici koji su radili za računalom više od 5 sati dnevno, nakon njih oni koji su radili od 3 do 5 sati, manji broj učenika kod rada za računalom od 1 do 3 sata te najmanje oni koji su radili do 1 sata dnevno, prema spolu više djevojčice nego dječaci. Bolove u leđima prilikom rada za računalom osjeća 27% učenika (dječaci 17,9%, djevojčice 35,2% od ukupnog broja ispitanih učenika) te su i u ovom slučaju to odgovorili, prema intenzitetu frekvencije, najviše učenici koji su radili za računalom više od 5 sati dnevno, nakon njih oni koji su radili od 3 do 5 sati, manji broj učenika kod rada za računalom od 1 do 3 sata te najmanje oni koji su radili do 1 sata dnevno, prema spolu više djevojčice nego dječaci. Bolove u rukama (ruci) prilikom rada na računalu osjeća 20% učenika (dječaci 13,7%, djevojčice 25,7% od ukupnog broja ispitanih učenika). I u ovom su slučaju to odgovorili učenici, prema intenzitetu frekvencije, najviše učenici koji su radili za računalom više od 5 sati dnevno, nakon njih oni koji su radili od 3 do 5 sati, manji broj učenika kod rada za računalom od 1 do 3 sata te najmanje oni koji su radili do 1 sata dnevno, prema spolu i ovdje više djevojčice nego dječaci.

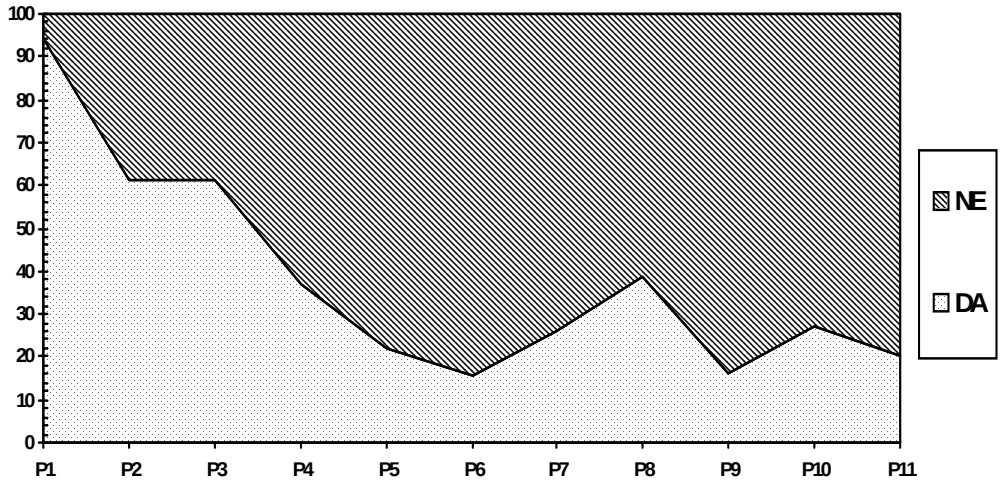
Uspoređujući razrednu nastavu i predmetnu nastavu u posjedovanju računala kod kuće, nema razlika, odnosno 94,5% učenika posjeduje računalo kod kuće. Računalo svakodnevno koristi 72,5% učenika predmetne nastave te 48,4% razredne nastave, učenici u nešto većem broju koriste računalo od učenica u razrednoj nastavi. Do 1 sata dnevno računalo koristi 68,1% učenika razredne te 56% učenika predmetne nastave, učenice u nešto manjem broju od učenika. Od 1 do 3 sata dnevno računalom se koristi 27,5% ispitanih učenika razredne nastave te u većem broju učenici predmetne nastave (45%). Manji broj učenika, njih 16,5%, ispitanih u razrednoj nastavi, koristi računalo od 3 do 5 sati, te njih 26,6% u predmetnoj nastavi, većinom su to dječaci (22% u razrednoj i 37% u predmetnoj nastavi) u odnosu na djevojčice (12% u razrednoj i 16,4% u predmetnoj nastavi). Na pitanje koriste li računalo više od 5 sati dnevno, u razrednoj nastavi potvrdno je odgovorilo 16,5% ispitanih učenika 19,5% učenici, 14% učenice), a u predmetnoj nastavi 14,7% (18,5% učenici, 10,9% učenice). Kod dužeg rada za računalom, probleme s očima primjećuje 35,2% ispitanih učenika razredne nastave, od kojih učenice u većem broju (48%) od učenika (19,5%), te 18,3% ispitanih učenika predmetne nastave, od kojih također

učenice u većem broju (23,6%) od učenika (13%). Kod ovog su pitanja učenici razredne i predmetne nastave u velikoj većini zaokruživali ili podvlačili redoslijed ponuđenih odgovora (svrbež, peckanje, suzenje očiju) te je temeljem toga određen odgovor, prema intenzitetu frekvencije prikazanih odgovora, bio na prvom mjestu suzenje očiju, peckanje, a tek onda svrbež, prema spolu više djevojčice nego dječaci i kod razredne nastave i kod predmetne nastave. Na pitanje o očnim vježbama (gledanje u daljinu), 46,2% učenika razredne nastave odgovorilo je potvrdno da rade takve vježbe (dječaci 36,6%, djevojčice 54%), dok je tek 33% učenika predmetne nastave odgovorilo potvrdno da rade takve vježbe (dječaci 40,7%, a djevojčice tek 25,5%). Na pitanje osjeća li glavobolju nakon rada za računalom, odgovorili su potvrdno u većem broju učenici razredne nastave 20,9% (dječaci 12,2%, djevojčice 28% od ukupnog broja ispitanih učenika) u odnosu na 11,9% učenika predmetne nastave (dječaci 9,3%, djevojčice 14,5% od ukupnog broja ispitanih učenika). I u ovome su slučaju to odgovorili učenici, prema intenzitetu frekvencije, najviše učenici koji su radili za računalom više od 5 sati dnevno, nakon njih oni koji su radili od 3 do 5 sati, manji broj učenika kod rada za računalom od 1 do 3 sata te najmanje oni koji su radili do 1 sata dnevno, prema spolu više djevojčice nego dječaci. Bolove u leđima prilikom rada za računalom osjećali su u većem broju učenici predmetne nastave 28,4% (dječaci 16,7%, djevojčice 40% od ukupnog broja ispitanih učenika) u odnosu na 25,3% učenika razredne nastave (dječaci 19,5%, djevojčice 30% od ukupnog broja ispitanih učenika). Prema intenzitetu frekvencije, u ovome slučaju to su bili najviše učenici koji su radili za računalom više od 5 sati dnevno, nakon njih oni koji su radili od 3 do 5 sati, manji broj učenika kod rada za računalom od 1 do 3 sata te najmanje oni koji su radili do 1 sata dnevno, prema spolu i ovdje više djevojčice nego dječaci. Bolove u rukama (ruci) prilikom rada za računalom osjećali su u većem broju učenici razredne nastave 30,8% (dječaci 22%, djevojčice 38% od ukupnog broja ispitanih učenika) u odnosu na 11% učenika predmetne nastave (dječaci 7,4%, djevojčice 14,5% od ukupnog broja ispitanih učenika). Prema intenzitetu frekvencije i u ovome su slučaju to bili najviše učenici koji su radili za računalom više od 5 sati dnevno, nakon njih oni koji su radili od 3 do 5 sati, manji broj učenika kod rada za računalom od 1 do 3 sata te najmanje oni koji su radili do 1 sata dnevno, prema spolu i ovdje više djevojčice nego dječaci. Podaci anketnog istraživanja prikazani su u tabličnom prikazu 1. te u grafičkim prikazima na slikama 1., 2. i 3.

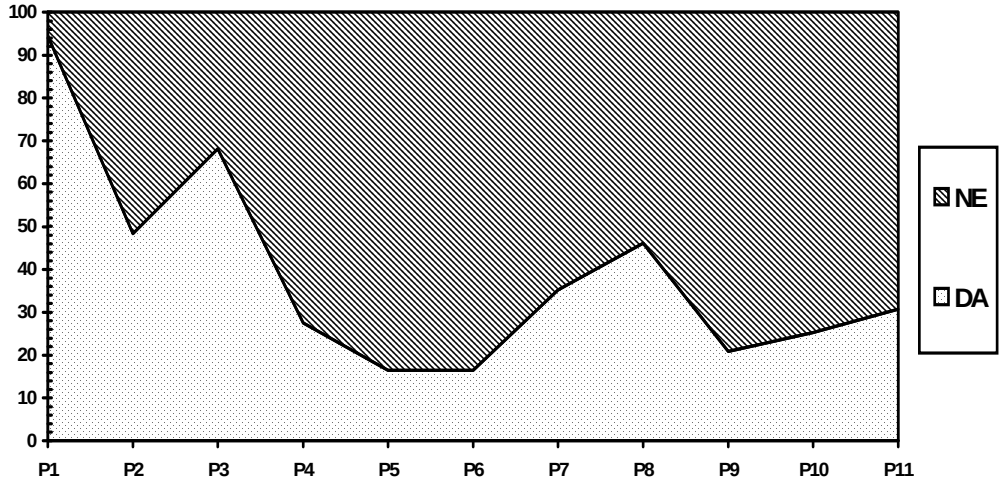
	Pitanje	Razredna nastava (RN):				Predmetna nastava (PN):				Ukupno (RN+PN):			
		DA [%]		NE [%]		DA [%]		NE [%]		DA [%]		NE [%]	
		M [%]	Ž [%]	M [%]	Ž [%]	M [%]	Ž [%]	M [%]	Ž [%]	M [%]	Ž [%]	M [%]	Ž [%]
P1	Imaš li kod kuće računalo?	94,5		5,5		94,5		5,5		94,5		5,5	
		95,1	94,0	4,9	6,0	92,6	96,4	7,4	3,6	93,7	95,2	6,3	4,8
P2	Koristiš li računalo svakodnevno?	48,4		51,6		72,5		27,5		61,5		38,5	
		56,1	42,0	43,9	58,0	72,2	72,7	27,8	27,3	65,3	58,1	34,7	41,9
P3	Koristiš li računalo do 1 sata dnevno?	68,1		31,9		56,0		44,0		61,5		38,5	
		73,2	64,0	26,8	36,0	59,3	52,7	40,7	47,3	65,3	58,1	34,7	41,9
P4	Koristiš li računalo od 1 do 3 sata dnevno?	27,5		72,5		45,0		55,0		37,0		63,0	
		34,1	22,0	65,9	78,0	40,7	49,1	59,3	50,9	37,9	36,2	62,1	63,8
P5	Koristiš li računalo od 3 do 5 sati dnevno?	16,5		83,5		26,6		73,4		22,0		78,0	
		22,0	12,0	78,0	88,0	37,0	16,4	63,0	83,6	30,5	14,3	69,5	85,7
P6	Koristiš li računalo više od 5 sati dnevno?	16,5		83,5		14,7		85,3		15,5		84,5	
		19,5	14,0	80,5	86,0	18,5	10,9	81,5	89,1	18,9	12,4	81,1	87,6
P7	Kod rada na računalu osjećaš li probleme s očima? (svrbež, peckanje, suzenje očiju)	35,2		64,8		18,3		81,7		26,0		74,0	
		19,5	48,0	80,5	52,0	13,0	23,6	87,0	76,4	15,8	35,2	84,2	64,8
P8	Radiš li očne vježbe? (gledanje u daljinu)	46,2		53,8		33,0		67,0		39,0		61,0	
		36,6	54,0	63,4	46,0	40,7	25,5	59,3	74,5	38,9	39,0	61,1	61,0
P9	Osjećaš li glavobolju nakon rada za računalom?	20,9		79,1		11,9		88,1		16,0		84,0	
		12,2	28,0	87,8	72,0	9,3	14,5	90,7	85,5	10,5	21,0	89,5	79,0
P10	Osjećaš li bolove u leđima prilikom rada za računalom?	25,3		74,7		28,4		71,6		27,0		73,0	
		19,5	30,0	80,5	70,0	16,7	40,0	83,3	60,0	17,9	35,2	82,1	64,8
P11	Osjećaš li bol u rukama (ruci) prilikom rada za računalom?	30,8		69,2		11,0		89,0		20,0		80,0	
		22,0	38,0	78,0	62,0	7,4	14,5	92,6	85,5	13,7	25,7	86,3	74,3

Tablica 1. Tablični prikaz rezultata

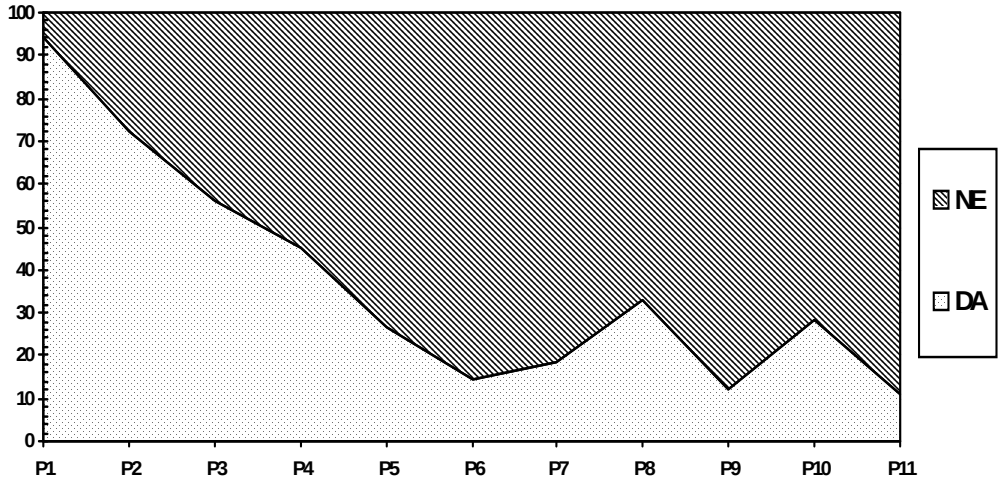
Na sljedećim se grafovima vidi odnos pozitivnih i negativnih odgovora svih učenika škole na postavljena DA-NE pitanja.



Slika 1. Pozitivni i negativni (u %) odgovori učenika na postavljena pitanja



Slika 2. Pozitivni i negativni (u %) odgovori učenika razredne nastave (1. - 4. razreda) na postavljena pitanja



Slika 3. Pozitivni i negativni (u %) odgovori učenika predmetne nastave (5. - 8. razreda) na postavljena pitanja

IV. ZAKLJUČAK

Obradom podataka dobiveni su rezultati koji su dali odgovore na postavljena pitanja o utjecaju rada za računalm na zdravlje učenika Osnovne škole Centar u Puli. Uspoređujući duljinu rada za računalom i tegobe navedene u pitanjima ustanovljeno je da duljina rada za računalom utječe na intenzitet frekvencija navedenih zdravstvenih tegoba u pitanjima te se nalaze u proporcionalnom odnosu. Učenici koji više sati dnevno provode za računalom, osjećaju i više tegoba, osobito suzenje očiju, glavobolju i bolove u leđima i rukama. Što manje vremena učenici provode za računalom to osjećaju manje tegoba ili ih uopće ne osjećaju na što ukazuju i dobiveni rezultati. Analiza rezultata pokazuje da djevojčice osjećaju navedene tegobe u većem broju od dječaka. Učenici predmetne nastave više vremena provode u radu za računalom, ali manje osjećaju navedene zdravstvene tegobe u odnosu na učenike nižih razreda, osim kod problema s bolovima u leđima.

V. ZAHVALE

Autori su zahvalni Marleni Lakić, prof. što je pročitala rad i dala im korisne povratne informacije i sugestije.

LITERATURA

[1] LJ. Baban, S. Jelinić, M. Lamza-Maronić, K. Ivić, "Primjena metodologije znanstvenog istraživanja", Ekonomski fakultet Osijek, Osijek, 1993.
[2] B.L. Cole, J.D. Maddocks, K. Sharpe, „Effects of VDUs on the eyes: Report of a 6-year epidemiological study", Optom Vis Sci, 73(8):512-28,1996.
[3] [4] Healy, JM (1998). *Failure to connect: How computers affect our children's minds—for better and worse* . New York: Simon & Schuster.

OpenProj - open source softverski alat za upravljanje IT projektima

D. Valenčić*, D. Kažović** i A. Radovan*
* Veleučilište Velika Gorica, Velika Gorica, Hrvatska
** Ekonomski fakultet Rijeka, Rijeka, Hrvatska
davorin_valencic@yahoo.com, dkazovic@gmail.com, aleksander.radovan@gmail.com

Sažetak - OpenProj je open-source softverski alat namijenjen vođenju projekata i alternativa komercijalnom softveru MS Project. U ovom članku su prikazane osnovne karakteristike OpenProj softvera, dana je usporedba s MS Project softverom i analizirane mogućnosti primjene OpenProj softvera u praktičnom dijelu edukacije iz područja upravljanja projektima na visokoškolskom IT stručnom studiju.

I. UVOD

Danas je teško zamislivo vođenje projekata bez upotrebe informacijskih i komunikacijskih tehnologija. Osim vještina upravljanja projektima, potrebne su nam i IT vještine i alati za prijavu, vođenje i dokumentiranje projekata.

Efikasno upravljanje projektima zahtjeva više od kvalitetnog planiranja. Ono traži da sve relevantne informacije budu dobavljene, analizirane i dostupne u što kraćem vremenskom roku. Takav pristup omogućava otkrivanje problema u ranoj fazi i efikasan pronalazak alternativnih rješenja.

Danas su raspoloživi softverski alati za upravljanje projektima pomoću kojih je moguće praćenje i kontroliranje projekata. Softver za upravljanje projektima pruža izuzetnu pomoć voditelju projekta u praćenju različitih parametara i zadataka koji su vrlo često međusobno povezani.

Ipak mora biti jasno da najsloženiji i najefikasniji softverski paket ne može zamijeniti potrebna znanja i vještine u upravljanju projektima. Zbog toga će se na početku ovog članka ukratko prikazati najčešće metodologije i pristupi korišteni pri upravljanju projektima. Nakon toga će se razmotriti karakteristike softverskih alata za upravljanje projektima. Zatim će se analizirati besplatni softver OpenProj i mogućnost njegove primjene u edukaciji upravljanja projektima na visokoškolskom stručnom studiju.

A. *Vještine i znanja upravljanja projektima*

Voditelji projekata informacijskih sustava za upravljanje projektima danas koriste različite metodologije (na primjer Prince2, Agile, i druge). Metodologija se definira kao skup smjernica i principa koji se mogu skrojiti i primijeniti specifičnoj situaciji [1].

Ovaj rad je podržan od Veleučilišta Velika Gorica, Velika Gorica, Hrvatska

Najvažnije karakteristike dobre metodologije upravljanja projektima su uporaba predložaka, standardizirane tehnike planiranja, vremenskoga određivanja i kontrole troškova, standardizirani oblik izvještavanja, fleksibilnost metodologije za moguću primjenu na različitim vrstama projekata, jednostavnost i razumljivost za korisnike, uporaba standardiziranih faza životnoga ciklusa kao i temeljenost na smjernicama (umjesto na procedurama) i na profesionalnoj etici. Ukoliko se sve dobro primijeni, zbog bolje kontrole nad opsegom projekta može se očekivati brže vrijeme izlaska na tržište, smanjeni rizik projekta, bolji proces donošenja odluka, veće zadovoljstvo kupca te više vremena za dodane vrijednosti.

Postoji i nekoliko udruženja koja se bave upravljanjem projektima, na primjer Association for Project Management (APM), The International Project Management Association (IPMA) i Project Management Institute (PMI). Ova udruženja nastoje razviti međunarodne standarde za upravljanje projektima, gdje je standard formalni dokument koji opisuje ustanovljene norme, metode, procese i prakse [2].

Tako je Project Management Institute (PMI) sa svojim Vodičem kroz znanje upravljanja projektima (Vodič kroz PMBOK) postao priznati IEEE standard [3] .

Vodič kroz PMBOK identificira onu podskupinu znanja (procese, vještine, alate i tehnike) o upravljanju projektima koja je općenito priznata kao dokazana praksa [3]. Koristeći izraz „općenito priznata“ PMI izražava da su opisano znanje i praksa primjenjivi na većinu projekata većinu vremena i da postoji konsenzus o njihovoj vrijednosti i korisnosti.

Prema PMI definiciji projekt je „privremeni poduhvat poduzet u svrhu stvaranja jedinstvenog proizvoda, usluge ili rezultata“ [3]. Osnovne karakteristike projekta su njegova privremenosti i jedinstvenost, jer svaki projekt ima svoj određeni početak i određeni kraj, i u tome se razlikuje od operativnog rada, odnosno proizvodnje. Uspješan projekt traje što kraće da bi se s zadanim troškovima postigao neki određeni cilj (proizvod, usluga), operativni rad je uspješniji ako što duže ostvarujemo dobit od proizvodnje nekog proizvoda ili davanje neke usluge. Upravljanje projektima je primjena znanja, vještina, alata i tehnike da bi projektne aktivnosti zadovoljile zahtjevima projekta. Upravljanje projektima ostvaruje se kroz primjenu i integraciju grupa procesa upravljanja

projektima pokretanja, izvršavanje, nadzor, kontrola i zatvaranje [3].

Vodič kroz PMBOK objašnjava da softver za upravljanje projektima može pomoći u planiranju, organiziranju i upravljanju resursima i razvoju procjena resursa. Ovisno o osobinama softvera, mogu se definirati struktura raščlambe resursa, dostupnost resursa, cjenici resursa i različiti resursni kalendari kako bi pomogli u optimizaciji iskoristivosti resursa [3].

II. SOFTVER ZA UPRAVLJANJE PROJEKTIMA

Softver za upravljanje projektima su programi razvijeni da bi pomogli voditelju projekta u razvijanju plana upravljanja projektom, dodjeljivanju resursa za svaki zadatak, praćenje progresa projekta, upravljanje budžetom kao i u analiziranju radnog opterećenja članova projektnog tima.

Softverski alati za upravljanje projektima mogu se klasificirati na više načina:

1. Prema vrsti :
   - *Desk-top* softver koji se instalira na računalu korisnika
   - Mrežni softver koji se instalira na poslužitelju i može biti dostupan korisnicima s njihovih računala. Može se implementirati i kao web aplikacija kojoj se pristupa pomoću web preglednika korisnika.
2. Prema cijeni:
   - *Open-source* (besplatni) softver
   - Komercijalni softver
3. Prema mogućnostima [4]:
   - Razina 1. – softverski alati koji omogućavaju planiranje jednog projekta, a karakterizira ih jednostavnost, lakoća korištenja i ograničena analiza podataka. Osim toga odstupanje od početnog plana zahtjeva promjenu svih podataka i ponovno upisivanje u računalo.
   - Razina 2 – alati koji omogućavaju planiranje jednog projekta, i pružaju pomoć voditeljima projekta u planiranju, praćenju i izvještavanju o projektu. Ovi alati omogućavaju složenu analizu projekta, napretka i reviziju plana temeljenu na stvarnim performansama. Osim toga pružaju poluautomatsku kontrola projekta.
   - Razina 3 – alati koji omogućavaju planiranje, nadzor i kontrolu više projekata koristeći zajedničku bazu podataka. Istovremeno pružaju mogućnost sofisticiranog nadzora i izvještavanja između različitih projekata.

Primjeri mogućnosti današnjih softverskih paketa za upravljanje projektima su sljedeći [4]:

- Planiranje, praćenje i nadgledanje projektnih zadataka, resursa i troškova.
- Izvještaji o projektu. Primjeri izvještaja su: analiza kritičkog puta, analiza ostvarene vrijednosti, budžetirani trošak obavljenog rada, budžetirani trošak planiranog rada, usporedba planiranog i stvarnog rashoda i drugi.
- Projektni kalendar koji omogućava korisniku definiranje raspoloživog radnog tjedna na osnovu aktualnih podataka (vikendi, praznici, godišnji odmori).
- "Što ako" analiza. Na osnovu raspoloživih podataka softver izvršava usporednu analizu i prikazuje podatke dvije verzije projektnog plana u tabličnom ili grafičkom obliku s ciljem brze i jednostavne usporedbe.
- Analiza više projekata gdje softver pomoću zajedničke baze podataka za više projekata omogućava istovremenu analizu i izvještavanje o više projekata.

Primjeri softvera za upravljanje projektima su sljedeći:

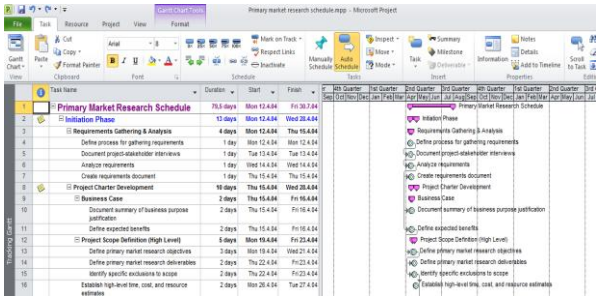
- komercijalni softver: Primavera P6, Microsoft Project;
- besplatni softver: Gantt project, Open Workbench i OpenProj.

A. *Microsoft Project*

Microsoft Project (MS Project) je najčešće korišteni softver za upravljanje projektima. MS Project je bila treća po redu aplikacija koju je Microsoft izradio za Windows operacijski sustav ali nikada nije bio dio MS Office paketa [5].

Posljednja verzija softvera je MS Project 2013 i izdana je krajem 2012. godine.

Verzija MS Project 2010 (slika 1.) je uz verziju MS Project 2007 najrašireniji i izrađuje se u dvije verzije: Standard i Professional. MS Project je vrlo koristan softverski alat za voditelje projekata. On pomaže voditeljima projekata planirati projekt, uvesti zadatke i resurse, te pratiti rokove i troškove projekta. MS Project pomaže pri unosu informacija koje su potrebne za izradu vremenskog rasporeda, izračun troškova projekta i praćenje promjena u projektu, te u sagledavanju projektnog plana i kontrolnih točaka. MS Project je



Slika1. Sučelje softvera MS Project 2010

pogotovo efikasan u fazi izvođenja kada dođe do promjena i potrebe za uspoređivanjem postojećeg stanja s planiranim. Korisničko sučelje većim dijelom je poznato korisnicima MS Office paketa (na primjer aplikacija MS Word i MS Excel). Naredbe su organizirane u menijima, a one češće korištene lako su dostupne preko trake s alatima.

Microsoft izrađuje i MS Project Server softver. MS Project Server softver se instalira na poslužitelju i kao osnovu koristi Microsoft SharePoint. Za pristup korisnika s klijent računala MS Project Server podržava MS Project softver kao klijent ili web preglednik s Project Web App (PWA) komponentom [6].

III. OPENPROJ

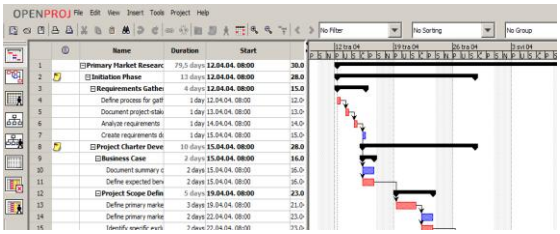
OpenProj softver je razvijen od firme Projity i predstavlja besplatnu alternativu za Microsoft Project. Od svoje prve verzije softvera u siječnju 2008. godine, softverski alat OpenProj vrlo brzo je postao jedan od najpopularnijih open-source rješenja za upravljanje projektima. OpenProj se bazira na java platformi što mu omogućava rad na raznim operacijskim sustavima. Trenutno posljednja izdana verzija ovog softvera je OpenProj 1.4. [7].

OpenProj blisko oponaša MS Project. OpenProj ima slično korisničko sučelje (slika 2), kao i sličan pristup izgradnji plana projekta. Izbornici i stupci su slični kao i u MS Project softveru tako da korisnik ne bi trebao imati problema prilikom prilagođavanja.

A. Usporedba MS Project i OpenProj softvera

Oba softverska paketa Microsoft Project i OpenProj su vrlo efikasne i moćne aplikacije za upravljanje projektima. Zajedničke su im sljedeće karakteristike:

- Gantov dijagram: grafički prikaz informacija vezanih uz kalendar. U tipičnom stupčanom grafikonu planirane aktivnosti ili komponente WBS-a popisane su lijevo od prikaza, datumi su prikazani na vrhu a trajanje aktivnosti je prikazano vodoravnim stupovima.
- PERT dijagram: grafički prikaz rasporeda zadataka u projektu. Prikazuje sekvencu zadataka i koji zadaci mogu biti izvršavani simultano, kao i sve zadatke na kritičkom putu koji moraju biti izvršeni na vrijeme tako da projekt bude završen u predodređenom vremenu.
- Dijagram resursa: lista raspoloživih resursa s odgovarajućim detaljima. Resursi koji se nalaze u listi mogu biti dodijeljeni u Gantov dijagram.
- WBS ili strukturna raščlamba poslova: hijerarhijska raščlamba posla koji projektni tim treba izvršiti, orijentirana na isporuke, kako bi se postigli projektni ciljevi i stvorile tražene isporuke. Organizira i definira ukupni projektni opseg. Svaka niža razina predstavlja detaljniju definiciju projektnoga rada. WBS se raščlanjuje na radne pakete. Hijerarhijska orijentacija prema isporukama uključuje i unutrašnje i vanjske isporuke.



Slika 2. Sučelje softvera OpenProj 1.4

- RBS ili strukturna raščlamba resursa: hijerarhijska strukture resursa prikazana prema kategoriji resursa i vrsti resursa koji se koristi u rasporedu raspodjele resursa te pri izradi rasporeda ograničenih resursa i koja se može koristiti za određivanje i analizu raspodjele ljudskih resursa na projektu.
- Generiranje izvještaja: oba softvera (Microsoft Project i OpenProj) omogućavaju različite vrste izvještaja, kao na primjer izvještaj o izvršavanju zadataka i odgovarajućim resursima.

Neke od prednosti MS Project softvera su sljedeće:

- Dokumentacija i pomoć: MS Project ima izuzetno dobru *online* i *offline* dokumentaciju i pomoć, tako da je učenje relativno jednostavno.
- Prijenos u PDF format: moguć je prijenos Microsoft Project datoteka u PDF format, kao i u ostale formate dokumenata MS Office paketa softvera.
- Java nije potrebna: nije potrebno instalirati Java softver da bi se koristio MS Project softver, kao što je to potrebno za OpenProj.

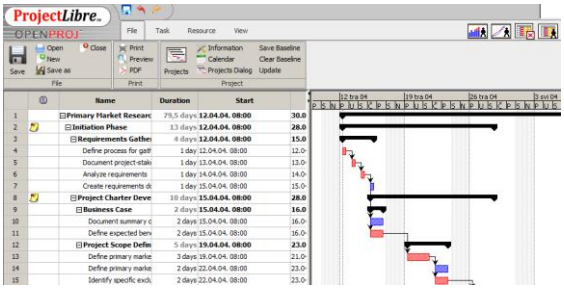
Neke od prednosti OpenProj softvera su sljedeće:

- Cijena: OpenProj je besplatan softver za upravljanje projektima.
- Prilagodljivost: OpenProj ima mogućnost otvaranja MS Project datoteka i spremanja u svoj format datoteka ili u format Excel softvera.
- Potrebni resursi: OpenProj u odnosu na MS Project zahtjeva mnogo manje resursa računala za instalaciju i izvršavanje programa.
- Neovisnost o operacijskom sustavu: OpenProj softver se izvršava na Java platformi i zbog toga može biti instaliran na različitim operacijskim sustavima (Windows, Linux, Mac OS /OS X)

Može se reći da je najvažnija razlika između Microsoft Project i OpenProj softvera je cijena.

B. Daljnji razvoj OpenProj softvera - ProjectLibre

Krajem 2008. godine firmu Projity, koja je razvila OpenProj softver, kupila je firma Serena Software. Nakon preuzimanja, Serena Software je prestala s daljnjim razvojem OpenProj softvera [7].



Slika 3. Sučelje softvera ProjectLibre

Tokom 2012. godine osnivači firme Projity (koji su istovremeno i programeri OpenProj softvera) objavili su da su daljnji razvoj OpenProj softvera, ali pod novim imenom: ProjectLibre. Oni su napravili takozvani „project fork", odnosno kopirali su izvorni kod OpenProj softvera i nastavili s njegovim razvojem kreirajući novi softver (što je dozvoljeno za programere izvornog open-source softvera) kojeg su nazvali ProjectLibre.

Prvotni cilj s ProjectLibre softverom bilo je kreiranje poslužiteljskog ProjectLibre Project Server softvera kao alternativno rješenje i zamjena za Microsoft Project Server. Zbog toga što je potreban i desk-top softver, a OpenProj u međuvremenu nije ažuriran, napravljena je nova verzija OpenProj softvera (verzija 1.5.3, nazvana i ProjectLibre) [8].

ProjectLibre dodaje nove mogućnosti softveru OpenProj1.4. [8]:

- potpuna kompatibilnost s Microsoft Project 2010 softverom (prijenos podataka u i iz MS Project 2010 datoteka),
- poboljšano korisničko sučelje softvera (slika 3),
- mogućnost štampanja (veliki nedostatak OpenProj 1.4. softvera),
- prijenos podataka u PDF format (bez ograničenja),
- rješavanje nekoliko softverskih grešaka u OpenProj 1.4.

C. Primjena OpenProj (ProjectLibre) softvera u edukaciji

OpenProj softver (i njegova novija verzija ProjectLibre) može biti korišten od strane studenata tokom nastave i vježbi iz predmeta Upravljanje projektima na visokoškolskom IT stručnom studiju.

Prednost OpenProj softvera što je besplatan omogućava studentima jednostavnu nabavu i instalaciju softvera. Tako studenti na svojim kućnim računalima mogu vježbati rad na softveru za upravljanje projektima i izrađivati seminare i ostale zadatke vezane uz predmet Upravljanje projektima.

Isto tako OpenProj softver ima potrebu za manjim resursima računala, tako da ga studenti za vježbanje mogu instalirati na starijim računalima (manja radna memorija, sporiji CPU).

Osim toga OpenProj softver podržava različite operacijske sustave, pa studenti mogu koristiti OpenProj na Linux i Mac OS (OS X) platformama i tako proširiti svoje znanje iz informacijskih tehnologija koje studiraju.

Jedna od najvećih prednosti MS Project softvera je dostupnost edukacijskih materijala i dokumentacija. Tokom predavanja i vježbi iz predmeta Upravljanje projektima studenti mogu dobiti osnovna znanja o OpenProj softveru i tako može biti nadoknađen nedostatak edukacijskih materijala za taj softver.

MS Project i OpenProj su dva vrlo slična softvera za upravljanje projektima. Kada se prilagodi na sučelje i nauče osnovne funkcije jednog softvera, vrlo je jednostavno korištenje drugog softvera. Uz vrlo mali napor i u kratko vrijeme se može prijeći u korištenju iz jednog u drugi softver za upravljanje projektima.

Prijenos podataka između OpenProj i MS Project softvera je moguć. OpenProj softver ima mogućnost otvaranja datoteka MS Project softvera (s ekstenzijama .MPP i .MPT) i tako omogućava pregledavanje MS Project podataka i daljnji rad na njima u OpenProj softveru. Isto tako nakon rada u OpenProj softveru podaci mogu biti spremljeni u XML formatu i naknadno uneseni u MS Project (i ostale programe koji imaju mogućnost unosa XML formata).

IV. ZAKLJUČAK

Današnji softverski alati za upravljanje projektima pružaju izuzetnu pomoć voditelju projekta pri praćenju i kontroliranju projekta. Ipak da bi se koristio softver za upravljanje projektima, voditelj projekta mora posjedovati potrebna znanja i vještine u upravljanju projektima.

MS Project je najrašireniji softver za upravljanje projektima zbog svojih složenih mogućnosti, jednostavne uporabe i dostupne dokumentacije. OpenProj je open-source besplatan softver i vrlo dobra alternativa MS Project softveru. OpenProj ima vrlo slično sučelje i većinu istih mogućnosti kao i MS Project softver. Iako postoje nedostaci OpenProj softvera (nedostatak dokumentacije, manjkavo održavanje i razvoj softvera), zbog svojih prednosti (cijena, mogućnosti), OpenProj softver (ili njegova novija verzija ProjectLibre) može biti efikasno korišten od strane studenata tokom nastave i vježbi iz predmeta Upravljanje projektima na visokoškolskom IT stručnom studiju. Softver OpenProj isto tako može biti besplatno instaliran na kućnim računalima studenata, tako da studenti imaju mogućnost detaljnijeg upoznavanja naprednog softvera za upravljanje projektima.

LITERATURA

[1] Charvat, J.: „Project Management Methodologies: Selecting, Implementing, and Supporting Methodologies and Processes for Projects", Hoboken, John Wiley & Sons, Inc, New York, 2003.

[2] Mignerat M., Rivard, S: „The institutionalization of information system project management practices, Information and Organization", Volume 22, Issue 2, April 2012,

[3] PMI: „A Guide to the Project Management Body of Knowledge (PMBOK Guide) „, Project Management Institute, Fourth edition, 2008.

[4] Kerzner, H.: Project Management: „A Systems Approach to Planning, Scheduling, and Controlling“, Eighth Edition, John Wiley & Sons, New York, 2003.

[5] http://en.wikipedia.org/wiki/MS_project, učitano 25.01.2013.

[6] http://www.microsoft.com/project/en-us/Preview/, učitano 29.01.2013.

[7] http://en.wikipedia.org/wiki/OpenProj, učitano 30.01.2013.

[8] http://www.projectlibre.org/, učitano 30.01.2013.

ZAŠTITA I SIGURNOST DJECE NA INTERNETU

Maja Kosovec*, C. Mario Dumančić **
* Osnovna škola „Remete“, Remete 99a, Zagreb, Hrvatska
** Učiteljski fakultet u Zagrebu, Savska cesta 77, Zagreb, Hrvatska
maja.kosovec@gmail.com, mario.dumancic@ufzg.hr

Sažetak - Internet, kao jedan od najrasprostranjenijih medija, postao je dio svakodnevice i djece i odraslih. Njegove brojne mogućnosti čine ga zanimljivim, ali i potencijalno opasnim. Budući da je riječ o mediju gdje je dopuštena potpuna anonimnost, prostora za nezakonite radnje je mnogo. Neprestana edukacija odraslih, djece i mladih preduvjet je za sigurno odrastanje djece u svijetu informacijsko-komunikacijske tehnologije, u kojem mediji postaju dio svakodnevnice u obrazovanju i odrastanju. Ovim radom se željelo ustanoviti koliko su djeca osnovnoškolske dobi upoznata s opasnostima na internetu. Pritom su kao čimbenici narušavanja sigurnosti uzeta određena rizična ponašanja djece te izloženost rizicima pri najučestalijim aktivnostima na internetu. Prilikom procjene rizičnog ponašanja djece naglasak je stavljen na društvenu mrežu Facebook zbog njezine popularnosti među učenicima. U radu je prikazano istraživanje na učenicima osnovnih škola u Zagrebu i izvan njega, koji pohađaju 5. i 8. razred.

I. UVOD

Mogućnosti interneta danas su neograničene. Počevši od razmjene informacija putem elektroničkih poruka, preko pretraživanja web-a, poput velike knjižnice podataka, sve do ostvarivanja socijalnih kontakata i stvaranja novih poznanstava putem društvenih mreža te e-learninga kao novog načina učenja pomoću računala, odnosno interneta. Jedan od prvih ciljeva interneta bio je ostvariti brzu komunikaciju između dva ili više računala. Danas možemo reći da je to bila glavna prednost interneta koja je s vremenom evoluirala u brzu komunikaciju između korisnika. Uz mnoge pozitivne strane, internet sa sobom nosi i opasnosti. Neželjene radnje na internetu nazivaju se internetski kriminal [1]. Pritom se razlikuju neželjene radnje kojima je cilj uništiti informacijsko-komunikacijsku tehnologiju i one kojima je informacijsko- komunikacijska tehnologija sredstvo za napad korisnika. Opasnosti za korisnike su cyberbullying [2], pornografija i pedofilija [3] te ovisnost o internetu [4]. Najugroženiji i najosjetljiviji korisnici interneta su dakako djeca. Da bismo ih uspjeli zaštiti i njihove aktivnost učiniti što sigurnijima, potrebno ih je educirati o opasnostima koje im prijete. Pritom je važno odrediti koje su njihove najčešće aktivnosti na internetu i gdje su najugroženiji. U posljednjih pet godina društvene mreže, kao što je Facebook, postale su glavni način zabave i svrha korištenja interneta kod djece i mladih. Zanimanje djece je veliko, a svijest o opasnostima još uvijek dosta mala. Nedovoljno razvijene vještine donošenja odluka, impulzivne reakcije te naivnost djece osnovnoškolske dobi dovode u pitanje sve prednosti koje nam pružaju društvene mreže. Učiniti internet sigurnim okruženjem osnovna je zadaća svih osoba koje sudjeluju u odrastanju djece. Na osnovu ovih promišljanja postavili smo cilj i problem našeg istraživanja.

II. METODOLOGIJA ISTRAŽIVANJA

A. Cilj, problemi i hipoteze

Cilj ovog istraživanja bio je utvrditi koliko su djeca osnovnoškolske dobi sigurna i zaštićena na internetu. U središte istraživanja stavljena su rizična ponašanja djece na društvenoj mreži Facebook i izloženost rizicima pri najučestalijim aktivnostima na internetu.

Istraživanjem smo željeli ispitati slijedeći problemi: učestalost i svrha korištenja internetom, te mjesto s kojeg djeca odlaze na internet, rizična ponašanja djece na društvenoj mreži Facebook, dosadašnja iskustva koja su potencijalno ugrozila sigurnost djece na internetu, djetetovo poznavanje privatnosti na internetu i informiranost o opasnostima na internetu te kome bi se učenici povjerili u slučaju uznemiravanja na internetu. Uz spomenute probleme željele su se utvrditi i razlike u rizičnom ponašanju djece s obzirom na dob i spol te razlike u informiranosti djece o opasnostima na internetu s obzirom na školu koju pohađaju.[1]

Na osnovu cilja i problema koje smo si postavili kreirane su osnovne tri hipoteze rada:

Hipoteza 1: Postoji statistički značajna razlika u sigurnosti djece na internetu s obzirom na dob.

Hipoteza 2: Postoji statistički značajna razlika u sigurnosti djece na internetu s obzirom na spol.

Hipoteza 3: Postoji statistički značajna razlika u informiranju djece o opasnostima na internetu s obzirom na školu koju pohađaju.

[1] Nezavisne varijable u ovom istraživanju su dob i spol djeteta te škola koju pohađa dok su zavisne varijable sigurnost djece na internetu i informiranost o opasnostima na internetu.

B. Uzorak ispitanika, instrumenti i metode obrade podataka

Uzorak ispitanika sastojao se od 205 učenika osnovnih škola, 100 dječaka i 105 djevojčica. Istraživanjem su obuhvaćeni učenici 5. i 8. razreda, starosti 11 i 14 godina. Ispitano je 116 učenika 5. razreda (54 dječaka i 62 djevojčice) i 89 učenika 8. razreda (46 dječaka i 43 djevojčice).

U istraživanju su sudjelovale Osnovna škola Otok iz Zagreba, Osnovna škola Mato Lovrak iz Zagreba, Osnovna škola Vugrovec- Kašina iz Kašine te III. osnovna škola Čakovec iz Čakovca.

U Osnovnoj školi Otok ispitano je ukupno 80 učenika, od toga 37 učenika 5. razreda i 43 učenika 8. razreda.

U Osnovnoj školi Mato Lovrak iz Zagreba ispitano je ukupno 48 učenika, od toga 26 učenika 5. razreda i 22 učenika 8. razreda.

U Osnovnoj školi Vugrovec- Kašina iz Kašine ispitano je ukupno 38 učenika, od toga 14 učenika 5. razreda i 24 učenika 8. razreda.

U III. osnovnoj školi Čakovec ispitano je 39 učenika, svi su pohađali 5. razred.

Od ukupnog broja ispitanika njih 182 pohađa izbornu nastavu informatike, dok svega 23 učenika ne pohađa izbornu nastavu informatike.

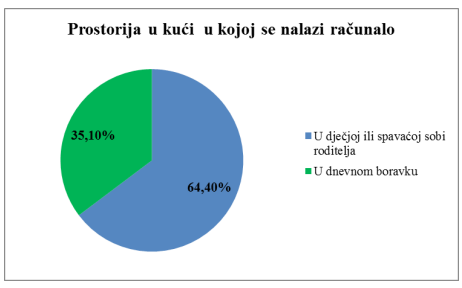
U istraživanju je korišten anoniman anketni upitnik za učenike, po principu papir- olovka. Upitnik se sastojao od 20 pitanja zatvorenog tipa. U određenim pitanjima bilo je moguće zaokružiti više od jednog odgovora. Također, određena pitanja su uz ponuđene odgovore nudila i mogućnosti samostalnog upisivanja odgovora (pitanja otvorenog tipa).

U obradi podataka korištena je deskriptivna analiza podataka i t-test za nezavisne uzorke, kojim su utvrđene razlike između aritmetičkih sredina. Razina statističke značajnosti je postavljena na $p<0,05$.

III. REZULTATI I RASPRAVA

Istraživanje je pokazalo da većina učenika (96,6%) obuhvaćenih ovim istraživanjem kod kuće posjeduje računalo s pristupom internetu. U skladu s tim, kao najučestalije mjesto pristupa internetu učenici su naveli kućno računalo (55,6%), a odmah iza toga mobilni telefon (22,9%).

Grafički prikaz I. Mjesto držanja računala u kući



Kućno računalo se najčešće nalazi u dječjoj sobi (64,4%) ili u spavaćim sobama roditelja, braće i sestara, što predstavlja potencijalnu opasnost, budući da je dijete u svojoj sobi uglavnom van roditeljskog pogleda i nadzora. Ovaj problem javlja se češće kod starijih učenika. Dobivena je statistički značajna razlika s obzirom na dob učenika ($p<0,05$) i prostoriju gdje se nalazi računalo (Tablica I). Manji broj učenika (35,1%) naveo je da se računalo nalazi u dnevnom boravku, koji je puno sigurnije mjesto za učenikov boravak na internetu. Dnevni je boravak mjesto okupljanja obitelji te je veća vjerojatnost da će učenik biti pod nadzorom odrasle osobe za vrijeme boravka na internetu. Svakodnevno se internetom koristi čak 69,8% učenika, a svega 1,5% koristi internet manje od jednom mjesečno. Od ukupnog broja učenika njih 46,3% provodi na internetu 1-2 sata dnevno. Ostatak učenika provodi manje od jednog sata (21,5%), odnosno 2-3 sata (21%).

Tablica I. T-test za nezavisne uzorke. Razlika u prostoriji kod kuće u kojoj se nalazi računalo s obzirom na dob učenika.

		Levene's Test for		t-test for Equality of Means						
									95% Confidence	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Prostorija kod kuće gdje se nalazi računalo	Equal variances assumed	4,914	,028	2,605	203	,010	,25823	,09913	,06277	,45369
	Equal variances not assumed			2,673	202,016	,008	,25823	,09662	,06772	,44875

Ispitano je i poznavanje privatnih podataka te informiranost o zaštiti sebe i svojih podataka na internetu. Dobiveni su slijedeći podatci. Najveći broj učenika privatnim podatkom smatra korisničko ime i lozinku (90,7%), zatim slijedi adresa stanovanja (83,4%), broj mobitela (82%) i broj telefona (81%), vlastite fotografije (65,9%), e-mail adresa (52,7%), vlastito ime i prezime (43,9%) te ime škole (35,1%). Postoji statistički značajna razlika ($p<0,05$) s obzirom na dob u shvaćanju imena i prezimena te e-mail adrese kao osobnog podatka (Tablica II). Više mlađih učenika (74,4% za ime i prezime, 66,7% za e-mail), nego starijih (25,6% za ime i prezime, 33,3% za e-mail) smatra ta dva podatka privatnim podatcima. Važno je napomenuti kako se svi navedeni podatci smatraju privatnim podatcima te njihovim objavljivanjem učenici ugrožavaju svoju sigurnost. Prema ovdje dobivenim podatcima na se taj način još uvijek dio učenika iz ovog istraživanja izlaže rizicima. Zbog toga je potrebno proraditi na svijesti o privatnim podatcima. Pri čemu je potrebna suradnja škole i roditelja.

Profil na Facebook-u posjeduje 83,4% učenika, češće stariji učenici od mlađih učenika. T-testom za nezavisne uzorke utvrđena je statistički značajna razlika ($p<0,05$) u posjedovanju Facebook profila s obzirom na dob (Tablica III).

Ovaj podatak možemo pripisati sazrijevanju učenika, njihovoj radoznalosti i potrebi za istraživanjem, za sklapanjem prijateljstava u virtualnom svijetu gdje se mogu prikazati onako kako žele.

Stoga ne čudi da je najčešća internet aktivnost 55,1% učenika posjećivanje društvenih mreža poput Facebook-a i Twitter-a.

Tablica II. T-test za nezavisne uzorke. Razlika u procjeni privatnih podataka s obzirom na dob.

		Levene's Test for Equality of		t-test for Equality of Means						
									95% Confidence Interval of the	
		F	Sig.	t	df	Sig. (2-tailed)	Difference	Error Difference	Lower	Upper
Osobni podatak-ime i prezime	Equal variance	24,023	,000	-4,792	203	,000	-,31916	,06661	-,45049	-,18783
	Equal variance			-4,868	198,684	,000	-,31916	,06557	-,44846	-,18986
Osobni podatak-e-mail adresa	Equal variance	,511	,476	-3,131	203	,002	-,21620	,06905	-,35235	-,08004
	Equal variance			-3,126	188,271	,002	-,21620	,06917	-,35264	-,07975

Ovom aktivnošću učenici „izgrađuju“ svoj identitet, eksperimentiraju te započinju socijalizaciju u drugačijem prostoru od onog na koji su navikli do tada. Roditelji ispitanih učenika su u velikom broju upoznati i upućeni u aktivnosti svoje djece na društvenoj mreži, što je prvi korak u prevenciji opasnosti na internetu. Unatoč tome, ponašanje učenika na društvenoj mreži nije posve odgovorno. Od 83,4% učenika koji imaju Facebook profil njih 21% ne poznaje sve svoje prijatelje s Facebooka.

Tablica III. T- test za nezavisne uzorke. Razlika u posjedovanju Facebook profila s obzirom na dob.

		Levene's Test for Equality of		t-test for Equality of Means						
									95% Confidence Interval of the	
		F	Sig.	t	df	Sig. (2-tailed)	Difference	Error Difference	Lower	Upper
Imaš li Facebook profil?	Equal variance	108,481	,000	4,232	203	,000	,21368	,05049	,11412	,31323
	Equal variance			4,603	172,803	,000	,21368	,04642	,12205	,30530

Od ukupnog postotka učenika koji ne poznaju sve svoje prijatelje s Facebooka većinu čine stariji učenici (81,4%), što je dokazano i t-testom za nezavisne uzorke ($p<0,05$). Dakle, postoji i statistički značajna razlika u nepoznavanju prijatelja s Facebooka s obzirom na dob učenika (Tablica IV). Ovaj podatak daje nam informaciju da se stariji učenici rizičnije i manje odgovorno ponašaju na društvenim mrežama te da im društvena mreža služi i za upoznavanje novih osoba, a ne samo za komunikaciju s osobama koje poznaju.

Većina učenika upoznata je sa sigurnosnim postavkama na društvenoj mreži. Zaštićene profile ima 68,3% učenika, a tajnu lozinku i korisničko ime njih 76,6%. Zahtjev za prijateljstvom od nepoznate osobe 35,1% učenika obriše bez gledanja tuđeg profila, dok 27,8% prvo prihvati zaahtjev i pogleda profil osobe, a zatim ga izbrišu. Jedan dio učenika (14,1%) prihvaća zahtjev za prijateljstvom od nepoznate osobe te se tom nepromišljenom radnjom dovode u opasnost.

Tablica IV. T-test za nezavisne uzorke. Razlika u nepoznavanju prijatelja na Facebooku s obzirom na dob.

		Levene's Test for Equality of		t-test for Equality of Means						
									95% Confidence Interval of the	
		F	Sig.	t	df	Sig. (2-tailed)	Difference	Error Difference	Lower	Upper
Poznaješ li sve prijatelje s FB?	Equal variance	132,150	,000	-5,134	169	,000	-,31874	,06208	-,44130	-,19619
	Equal variance			-5,120	135,873	,000	-,31874	,06226	-,44186	-,19562

Istraživanjem se između ostalog željelo ustanoviti koriste li učenici web kameru pri svojim internet aktivnostima te dovode li se na taj način u opasnost. Polovica učenika (50,2%) obuhvaćenih istraživanjem posjeduju web kameru. U najvećoj mjeri (32,2%) koriste ju za video-poziv s prijateljima iz škole ili ulice, zatim (23,9%) za video-poziv s roditeljima ili rođacima, a vrlo mali dio (5,4%), no ne i zanemariv, za video-poziv s prijateljima koje upozna na internetu. Prema tome, valja obratiti pozornost na to da postoje učenici koji nesavjesno koriste web kameru, iako u znatno manjem broju od onih koji ju savjesno koriste.

Djevojčice, kao ranjiviji spol, su češće žrtve uznemiravanja na internetu, nego dječaci. Zbog određenih sadržaja na internetu 17,1% učenika se osjećalo uznemireno, od čega je veći postotak djevojčica (80%), naspram dječaka (20%). Dobivena je i statistički značajna razlika ($p<0,05$) s obzirom na spol.

Od 19,7% učenika koji su primili poziv za sastanak s nepoznatom osobom, djevojčice čine većinu (71,8%) te je i tu utvrđena i statistički značajna razlika ($p<0,05$) s obzirom na spol (Tablica V). Nadalje, njih 12,7% je pristalo na takav sastanak, i to većina starijih učenika (84,6%). Utvrđena je i statistički značajna razlika ($p<0,05$) s obzirom na dob (Tablica VI). Ovaj podatak ponovno možemo pripisati potrebi za dokazivanjem i eksperimentiranjem starijih učenika te njihovoj potrebi za upoznavanjem novih ljudi. Mlađi učenici su pažljiviji i imaju veći osjećaj straha kada je u pitanju nepoznata osoba.

Tablica V. T-test za nezavisne uzorke. Razlika u dobivanju poziva za sastanak s nepoznatom osobom s obzirom na spol.

		Levene's Test for Equality of		t-test for Equality of Means						
									95% Confidence Interval of the	
		F	Sig.	t	df	Sig. (2-tailed)	Difference	Error Difference	Lower	Upper
Poziv za sastanak od nepoznate osobe	Equal variance	25,861	,000	3,057	203	,003	,17667	,05779	,06272	,29061
	Equal variance			3,069	200,560	,002	,17667	,05756	,06316	,29018

Tablica VI. T-test za nezavisne uzorke. Razlika u sastajanju s nepoznatom osobom s obzirom na dob.

		Levene's Test for Equality of		t-test for Equality of Means						
									95% Confidence Interval of the	
		F	Sig.	t	df	Sig. (2-tailed)	Difference	Error Difference	Lower	Upper
Susretanje s osobom koju ne poznaje	Equal variance	125,919	,000	4,759	203	,000	,21271	,04470	,12458	,30084
	Equal variance			4,338	112,137	,000	,21271	,04903	,11556	,30986

Od ukupnog broja ispitanika njih 54,1% primilo je zahtjev za prijateljstvom od nepoznatih ljudi, a 31,2% odgovorio je da su mu se na internetu obraćali neprimjerenim riječima. Nadalje, od ukupnog broja ispitanika njih 22,9% odgovorilo je da mu se na internetu otvorila stranica s pornografskim sadržajem, češće dječacima (61,7%), nego djevojčicama (38,3%). I u dosadašnjim istraživanjima dokazan je gore navedeni podatak [5].

Ovi ispitanici su najviše povjerenja pokazali u svoje roditelje. Najveći postotak učenika (48,3%) bi se u slučaju uznemiravanja na internetu obratio roditeljima, zatim slijede prijatelji (42%). Mali postotak učenika (8,3%), ali ne i zanemariv, se ne bi nikome obratio, dok bi se svega 3% učenika obratilo bratu/sestri ili baki, a niti jedan učenik se ne bi obratio učitelju/profesoru. Ovaj podatak govori nam da se u ovim školama još uvijek malo vremena posvećuje zaštiti i sigurnosti djece na internetu, da djeca nemaju povjerenja u učitelje i profesore te otvara puno prostora u budućem radu učitelja i profesora na temu sigurnosti i zaštite djece na internetu.

Ove potvrđuju i podatci koji govore da najveći broj učenka informacije o zaštiti i odgovornom ponašanju na internetu dobiva kod kuće od roditelja (44,4%). U školi na satu informatike tek 29,3%, odnosno na satu razrednika 7,3% učenika. Predavanje o sigurnosti i zaštiti na internetu odslušalo je svega 4,4% učenika. Jedan dio učenika (14,6%) nije uopće informiran o zaštiti i sigurnosti na internetu. Prilikom usporedbe ove 4 škole koje su sudjelovale u istraživanju utvrđeno je da najveći broj učenika koji su dobili informacije u školi na satu informatike pripada III. osnovnoj školi u Čakovcu (51,7%). Dokazana je i statistički značajna razlika ($p<0,05\%$) s obzirom na ostale škole iz istraživanja. Učenici ostalih škola, Otok, Mato Lovrak i Vugrovec-Kašina, uglavnom su dobili informacije kod kuće. Ovaj podatak je bio očekivan budući da je III. osnovna škola Čakovec aktivan sudionik Unicefovog projekta „Prekini lanac! - Zaustavimo elektroničko nasilje“ te sudjeluje u akcijama „Sigurniji internet za djecu“ koje svake godine organizira udruga Suradnici u učenju. U ovoj se školi s učenicima provode razne radionice na temu sigurnijeg korištenja interneta te i učenici sami izrađuju multimedijalne radove na istu temu. Učitelji, profesori i cjelokupni nastavni kadar ulažu vrijeme i trud kako bi učenike što više educirali o sigurnom korištenju interneta.

Učenici obuhvaćeni ovim istraživanjem pokazali su određen postotak rizičnog ponašanja u svim postavljenim elementima, s tim da je samo kod određenih elemenata ustanovljena statistički značajna razlika u odnosu na dob učenika, koja je dokazana t-testom za nezavisne uzorke ($p<0,05$). Riječ je o starijim učenicima koji u većem postotku imaju profil na društvenoj mreži Facebook te ne poznaju sve prijatelje na Facebook-u i češće od mlađih učenika pristaju na sastanak s nepoznatim ljudima.

U ostalim elementima rizičnog ponašanja, kao što su nezaštićen Facebook profil, otkrivanje korisničkog imena i lozinke te korištenje web kamere za video pozive s nepoznatim osobama, učenici nisu pokazali statistički značajnu razliku s obzirom na dob.

Kada je riječ o izloženosti rizicima u svim postavljenim elementima ustanovljen je određen postotak izloženosti riziku, a statistički značajna razlika s obzirom na dob utvrđena je samo u određenim elementima. Utvrđeno je da stariji učenici iz ovog uzorka ispitanika u većem postotku nego mlađi pristupaju internetu s računala koje se nalazi u njihovoj sobi, van roditeljskog nadzora, i ne smatraju svoje ime i prezime privatnim podatkom te ih češće od mlađih učenika dijele s drugima putem interneta. U ostalim elementima izloženosti riziku, kao što su primanje zahtjeva za prijateljstvom od nepoznatih ljudi, otvaranje stranica s pornografskim sadržajem te obraćanje neprimjerenim riječima, nije utvrđena statistički značajna razlika s obzirom na dob.

Budući da su statistički značajne razlike u odnosu na dob i spol dobivene samo u pojedinim elementima rizičnog ponašanja i izloženosti rizicima, Hipoteza 1: „Postoji statistički značajna razlika u sigurnosti djece na internetu s obzirom na dob“ može se tek djelomično potvrditi.

Nadalje, statistički značajna razlika s obzirom na spol nije dobivena niti u jednom elementu rizičnog ponašanja, ali je dobivena u određenim elementima koji se tiču izloženosti rizicima. Riječ je o osjećaju uznemirenosti zbog određenih sadržaja na internetu te pozivu za sastanak s nepoznatim ljudima, pri čemu su djevojčice češće primale pozive za sastanak s nepoznatim ljudima te se time i statistički u većem postotku od dječaka osjećale uznemirenima na internetu. Od ostalih čimbenika koji se tiču izloženosti riziku, među ovom skupinom ispitanika nije ustanovljena statistički značajna razlika s obzirom na spol. U skladu s tim i Hipoteza 2: „Postoji statistički značajna razlika u sigurnosti djece na internetu s obzirom na spol“ može se tek djelomice potvrditi. Ovime se otvara prostor za daljnja istraživanja na ovu temu koja bi trebala biti opširnija i obuhvatiti veći uzorak ispitanika.

T-testom za nezavisne uzorke u potpunosti je potvrđena Hipoteza 3: „Postoji statistički značajna razlika u informiranosti učenika s obzirom na školu koju pohađaju“ i može se sa sigurnošću prihvatiti. Statistički je dokazano da su učenici III. osnovne škole u Čakovcu u najvećem broju dobili informacije o sigurnosti i zaštiti na internetu u školi. U prilog tome ide i činjenica da III. osnovna škola u Čakovcu jedina od škola iz ovog uzorka aktivno provodi program zaštite djece na internetu.

IV. ZAKLJUČAK

Sigurnost i zaštita djece na internetu, kao aktualna tema i problem novog doba, pred sve odrasle osobe stavlja zadatak stalne edukacije i zanimanja za aktivnosti djece na internetu.

Naglasak u edukaciji potrebno je staviti na odgovorno korištenje društvenih mreža, osobito kada je riječ o sklapanju prijateljstava s nepoznatim ljudima i zaštiti osobnih podataka. Ovo je istraživanje i statistički dokazalo rizičnije ponašanje starijih učenika, i možemo zaključiti da je edukacija potrebna podjednako i jednima i drugima.

Ovo nam istraživanje ukazuje na potrebu rane edukacije i prevencije kako bi se smanjila rizična ponašanja djece na internetu. Dakako, u najkritičnijoj fazi djetetova sazrijevanja, pred kraj osnovnoškolskog razdoblja, u kojem dijete pronalazi sebe i više od svega voli eksperimentirati, potrebno je dodatno povećani nadzor i edukacija. Ono što bi trebalo naglasiti u vezi neodgovornog ponašanja starijih učenika je njihova potreba za dokazivanjem i predstavljanjem onakvim kakvi nisu pa stoga i taj čimbenik moramo uzeti u obzir prilikom ovog istraživanja. Podatci dobiveni ovim istraživanjem, zbog ograničenog uzorka, ne mogu se primjeniti na cjelokupnu populaciju.

Virtualni svijet nije ništa opasniji od stvarnog svijeta, ali ga čovjekovo olako shvaćanje može učiniti takvim. Sve mjere prevencije koje odrasli koriste kako bi zaštitili dijete u stvarnom svijetu, podjednako su potrebne i u virtualnom svijetu. Pritom je, uz roditeljsku brigu i zaštitu djece na internetu, vrlo važna i stručna edukacija djece koja bi se trebala provoditi u svim školama.

LITERATURA

[1] Đ. Težak, Internet- poslije oduševljenja. Zagreb, Hrvatska: Sveučilišna naknada, 2010.

[2] Z. Miliša, M. Tolić, N. Vertovšek, Nenad, Mladi- odgoj za medije- Priručnik za stjecanje medijskih kompetencija, M.E.P d.o.o., Zagreb, 2010.

[3] Kukec, Tomislav, Policija u ratu s internet pedofilima: S Facebooka ukrali 36 fotografije gole djece pa ih stavili na pornostranice, Jutarnji list, 18.veljače, str. 6-7, Zagreb, 2012.

[4] Eduvizija.hr, ovisnost o internetu. Dostupno na: http://www.eduvizija.hr/portal/sadrzaj/ovisnost-o-internetu. (Preuzeto:18.12.2912.)

[5] Istraživanje o korištenju Interneta, mobitela i drugih tehnologija, Poliklinika za zaštitu djece grada Zagreba. Dostupno na: http://www.poliklinika-djeca.hr/istrazivanja/istrazivanje-okoristenj u interneta mobitela-i-drugih-tehnologija/ (Preuzeto 15.11. 2012.)

Mogućnosti primjene informacijsko-komunikacijske tehnologije u primarnom obrazovanju

Marina Lukić*, dr.sc. Mario Dumančić**
* Osnovna Škola "Pušća", Zagorska 2, Donja Pušća
** Učiteljski fakultet u Zagrebu, Savska cesta 77., Zagreb, Hrvatska
lukicmarina23@gmail.com, mario.dumancic@ufzg.hr

Sažetak - Utjecaj informacijsko-komunikacijske tehnologije na cjelokupan ljudski život i rad postaje svakim danom sve veći. U današnje vrijeme upotreba informacijsko-komunikacijske tehnologije u nastavom procesu je nezaobilazna, stoga je važno da učitelji shvate važnost stalne edukacije za njenu primjenu. U ovom radu prikazane su mogućnosti primjene informacijsko-komunikacijske tehnologije u primarnom obrazovanju. Korištenje informacijsko-komunikacijske tehnologije konkretizirano je kroz nekoliko čimbenika - stručna osposobljenost učitelja za njenu primjenu, mogućnost i učestalost upotrebe informacijsko-komunikacijske tehnologije u nastavi te njena važnost za kvalitetu nastavnog procesa. Na uzorku od 117 učitelja koji rade u Zagrebu i ruralnim sredinama ispitivali su se stavovi učitelja o korištenju tehnologije u nastavi te učestalost korištenja određenih elemenata informacijsko-komunikacijske tehnologije u nastavi s obzirom na dob, radni staž ispitanika te sredinu u kojoj se nalazi škola.

I. UVOD

ICT[1] u današnjem svijetu ima velik utjecaj na život i rad ljudi. Postala je svakodnevnicom u svim područjima rada i djelovanja, pa tako i u obrazovanju te u skladu s tim treba oblikovati i školski sustav [1]. ICT u obrazovanju služi za stjecanje osnovnih vještina te za cjeloživotno učenje kojem se u današnje vrijeme pridaje mnogo pozornosti [2]. Njena primjena zahtijeva svjesnost o svrsi njezina korištenja, učiteljevu intelektualnu otvorenost, kreativnost u radu te spremnost na rizik [3]. Korištenjem ICT-a otvaraju se nove mogućnosti unutar odgojno-obrazovnog procesa. Tehnološki napredak, koji je usko vezan uz ICT, uvelike utječe na obrazovanje i znatno mijenja postojeći proces učenja i poučavanja. Nove obrazovne paradigme zahtijevaju i novu tehnologiju koja obuhvaća korištenje računala i svih njegovih mogućnosti u svrhu poboljšanja kvalitete odgojno-obrazovnog procesa [4].

II. METODOLOGIJA ISTRAŽIVANJA

A. *Cilj, problemi i hipoteze*

Cilj ovog istraživanja bio je utvrditi postoji li razlika u korištenju ICT-a u nastavnom procesu s obzirom na dob i staž anketiranih učitelja te s obzirom na sredinu u kojoj je škola.

Istraživanje obuhvaća i određene probleme i hipoteze koje je potrebno ispitati, a to su stavovi učitelja o osobnoj stučnoj osposobljenosti za korištenje ICT-a u nastavnom procesu, mogućnost upotrebe računala u nastavi te učestalost uporabe iste, učestalost korištenja računala za PowerPoint prezentacije, slušanje glazbe te za gledanje crtanih ili dokumentarnih filmova u nastavi, učestalost korištenja interneta u nastavi. Nadalje, ispitani su i stavovi učitelja o zanimljivosti nastave uz primjenu ICT-a, o važnosti korištenja ICT-a u nastavnom procesu, o važnosti osposobljavanja učiteljskog kadra za korištenje ICT-a te stavovi učitelja o uvođenju Informatike kao obveznog predmeta u primarno obrazovanje.

Nezavisne varijable u ovom istraživanju su dob i staž učitelja te sredina u kojoj se škola nalazi.
Zavisna varijabla je korištenje ICT-a u nastavnom procesu.

Postavljene su tri hipoteze:

Hipoteza 1: Korištenje ICT-a u nastavnom procesu povezano je s dobi ispitanika.

Hipoteza 2: Korištenje ICT-a u nastavnom procesu povezano je sa stažom ispitanika.

Hipoteza 3: Korištenje ICT-a u nastavnom procesu povezano je sa sredinom u kojoj je škola.

[1] Informacijsko-komunikacijska tehnologija

B. Uzorak ispitanika, instrumenti i metode obrade podataka

Uzorak ispitanika sastojao se od 117 učitelja osnovnih škola na području grada Zagreba i izvan njega. Od ukupnog broja, 3 ih je učitelja i 114 učiteljica.

U istraživanju je sudjelovalo 10 škola, od kojih je 6 u gradu Zagrebu (OŠ[2] Trnsko, OŠ Davorina Trstenjaka, OŠ Otok, OŠ Mate Lovraka, OŠ Tina Ujevića, OŠ Prečko), a 4 u ruralnim sredinama (OŠ Ljudevita Gaja Zaprešić, OŠ Antuna Augustinčića Zaprešić, OŠ Vugrovec-Kašina, OŠ Čakovec).

U OŠ Trnsko ispitano je 12 ispitanika, u OŠ Davorina Trstenjaka 16, u OŠ Otok ispitano ih je 10, a u OŠ Mate Lovraka 11. U OŠ Tina Ujevića sudjelovalo je 7 ispitanika, a u OŠ Prečko jedan više od prethodne škole, tj. njih 8.

Što se tiče ruralnih sredina u OŠ Ljudevita Gaja ispitano je 10 učitelja, dok je u OŠ Antuna Augustinčića sudjelovalo njih 11. U OŠ Vugrovec-Kašina anketu je ispunjavalo 17 učitelja, a u OŠ Čakovec njih 8.

Starosna dob ispitanika podijeljena je na 4 kategorije (<30, 30-40, 40-50 te 50-60> godina). 17 je ispitanika mlađe od 30 godina dok ih je 32 između 30 i 40 godina. Najviše je ispitanika starosne dobi između 40 i 50 godina, njih 44. 24 ispitana učitelja pripada četvrtoj kategoriji starosne dobi, od 50-60> godina.

Radni staž ispitanika u anketnom upitniku također je podijeljen u 4 kategorije (<1-10, 11-20, 21-30, 31-40> godina radnog staža). 32 ispitanika pripadaju prvoj kategoriji. Najviše učitelja, njih 36, ima između 11 i 20 godina radnog staža. 31 učitelj radi u školi između 21-30 godina, dok je najmanje onih koji rade u školi 31-40> godina, njih 18.

Istraživanje je provedeno kroz anonimni anketni upitnik za učitelje razredne nastave. Upitnik se sastojao od 16 pitanja zatvorenog tipa. U pitanjima se od ispitanika tražilo da zaokruže isključivo jedan odgovor, onaj koji najbolje izražava njihove stavove i mogućnosti o navedenim pitanjima.

U obradi podataka korištena je deskriptivna analiza podataka i t-test za nezavisne uzorke, kojim su utvrđene razlike između aritmetičkih sredina. Razina statističke značajnosti je postavljena na $p<0,05$.

III. REZULTATI I RASPRAVA

Na temelju provedenog istraživanja utvrđeno je kako se većina učitelja djelomično slaže s tvrdnjom da su dovoljno stručno osposobljeni za korištenje ICT-a u nastavnom procesu (39,3%). Slično je i sa stavovima ispitanika o važnosti osposobljavanja svih učitelja za korištenje ICT-a u nastavi. 41,9% učitelja se u potpunosti slaže s tom tvrdnjom čime pokazuju svoju svijest o važnosti ICT-a u odgojno - obrazovnom procesu. T-testom za nezavisne uzorke utvrđena je statistički značajna razlika u stavovima o osobnoj stručnoj osposobljenosti za korištenje ICT-a s obzirom na dob ispitanika (Tablica I i II). Crosstablicom utvrđeno je kako su mlađi učitelji sigurniji u vještine upotrebe ICT-a u nastavi, dok je kod starijih učitelja još uvijek prisutna nesigurnost pri njenom korištenju što je posljedica tradicionalizma i stare škole, ali i nedovoljno stečenih umijeća za korištenje iste.

Tablica I: T-test za nezavisne uzorke. Razlika u stavovima o osobnoj stručnoj osposobljenosti za korištenje ICT-a s obzirom na dob (<30 i 40-50 godina)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Osposobljenost za korištenje ICT-a u nastavi	Equal variances assumed	,588	,446	2,956	59	,004	,71257	,24105	,23024	1,19490
	Equal variances not			3,164	33,695	,003	,71257	,22524	,25467	1,17047

Tablica II: T-test za nezavisne uzorke. Razlika u stavovima o osobnoj stručnoj osposobljenosti za korištenje ICT-a s obzirom na dob (<30 i 50-60> godina)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Osposobljenost za korištenje ICT-a u nastavi	Equal variances assumed	,032	,860	3,999	39	,000	,98529	,24638	,48694	1,48365
	Equal variances not			4,037	35,730	,000	,98529	,24408	,49016	1,48043

Radni staž također ima utjecaja na stavove o dovoljnoj, osobnoj stručnoj osposobljenosti za primjenu ICT-a pa je tako i u ovome slučaju dobivena statistički značajna razlika s obzirom na učitelje koji imaju do 10 godina radnog staža i onih koji imaju između 31-40> godina iskustva (Tablica III i IV). Utvrđeno je kako su učitelji s manje radnog staža sigurniji u izvođenju nastave uz ICT od učitelja s više radnog staža.

Tablica III: T-test za nezavisne uzorke. Razlika u stavovima o osobnoj stručnoj osposobljenosti za korištenje ICT-a s obzirom na staž (<1-10 i 21-30 godina staža)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Osposobljenost za korištenje ICT-a u nasatvi	Equal variances assumed	3,935	,052	3,131	61	,003	,64214	,20508	,23206	1,05222
	Equal variances not			3,117	55,419	,003	,64214	,20603	,22932	1,05496

Tablica IV: T-test za nezavisne uzorke. Razlika u stavovima o osobnoj stručnoj osposobljenosti za korištenje ICT-a s obzirom na staž (<1-10 i 31-40> godina staža)

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Osposobljenost za korištenje ICT-a u nastavi	Equal variances assumed	,649	,425	3,739	48	,000	,81597	,21821	,37723	1,25472
	Equal variances not			3,552	30,349	,001	,81597	,22974	,34700	1,28495

[2] Osnovna škola

Kada je riječ o mogućnosti i učestalosti upotrebe ICT-a u nastavi, najviše učitelja odgovorilo je kako ponekad ima mogućnot korištenja računala u nastavi (28,2%), dok njih 22,2% često ima mogućnost primjeniti računalo u nastavnom procesu. Skoro podjednak je broj onih koji rijetko imaju tu mogućnost (23,1%). 12% učitelja nema uopće mogućnost iskoristiti prednosti računala u nastavi. Zanimljivo je kako kumulativno 85,5% učitelja ima mogućnost koristiti računalo, a 29,1% učitelja se izjasnilo kako rijetko koriste računalo, iako imaju tu mogućnost.
Samo 0,9% učitelja uvijek upotrebljava računalo u nastavi, kada im se omogući rad s njime. Ovi postoci su zapravo poražavajući jer je očigledno da učitelji, iako imaju mogućnost unaprijediti nastavni proces i učiniti ga zanimljivijim uz računalo, radije ostaju pri ploči i kredi te suhoparnom predavanju u kojem su učenici pasivni subjekti. T-testom utvrđeno je da postoji statistički značajna razlika u učestalosti upotebe računala s obzirom na dob učitelja (između onih koji imaju manje od 30 godina i onih koji imaju između 50-60 i više godina) (Tablica V). To nije iznenađujuće s obzirom da su to dvije krajnje točke, mladi učitelji (<30 godina) koji imaju više znanja o uporabi računala i kontakta s njime te stariji učitelji (50-60> godina) koji nisu odrastali s tehnologijom te zbog neznanja i neprakticiranja imaju averziju prema njegovoj uporabi.

Tablica V: T-test za nezavisne uzorke. Razlika u učestalosti upotrebe računala u nastavi s obzirom na dob ispitanika(<30 i 50-60> godina)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Učestalost upotrebe računala u nastavi	Equal variances	1,862	,180	2,380	39	,022	,73529	,30901	,11026	1,36033
	Equal variances			2,482	38,552	,018	,73529	,29624	,13587	1,33472

Provedenim istraživanjem utvrđeno je da većina učitelja u nastavi ponekad koristi PowerPoint prezentacije (29,9%), dok je čak 25,6% posto onih koji to nikada ne čine. To nisu očekivani rezultati s obzirom da živimo u vremenu u kojem je tehnologija svepristuna i služi kao olakšavajući faktor u ljudskim djelatnostima. PowerPoint u nastavi pruža mnogo mogućnosti i učiteljima i učenicima, ali učitelji, nažalost, još uvijek toga nisu dovoljno svjesni. Kada je u pitanju korištenje računala za slušanje glazbe na nastavi Glazbene kulture dobiveni rezultati su poražavajući. Više od polovice učitelja (53,8%) izjasnilo se da ne koristi računalo za slušanje glazbe, a samo 6% učitelja to čini često, dok minimalnih 0, 9% ispitanika u nastavnom procesu to čini uvijek. Ako je u pitanju korištenje računala za gledanje crtanih ili dokumentarnih filmova u nastavi također ga najveći udio učitelja za to nikada ne koristi (39, 3%) dok 23, 9% ispitanika ponekad koristi računalo u tu svrhu. 12,8% učitelja često omogućuje učenicima gledanje filomova na računalu, dok ih samo 3,4% to čini uvijek. Nadalje, korištenje interneta u nastavi većini je učitelja očito nepotrebno i nevažno kada ih 41% nikada ne koristi internet. 20,5% ispitanih ga ponekad koristi, a 12,8% ih to čini često, što je vrlo malo s obzirom na bogatstvo informacija koje internet nudi učiteljima za nastavu.

Najveći postotak učitelja (48, 7%) slaže se s tvrdnjom da korištenje ICT-a u nastavnom procesu čini učenicima nastavu zanimljivijom od uobičajene. 29,1% ispitanika djelomično se slaže s navedenom tvrdnjom dok se 20,5% njih u potpunosti slaže da je nastava uz ICT učenicima zanimljivija od uobičajene. Ovim rezultatima se potvrđuje svjesnost učitelja o prednostima ICT-a u nastavi, ali ne i njihova svijest o primjeni iste jer je velik postotak onih koji je ne upotrebljavaju, a mogli bi.
Sukladno ovim rezultatima ni sljedeći nisu iznenađujući. Naime, 43,6% ispitanika slaže se s tvrdnjom da je korištenje ICT-a u nastavnom procesu primarnog obrazovanja korisno i prijeko potrebno. 38,5% njih djelomično se slaže s tom tvrdnjom dok se njih 12% u potpunosti slaže s time. Kumulativnih 6% učitelja se ne slaže s tom tvrdnjom čime potvrđuju zastarjelo razmišljanje da je za zanimljivu nastavu dovoljno imati ploču i kredu. T-testom utvrđena je statistički značajna razlika u stavovima o zanimljivosti nastave uz upotebu ICT-a s obzirom na dob (mladi učitelji (<30 godina) i učitelji u zrelim godinama (40-60> godina)) (Tablica VI i VII) i radni staž ispitanika (<1-10 i 31-40> godina radnog staža). Cross tablicom potvrđeno je kako se popriličan broj učitelja u zrelijim godinama djelomično slaže s tvrdnjom da je učenicima nastava uz upotrebu ICT-a zanimljivija dok se mlađi učitelji skoro svi slažu i u potpunosti slažu s tom tvrdnjom.

Tablica VI: T-test za nezavisne uzorke. Razlika u stavovima o zanimljivosti nastave uz upotrebu ICT-a s obzirom na dob ispitanika (<30 i 40-50 godina)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Stavovi o zanimljivosti nastave uz upotrebu ICT-a	Equal variances assumed	,321	,573	2,074	59	,042	,40775	,19659	,01437	,80113
	Equal variances not			2,273	35,600	,029	,40775	,17940	,04377	,77174

Tablica VII: T-test za nezavisne uzorke. Razlika u stavovima o zanimljivosti nastave uz upotrebu ICT-a s obzirom na dob ispitanika (<30 i 50-60> godina)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Stavovi o zanimljivosti nastave uz upotrebu ICT-a	Equal variances assumed	1,923	,173	3,652	39	,001	,96078	,26309	,42863	1,49293
	Equal variances not			3,956	38,319	,000	,96078	,24285	,46929	1,45227

Zadnjim t-testom u provedenom istraživanju je također utvrđeno da postoji statistički značajna razlika u stavovima o važnosti primjene ICT-a u nastavi s obzirom na dob (30-40 i 50-60> godina) što nije iznenađujuće s obzirom da su u svim prije navedenim čimbenicima mlađi učitelji pokazali prednost pred starijim učiteljima.

Ispitani stavovi o uvođenju Informatike kao obveznog predmeta u primarno obrazovanje potvrdili su svjesnost učitelja o zahtjevnosti vremena u kojem živimo, u kojem se djeca trebaju znati služiti tehnologijom, ali uz mjere i oprez roditelja i učitelja. Većina učitelja slaže se s tvrdnjom da bi Informatika trebala biti obvezan predmet u nižim razredima (37, 6%), a 29, 1% njih se u potpunosti slaže s time. 28, 2% učitelja djelomično se slaže s navedenom tvrdnjom što može biti posljedica mišljenja o preopterećenosti učenika nastavnim predmetima i obvezama. 4, 3% ih se ne slaže s tom idejom, a samo 0, 9% ih se uopće ne slaže s tom tvrdnjom.

Ovim istraživanjem ispitalo se utječu li dob te radni staž ispitanih učitelja te sredina u kojoj se nalazi škola na korištenje ICT-a u nastavi. Prvi čimbenik koji podrazumijeva korištenje ICT-a u nastavi je dovoljna stručna osposobljenost za korištenje iste. Za njega je t-testom utvrđeno kako postoji statistički značajna razlika s obzirom na dob (<30 i 40-50 godina te <30 i 50-60> godina) i s obzirom na radni staž (<1-10 i 21-30 godina radnog staža te <1-10 i 31-40> godina staža), ali nema statistički značajne razlike s obzirom na sredinu u kojoj je škola. Drugi čimbenik koji je važan za korištenje ICT-a u nastavi je mogućnost i učestalost upotrebe ICT-a u nastavi. T-testom za nezavisne uzorke utvrđena je statistički značajna razlika u učestalosti upotrebe ICT-a s obzirom na dob (<30 i 50-60> godina), dok statistički značajne razlike za radni staž i sredinu u kojoj je škola nema. Neizostavan čimbenik koji također definira korištenje ICT-a je i njen utjecaj na kvalitetu nastavnog procesa. Kada je riječ o stavovima učitelja o zanimljivosti nastave pri upotrebi ICT-a utvrđena je statistički značajna razlika s obzirom na dob ispitanika (<30 i 40-50 godina te <30 i 50-60> godina) te s obzirom na radni staž (<1-10 i 31-40> godina radnog staža). Nadalje, statistički značajna razlika s obzirom na dob ispitanika (30-40 i 50-60> godina) dobivena je i kod ispitivanja stavova o važnosti primjene ICT-a u nastavi. Važno je naglasiti kako se podaci dobiveni ovim istraživanjem, zbog ograničenog uzorka, ne mogu primjeniti na cjelokupnu populaciju.

Budući da je statistički značajna razlika za dob dobivena kod svih čimbenika važnih za korištenje ICT-a Hipoteza 1: „Korištenje ICT-a u nastavnom procesu povezano je s dobi ispitanika“ je prihvaćena. Statistički je dokazano kako mlađi učitelji više upotrebljavaju ICT u nastavi, vještiji su u njenoj uporabi te svjesniji njenog utjecaja na kvalitetu nastavnog procesa. Potvrđena hipoteza otvara prostor za daljnja istraživanja na ovu temu koja bi trebala biti opširnija i obuhvatiti veći uzorak ispitanika.
S obzirom na to da su statistički značajne razlike u odnosu na radni staž ispitanika dobivene samo kod nekih čimbenika korištenja ICT-a Hipoteza 2: „Korištenje ICT-a u nastavnom procesu povezano je sa stažom ispitanika“ može se djelomice potvrditi. Hipoteza 3: „Korištenje ICT-a u nastavnom procesu povezano je sa sredinom u kojoj je škola“ u potpunosti je odbačena jer nije utvrđena statistički značajna razlika između ruralne i gradske sredine.

IV. ZAKLJUČAK

U današnje vrijeme upotreba ICT-a u nastavi je nezaobilazna, stoga je važno da učitelji shvate važnost stalne edukacije za njenu primjenu kako bi učenicima, na taj način, postavili što kalitetnije temelje za cjeloživotno učenje.

Ovim istraživanjem utvrđeno je kako su mlađi učitelji sigurniji u svoju stručnu osposobljenost za korištenje ICT-a od starijih učitelja te češće upotrebljavaju računalo u nastavi. Nadalje, mlađi učitelji smatraju da je nastava uz ICT zanimljiva i korisna u primarnom obrazovanju, dok je kod starijih učitelja to mišljenje poljuljano. To je posljedica odrastanja u različitom vremenu. Mlađi učitelji stjecali su obrazovanje koristeći se tehnologijom pa im je stoga i lakše pratiti stalne inovacije, dok kod starijih učitelja to nije slučaj. Ipak, to nije izlika potonjim da se ne mogu stručno usavršavati i naučiti primjeniti saznanja o tehnologiji u nastavi.

Treba imati na umu kako ne postoji pravi način uporabe ICT-a u nastavi. Ona zapravo nije ni dobra, ni loša. Bitna je mjera u kojoj se ona upotrebljava te način uporabe koji mora biti sukladan ciljevima učenja i poučavanja, osobitostima učeničke populacije te individualnim potrebama svakog učenika.

LITERATURA

[1] J. Maravić, UNESCOV dokument o informacijskoj i komunikacijskoj tehnologiji u obrazovanju. Dostupno na: http://edupoint.carnet.hr/casopis/20/clanci/4.html (Preuzeto 30.03.2012.)

[2] Ž. Hutinski, Informacijska i komunikacijska tehnologija u obrazovanju: stanje i perspektive. Dostupno na: http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=66281. (Preuzeto 30.03.2012.)

[3] Babić, N., Irović, S., Krstović, J., Suvremene informacijske tehnologije i edukacija, Informatologia 36, 1, str.8.-13., 2003.

[4] V. Popović, Realizacija nastavnog programa „Mediji u obrazovanju“. Dostupno na: http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=33857. (Preuzeto 02.04.2012)

FACEBOOK U OBRAZOVANJU

Dorotea Drenjanac*, C. Mario Dumančić **
* Osnovna škola „Vugrovec-Kašina“, Ivana Mažuranića 43, Zagreb, Hrvatska
** Učiteljski fakultet u Zagrebu, Savska cesta 77, Zagreb, Hrvatska
dorotea.drenjanac@gmail.com, mario.dumancic@ufzg.hr

Sažetak - Posljednjih nekoliko godina društvene mreže su sve popularnije među učenicima te kao takve u velikoj mjeri utječu na njihov osobni razvoj. Zbog svoje popularnosti među učeničkom populacijom neki od učitelja su uvidjeli mogućnost njihove implementacije u obrazovni proces. U ovom radu je istražena implementacija Facebooka, kao najraširenije mreže među mladima, u obrazovnom procesu. Kako bi se Facebook želio implementirati u obrazovni sustav, prije svega potrebno je educirati učitelje o mogućnostima društvenih mreža kao nastavnih metoda i sredstava, te roditelje čija će djeca biti aktivni sudionici istih. U ovom radu cilj je istražiti i prikazati u kojoj se mjeri društvena mreža Facebook koristi u osnovnim školama s obzirom na mjesto stanovanja, te koristi li se u obrazovne svrhe i na koji način.

I. UVOD

Suvremena nastava temelji se na raznovrsnosti nastavnog sadržaja i nastavnih izvora. Uz udžbenike, sve učestaliji izvori znanja su upravo mediji. Mediji, kao sustavi javnog informiranja, prenosioci vijesti te raznih audio-vizualnih sadržaja, služe u svrhu informiranja, obrazovanja i zabave. Pod najučestalije medije ubrajaju se tisak, radio, televizija te najsuvremeniji medij internet. [1] Internet, kao novi medij, postao je neizostavan i gotovo najvažniji medij današnjeg društva, a samim tim i školskih ustanova. Djeca rastu okružena medijima, od rođenja su im dostupna televizija, radio, a sve više i internet. Djeca u današnje vrijeme sve ranije počinju koristiti internet koji sve više utječe na formiranje njihovih stavova i mišljenja te za njih predstavlja neiscrpan izvor znanja. Internet ima brojne pozitivne strane; integracija svih medija, zanimljivost, brzo pronalaženje informacija, učenje uz pomoć interneta, novi oblici učenja, povezanost s udaljenim školama, svakodnevno usavršavanje, uspostavljanje novih poznanstava, integracija odnosa učenik-učitelj-roditelj, mogućnost e-učenja i dr.[2] Shodno tome sve se češće spominje i ipmultimedijalna[1] nastava kao bogaćenje tradicionalne nastave s multimedijskim materijalima s interneta. Svjedoci smo razvoja i mogućnosti društvenih mreža koje su ne samo promijenile način komunikacije, nego i međusobno dijeljenje multimedijskih sadržaja. Autor Christakis određuje društvenu mrežu kao strukturu sastavljenu od pojedinaca ili organizacija koje su vezane preko jedne ili više međuzavisti kao što je prijateljstvo, zajednički interes, srodstvo ili sl. [3] Facebook je danas najveća internetska društvena mreža koju je 2004. osnovao student Mark Zuckerberg sa svojim cimerima i kolegama s faksa Eduardom Saverinom, Dustinom Moskovitzom i Chrisom Hugesom.[4] Facebook kao najpopularnija društvena mreža posjeduje brojne prednosti te mogućnosti implementacije u nastavu. Može li Facebook sa svojom interaktivnošću postati novi komunikacijski i obrazovni kanal usmjeren na učenika? Taj nov, kreativan obrazovni kanal može podržati brzo, lagano te zanimljivo savladavanje sadržaja. Ova društvena mreža posjeduje brojne aplikacije i alate koje su jednostave za uporabu, a predstavljaju zanimljiv potencijal u organizaciji i provedbi nastavnog procesa. Kako bi implementirali društvene mreže, a posebice Facebook u nastavi, učitelji bi trebali naučiti pedagoške i tehničke karakteristike Facebooka, socijalne aspekte učenja putem tehnologije te kako na što bolji mogući način implementirati Facebook u svakodnevnu nastavu. [5]

II. METODOLOGIJA ISTRAŽIVANJA

Cilj ovog istraživanja je utvrditi način korištenja Facebooka, kao jedne od najpopularnijih društvenih mreža, među učenicima viših razreda osnovnih škola. Osim načina korištenja, ispitivali su se stavovi učenika o korištenju Facebooka u nastavi i školi.
Istraživanjem su se željeli ispitati slijedeći problemi: korištenje društvene mreže Facebook, vrijeme provedeno na mreži, istinitost osobnih podataka na profilu, stavovi roditelja o korištenju mreže, najčešće aktivnosti na mreži, dostupnost učitelja/profesora na mreži, grupa na razini razreda ili škole, mogućnosti razredne grupe, mogućnosti Facebooka u obrazovanju te stavovi o korištenju Facebooka u obrazovanju.[2]

Na osnovu postavljenog cilja istraživanja kreirali smo osnovne četiri hipoteze:

Hipoteza 1: Učenici u gradskim sredinama češće koriste Facebook nego učenici u ruralnim sredinama.

Hipoteza 2: Učenici koriste Facebook, ali ne u obrazovne svrhe.

[1] IPmultimedija – prijenos multimedijskih sadržaja putem računalne mreže (interneta)

[2] Nezavisne varijable u ovom istraživanju su dob, spol djeteta te sredina u kojoj pohađaju školu. Zavisna varijabla bila je korištenje Facebooka u obrazovne svrhe.

Hipoteza 3: Ne postoji statistički značajna razlika u korištenju Facebooka u obrazovne svrhe s obzirom na spol ispitanika.

Hipoteza 4: Ne postoji statistički značajna razlika u korištenju Facebooka u obrazovne svrhe s obzirom na dob ispitanika.

A. Uzorak ispitanika, instrumenti i metode obrade podataka

Istraživanje smo proveli na uzorku od 187 ispitanika, od čega 82 učenika iz ruralnih sredina, a 105 iz gradskih. Pretpostavili smo da učenici sedmih i osmih razreda najviše koriste društvene mreže (dob 13 i 14 godina). Anketirano je 98 dječaka te 89 djevojčica, od toga 83 učenika ide u sedmi razred, a 104 u osmi razred osnovne škole.

U istraživanju su sudjelovale Osnovna škola Otok iz Zagreba, Osnovna škola Mato Lovrak iz Zagreba, Osnovna škola Vugrovec- Kašina iz Kašine te III. osnovna škola Čakovec iz Čakovca.
U Osnovnoj školi Otok iz Zagreba ispitano je ukupno 67 učenika.
U Osnovnoj školi Mato Lovrak iz Zagreba ispitano je ukupno 40 učenika.
U Osnovnoj školi Vugrovec- Kašina iz Kašine ispitano je ukupno 44 učenika.
U III. osnovnoj školi Čakovec iz Čakovca ispitano je 36 učenika.

U istraživanju smo koristili anoniman anketni upitnik za učenike. Upitnik se sastojao od 13 pitanja zatvorenog tipa te 1 pitanja otvorenog tipa. U određenim pitanjima moguće je zaokružiti više od jednog odgovora. Također, određena pitanja uz ponuđene odgovore nude i mogućnosti samostalnog upisivanja odgovora.

U obradi podataka korištena je deskriptivna analiza podataka i t-test za nezavisne uzorke, kojim su utvrđene razlike između aritmetičkih sredina. Razina statističke značajnosti je postavljena na $p<0,05$.

III. REZULTATI I RASPRAVA

Statističkom obradom podataka pokazalo se da većina učenika (93,6%) obuhvaćenih ovim istraživanjem posjeduje profil na društvenoj mreži Facebook što je potvrdilo našu pretpostavku o popularnosti i korištenju društvenih mreža kod učenika sedmih i osmih razreda. Od ukupnog broja učenika najveći dio odgovorio je da dnevno provodi manje od jedan sat na Facebooku, njih 69 (36,9%). 59 učenika na mreži provodi između 1 i 2 sata dnevno (32,6%) te njih 47 na preži provodi više od dva sata (25,1%). T- testom za nezavisne uzorke utvrđena je statistički značajna razlika u provođenju vremena na Facebooku s obzirom na spol ($p<0,05$).

Tablica I. T-test za nezavisne uzorke. Razlika u provođenju vremena na Facebooku s obzirom na spol

		Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Difference	Difference	Lower	Upper
VRIJEME NA MREŽI	Equal variances	,225	,636	-4,955	173	,000	-,56729	,11448	-,79324	-,34133
	Equal variances			-4,954	172,484	,000	-,56729	,11452	-,79332	-,34125

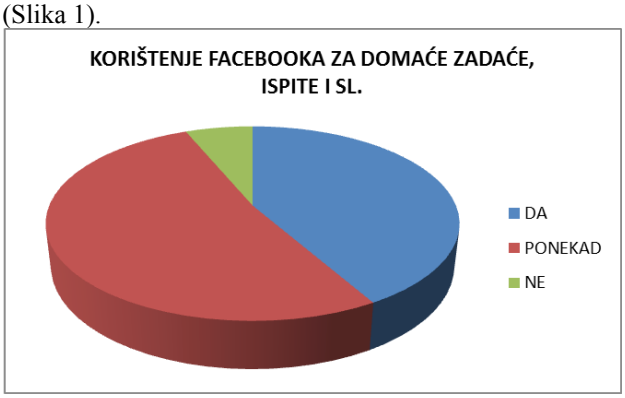
Iznenađujući je podatak da čak 90 učenika (48,1%) posjeduje profil s djelomično točnim podacima, 79 učenika ima potpuno istinite podatke o sebi (42,2%) dok ih je tek 6 odgovorilo da nema profil s osobnim podacima (3,2%), tj. posjeduje lažni profil.
T- testom za nezavisne uzorke utvrđena je statistički značajna razlika u posjedovanju Facebook profila s obzirom na osnovnu školu ($p<0,05$) (Tablica 4) dok s obrizom na spol i dob ne postoji statistički značajna razlika. Cross tablica pokazala je kako više učenika iz ruralnih sredina koristi Facebook nego iz gradskih.

Tablica II. T-test za nezavisne uzorke. Razlika u posjedovanju profila s obzirom na osnovnu školu.

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Koristiš li društvenu mrežu Facebook	Equal variances	31,118	,000	2,559	185	,011	,09143	,03573	,02093	,16193
	Equal variances			2,838	138,849	,005	,09143	,03222	,02772	,15513

Poražavajuća je informacija da je čak 44 učenika (23,5%) odgovorilo kako njihovi roditelji nemaju znanja o društvenoj mreži Facebook. 65 učenika (34,8%) odgovorilo je kako njihovi roditelji smatraju da previše koristi Facebook, dok roditelji 40 učenika (21,4%) smatra da je korištenje Facebooka u redu jer se saznaju važne informacije o školi i nastavi. Roditelji 32 učenika (17,1%) smatraju da je Facebook dobar način upoznavanja novih prijatelja dok se roditelji 16 učenika (8,6%) nikako ne slažu s korištenjem Facebooka.
Analiza podataka Cross tablicom pokazala je kako uglavnom roditelji djece iz ruralnih sredina imaju manje znanja o društvenoj mreži Facebook nego roditelji djece u gradskim sredinama.
Učenici su trebali označiti koliko često (uvijek, često, ponekad, rijetko, nikad) koriste pojedine aktivnosti (chat, igrice, nastavne grupe, pregledavanje zidova) na društvenoj mreži. Najveći broj učenika, njih 80 (42,8%) odgovorilo je da uvijek koristi Facebook za chat/dopisivanje dok ih je tek 5 (2,7%) odgovorilo da nikad ne koristi chat. Rezultati pokazuju da je većina učenika (29,9%) nikada ne koristi Facebook za igrice dok je 70 učenika (37,4%) odgovorilo kako nikada ne koristi nastavne grupe što je jedan od dokaza kako učenici koriste Facebook, ali ne u obrazovne svrhe. 107 učenika

(57,3%) odgovorilo je da uvijek i često pregledava zidove tuđih profila. Rezultati pokazuju kako su najzastupljenije aktivnosti na mreži chat/dopisivanje te pregledavanje grupa.
Na pitanje koriste li Facebook za razgovor s prijateljima o domaćoj zadaći, ispitima i sl. 73 učenika (39%) odgovorilo je da, 91 učenika (48,1%) odgovorilo je ponekad, a tek njih 11 (5,9%) odgovorilo je da ne koristi (Slika 1).



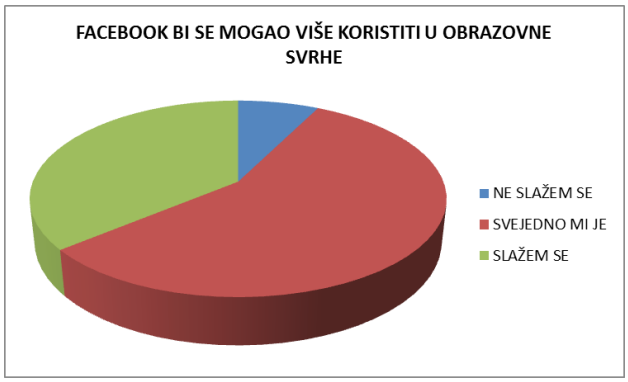
Slika I. Korištenje Facebooka za pitanja vezana uz nastavu

Na pitanje o dostupnosti profesora na društvenoj mreži učenici su odgovorili da većina profesora (72%) uopće nije dostupna na mreži radi pitanja i nejasnoća vezanih za nastavu, dok ih je samo 9,7% dostupno. Zaključak je, ako se želi implementirati Facebook u nastavu kao suvremeno bogaćenje, profesori i učitelji su ti koji prvi trebaju otvoriti profil i biti dostupni učenicima. T- testom za nezavisne uzorke utvrđena je statistički značajna razlika u dostupnosti profeosora s obzirom na školu. ($p<0,05$) Prema rezultatima iz Cross tablice u kojoj se uspoređivala dostupnost profesora s obzirom na školu, najviše profesora dostupno je učenicima iz Osnovne škole Vugrovec-Kašina.

Slika II. Dostupnost profesora na društvenoj mreži Facebook

Većina učenika (69,7%) odgovorilo je kako ne posjeduju grupu na Facebooku na razini razreda ili škole, a njih 30,3% odgovorilo je kako posjeduju. Većina učenika (19,3%) koji imaju grupu na razini razreda ili škole je odgovorilo kako im ta grupa omogućava uvid u podsjetnik za domaće zadaće, ispite i sl., 15,5% učenika pomoću te grupe ima uvid u sve edukativne materijale, 13,9% učenika služi tu grupu kao podsjetnik za školska događanja, a 0,5% ih je navelo kako im ta grupa omogućava jednostavno komunikaciju s prijateljima.
Prema rezultatima dobivenim iz Cross tablice utvrđeno je da više učenika iz ruralnih sredina posjeduje nastavnu grupu nego učenici iz gradskih sredina.

Većina učenika nemaju mišljenje i stav o upotrebi Facebooka u obrazovne svrhe, svejedno im je (56,6%). 13 učenika (7,4%) se ne slažu s upotrebom Facebooka u nastavi, dok se njih 63 (36%) slažu. Učenici u današnje vrijeme, kao ni učitelji, nisu svjesni svih mogućnosti društvene mreže kao edukativnog bogaćenja nastave. Većina učenika koji su odgovorili kako bi se društvena mreža mogla više koristiti u obrazovne svrhe, kao glavni razlog korištenja naveli su za informiranje o domaćim zadaćama, ispitima i sl. te smatraju kako je Facebook dobar način prenošenja informacija o nastavi učenicima koji su bolesni ili zbog nekog drugog razloga nisu u mogućnosti pohađati nastavu.



Slika III. Korištenje Facebooka u obrazovne svrhe

Prva hipoteza koju sam navela (Učenici u gradskim sredinama češće koriste Facebook nego učenici u ruralnim sredinama) nije prihvaćena. Kao što je u rezultatima istraživanja vidljivo, postoji statistički značajna razlika u korištenju Facebooka s obzirom na sredinu pohađanja škole. Više učenika u ruralnim sredinama koristi društvenu mrežu. Ovim istraživanjem razbijene su moje predrasude o ruralnim sredinama. Smatrala sam ih nedovoljno razvijenima za posjedovanje interneta uopće, ali dokazano je da učenici ne samo da posjeduju internet, nego su i aktivni korisnici društvenih mreža.
Druga hipoteza (Učenici koriste Facebook, ali ne u obrazovne svrhe) je potvrđena. Učenici Facebook uglavnom koriste za chat i komunikaciju s prijateljima ne vezano za školu i nastavu.
Prva nul hipoteza (Ne postoji statistički značajna razlika u korištenju Facebooka s obzirom na spol ispitanika) je prihvaćena. Facebook koriste podjednako i dječaci i djevojčice. Ali T-testom je utvrđeno kako dječaci dnevno na mreži provode manje od jedan sat, a djevojčice između jedan i dva ili više od dva.

Druga nul hipoteza (Ne postoji statistički značajna razlika u korištenju Facebooka s obzirom na dob ispitanika) je prihvaćena. Učenici sedmih i osmih razreda podjednako koriste društvenu mrežu.
Iznenadio me podatak da su profesori dostupni učenicima na mrežama uglavnom u ruralnim sredinama te da učenici iz ruralnih sredina češće imaju nastavne grupe nego učenici iz gradskih sredina. Dolazim do zaključka kako se u ruralnim sredinama nastava želi što više modernizirati i osuvremeniti. Da bi se to što kvalitetnije izrealiziralo potrebno je educirati roditelje o samoj društvenoj mreži jer, kako je istraživanjem utvrđeno, roditelji djece iz ruralnih sredina imaju manje znanja o Facebooku nego roditelji djece iz gradskih.
Smatram da uzorak nije dovoljno velik da bi se rezultati primijenili na cijelu populaciju te zbog toga za buduća istraživanja na ovu temu predlažem istraživanje s većim brojem uzorka.

IV. ZAKLJUČAK

Odgovornost je svakog učitelja i učiteljice, kao i svih osoba koje su na bilo koji način involvirane u odgoj i obrazovanje, da nastoje svojim učenicima pružiti najbolju moguću poduku i da ovladaju najsuvremenijim metodama kako bi ta poduka bila što efikasnija. Svijet u kojem živimo svijet je medija. Mediji su svuda oko nas i neprekidno nam šalju poruke te izravno ili neizravno utječu na oblikovanje mišljenja nas i naše djece. Zašto tu okruženost medijima ne bi pozitivno iskoristili i izvukli iz njih upravo ono najbolje? Vođena tom mišlju nastojala sam ovim radom dokazati kako je nužnost i potreba iskoristiti napredak elektroničkih medija i veliko zanimanje djece za društvene mreže kao suvremeno nastavno sredstvo. Društvene mreže naša su realnost i realnost su naše djece pa ih je stoga potrebno iskoristiti na što bolji način. Posebnosti koje nam taj novi suvremeni medij donosi neusporedive su sa svim ostalim suvremenim medijima. Nigdje se do sada nisu tako jasno brisali okviri tradicionalnih učionica i tradicionalne nastave kao kod upotrebe društvenih mreža. Ovaj novi način suradnje s učenicima i suradnje učenika međusobno doprinosi krajnjoj demokratičnosti i kreativnosti jer dopušta da se baš svako dijete izrazi onda i onako kako njemu odgovara. Nastava se ne treba više održavati samo i isključivo u učionici. Zašto učenik ne bi izvan nastavnog vremena pitao učitelja nešto što ga zanima?
Možda najvažnija ideja koja mi se kao budućoj učiteljici nekako nametnula sama od sebe dok sam istraživala fenomen društvenih mreža i njihovu iskoristivost kao suvremenog nastavnog sredstva ideja je da svaki učitelj preuzme na sebe odgovornost da educira djecu i roditelje o pravilnom i sigurnom korištenju interneta i društvenih mreža te njihovoj svakodnevnoj implementaciji u nastavu.
Valja odgovorno shvatiti svoju ulogu u odgoju i obrazovanju djece, istražiti mogućnosti modernih medija i iskoristiti njihove prednosti u nastavi. To pridonosi aktivnijem i boljem načinu učenja, ali i neusporedivo je odgojnije od mnogih drugih nastavnih metoda i sredstava. Ne treba se bojati modernizacije već ju treba iskoristiti za vlastite ciljeve.

LITERATURA

[1] Ilišin, V. & Marinović Bobinac, A. & Rodin, F., (2001), *Djeca i mediji*, Zagreb: Državni zavod za zaštitu obitelji, materinstva i mladeži Institut za društvena istraživanja u Zagrebu

[2] Miliša, Z. & Tolić, M. & Vertovšek, N., (2010), *Mladi-odgoj za medije*, Zagreb: M.E.P. d.o.o.

[3] Christakis, N.A., (2010), Povezani : iznenađujuća moć društvenih mreža i kako one utječu na naše živote, Zagreb: Algoritam

[4] Mamić, D., P., (2011), Društvena mreža Facebook, Diplomski rad, Fakultet elektronike i računarstva, Sveučilište u Zagrebu.

[5] Kirkpatrick, D., (2010), *The Facebook effect* [CD], NY: Simon & Shuster

Usporedba stavova dvije generacije budućih učitelja prema odgoju i obrazovanju u budućnosti

Vinkoslav Galešev
Udžbenik.hr, Zagreb, Hrvatska
vgalesev@enter.bug.hr

Sažetak - U sklopu praćenja promjena stavova novih generacija studenata prema promjenama školskog sustava i pedagoških vrijednosti u budućnosti kao posljedice uvođenja informacijsko-komunikacijske tehnologije u odgoj i obrazovanje, kao i u društvo u cjelini, ispitani su studenti[1] Učiteljskog fakulteta u Zagrebu, budući učitelji primarnog obrazovanja. Dobiveni rezultati uspoređeni su s rezultatima istraživanja provedenog na studentima prijašnjih generacija (prije dvije godine) istog fakulteta. Pomoću komponentne analize istražene su kvalitativne i kvantitativne promjene stavova studenata između dva mjerenja.

Cilj istraživanja je ispitati je li kod nove generacije studenata došlo do značajnijih kvalitativnih i/ili kvantitativnih promjena u stavovima prema odgoju i obrazovanju u budućnosti, što može ukazati na potrebu prilagodbe određenih nastavnih sadržaja tijekom studija.

I. UVOD

Budući da je strateška svrha obrazovanja rješavanje mikro i makro razvojnih problema društva [1], a koji u sadašnjosti nisu poznati i teško su predvidivi, planiranje ciljeva i sadržaja obrazovanja neminovno spada u područje futurologije [2]. Promjene zahtijevaju ne samo prilagođavanje i futurističko planiranje odgojno-obrazovnog sustava, nego i temeljno restrukturiranje škole. Gledajući na globalnoj razini, više nije pitanje kako unaprijediti školski sustav, već kako ga iznova osmisliti [3].

Postavlja se pitanje kako na ove promjene, koje se tako brzo odvijaju da o njima ne promišljaju samo futurolozi, gledaju studenti – budući učitelji. Kakav je njihov stav prema eksponencijalnom tehnološkom razvoju, a poglavito uvođenju tehnologije u odgojno-obrazovni proces?

U sklopu praćenja prilagođavanja studenta – budućih učitelja ovim promjenama, posebice u odnosu na primjenu informacijsko-komunikacijske tehnologije u odgojno-obrazovnom procesu u budućnosti, potrebno je longitudinalno pratiti prilagođavanje studenata kontinuiranim promjenama koje se odvijaju u domeni ove tehnologije, a koje imaju neposredni utjecaj na sudjelovanje budućih učitelja u kreiranju i realiziranju vlastite profesije.

[1] Uvažavajući rodnu ravnopravnost u radu se koriste generički nazivi studenti i učitelji u svrhu kratkoće izlaganja.

II. POSTUPAK

Kako je jedna od zadaća istraživanja usporedba stavova starije i novije generacije studenata – budućih učitelja, postupak i provedba ispitivanja u prvom i drugom ispitivanju temeljila se na ujednačenim (iako ne i istovjetnim) značajkama. Prvo ispitivanje provedeno je u 2009./2010., a drugo u 2011./2012. akademskoj godini.

A. Ispitanici

U oba ispitivanja uzorak su sačinjavali studenti prve godine diplomskog studija (četvrta godina studiranja) Učiteljskog fakulteta Sveučilišta u Zagrebu. U prvom ispitivanja veličina uzorka bila je 79 studenata, a u drugom 248 studenata. Oba uzorka nesrazmjerna su po spolnoj zastupljenosti, odnosno znatno je više studentica. U prvom uzorku studentica je bilo 91,1 %, a u drugom 88,3 %.

Potrebno je istaknuti da je i tijekom statističke obrade podataka zadržan ovaj nesrazmjer spolne zastupljenosti (nije izvršeno ponderiranje poduzoraka), budući da odražava stvarno stanje u populaciji studenata Učiteljskog fakulteta.

B. Mjerni instrument

Za ispitivanje stavova studenata prema odgoju i obrazovanju u budućnosti korištena je skala procjene koju je autor konstruirano za potrebe prvog mjerenja, a sastojala se od 29 čestica podijeljenih u četiri područja: *Odgoj i obrazovni sustav*, *Društveni položaj učitelja*, *Uloga učitelja u odgoju i obrazovanju* te *Utjecaj IK tehnologije na nastavu*. Prvotna inačica temeljila se na 7-stupanjskoj skali Likertovog tipa koja pokriva kontinuum od "potpuno se ne slažem " do "potpuno se slažem".

Nakon što su učeni određeni nedostaci u početnoj konstrukciji mjernog instrumenta, u drugom ispitivanju izostavljena je jedna čestica, a korištena je 5-stupanjska skala odgovora na istom kontinuumu.

C. Priprema i obrada podataka

Kako bi se mogli neposredno usporediti rezultati dobiveni u oba mjerenja, podaci prvog ispitivanja dobiveni na 7-stupanjskoj skali statističkim postupkom su rekodirani u 5-stupanjsku skladu.

Za potrebe ovog rada, izdvojene su i analizirane samo one čestice koje se odnose na stav studenata prema uvođenju informacijsko-komunikacijske tehnologije u

odgoj i obrazovanje u budućnosti. To je sljedećih deset čestica:

7. Nastava se neće odvijati u školi, nego u potpunosti preko interneta.
14. O djeci će se kod kuće brinuti roboti-dadilje.
20. Budući da će se sva nastava odvijati preko interneta, učitelji uopće neće upoznati svoje učenike.
21. Ako će obrazovanje preuzeti virtualni učitelji, samo je pitanja vremena kada će ih zamijeniti roboti.
23. Ne želim biti sudionik budućih promjena u obrazovanju jer je prevelik naglasak na ulozi i razvoju tehnologije.
24. Zabluda je da će uporaba moderne tehnologije u nastavi omogućiti učitelju da se više posveti učeniku.
25. Tiskani udžbenici će nestati, a koristit će se samo digitalni multimedijski sadržaji.
26. Moderna tehnologija doprinosi poboljšanju kvalitete odgoja i obrazovanja.
27. Učenici više neće učiti stane jezike jer će imati osobne uređaje za prijevod s bilo kojeg stranog jezika.
28. Umjesto bilježnica i pisala, učenici će koristiti razne digitalne uređaje za pisanje i pohranjivanje sadržaja.

Dobiveni podaci obrađeni su statističkim paketom IBM SPSS verzija 19.

III. REZULTATI

Najprije će biti prikazani i analizirani rezultati dobiveni u drugom mjerenju, nakon čega će se usporediti s rezultatima u prvom mjerenju.

Pojedinačni rezultati prvog mjerenja ovdje nisu prikazani, budući da su objavljeni u prijašnjem radu autora [4].

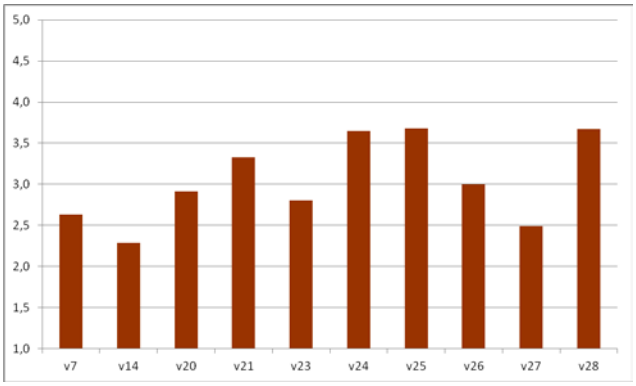
A. *Rezultati drugog mjerenja*

Za razliku od rezultata prvog istraživanja gdje su podaci znatno varirali, u drugom ispitivanju koeficijent varijabilnosti se kreće u rasponu od 32,6 do 54,4 (Tablica I.). To omogućava da prosječne vrijednosti, odnosno aritmetičke sredine smatramo kao zadovoljavajući prezentant podataka. Ujedno, to znači da su stavovi studenata u drugom mjerenju homogeniji nego u prvom ispitivanju provedenom prije dvije godine.

TABLICA I. REZULTATI U DRUGOM ISPITIVANJU

čestica	aritm. sredina	stand. devijac.	koef. varijab.
V7	2,63	1,20	45,6
V14	2,29	1,25	54,4
V20	2,91	1,47	50,4
V21	3,33	1,38	41,5
V23	2,80	1,41	50,3
V24	3,65	1,19	32,6
V25	3,68	1,28	34,7
V26	3,00	1,10	36,7
V27	2,49	1,31	52,5
V28	3,67	1,28	34,9

Na Slici 1. grafički su prikazane prosječne vrijednosti procjena studenata dobivene u drugom ispitivanju.



Slika 1. Grafički prikaz prosječnih rezultata u drugom ispitivanju

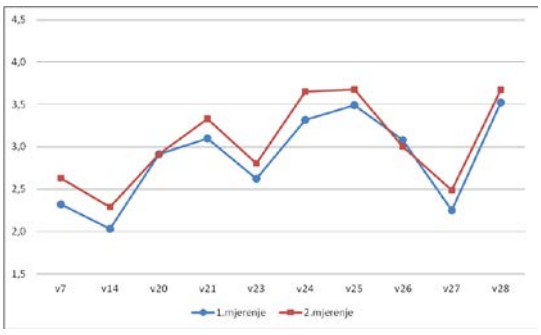
Na osnovu deskriptivne analize podataka na Slici 1. može se uočiti da samo u nekoliko čestica studenti imaju profiliran stav koji odstupa od neutralne središnje vrijednosti (3 - niti se slažem niti se ne slažem). Najizrazitije odstupanje zabilježeno je u čestici 14, što znači da su studenti prilično jedinstvenog mišljenja da u budućnosti o djeci kod kuće neće brinuti roboti-dadilje.

Međutim, studenti su svjesni infiltracije digitalnih medija u odgoj i obrazovanje pa smatraju da će tiskani udžbenici nestati, a koristit će se samo digitalni multimedijski sadržaji (V25) te da će u budućnosti umjesto bilježnica i pisala, učenici koristiti razne digitalne uređaje za pisanje i pohranjivanje sadržaja (V28). Također, mišljenja su da uporaba moderne tehnologije u nastavi neće omogućiti učitelju da se više posveti učeniku (V24). Time studenti izražavaju određenu razinu podozrivosti i negativan stav prema korisnosti uporabe moderne tehnologije u nastavi, odnosno smatraju da uporaba informacijsko-komunikacijske tehnologije ne osigurava kvalitetniji odnos učitelja s učenicima.

B. *Razlike između mjerenja*

Rezultati dobiveni na deset čestica koje ispituju stav prema uvođenju informacijsko-komunikacijske tehnologije u odgoj i obrazovanje u budućnosti uspoređeni su s obzirom na prvo i drugo mjerenje, odnosno ispitivanje, a prosječne vrijednosti prikazane su na Slici 2 (plavi kružići – prvo mjerenje, crveni kvadratići – drugo mjerenje).

Može se uočiti da je struktura, odnosno profil valencije stavova u prvom i drugom mjerenju prilično izjednačena i da prati gotovo jednaku krivulju. U smislu kvalitativnih razlika, na većini čestica stavovi su usmjereni prema neutralnom prosjeku skale procjene (3 - niti se slažem niti se ne slažem).



Slika 2. Grafički prikaz razlika između dva mjerenja

U četiri od deset čestica (V21, V24, V25 i V28), što čini 40%, mišljenje studenata se donekle udaljava od neodređenog stava i približava se višim vrijednostima, odnosno procjeni "slažem se". Što u kvalitativnom smislu znači ovaj pomak? Sve četiri navedene čestice odnose se na svijest o neminovnosti ekspanzije informacijsko-komunikacijske tehnologije u odgoj i obrazovanje, s time da su budući učitelji još više uvjereni da je zabluda da će uporaba moderne tehnologije u nastavi omogućiti učitelju da se više posveti učeniku.

Da bi se ispitala statistička značajnost uočenih razlika, primijenjen je neparametrijski test Mann-Whitney za testiranje razlika između prosječnih vrijednosti dva nezavisna uzoraka (Tablica II).

TABLICA II. TESTIRANJE ZNAČAJNOSTI RAZLIKA IZMEĐU DVA MJERENJA

Mann-Whitney test

	vrijednost	p
V7	8203,5	**0,025**
V14	8264,0	**0,029**
V20	9763,5	0,964
V21	9016,0	0,275
V23	9093,0	0,326
V24	8625,0	0,110
V25	9277,0	0,573
V26	9353,0	0,648
V27	8600,0	0,092
V28	9288,0	0,471

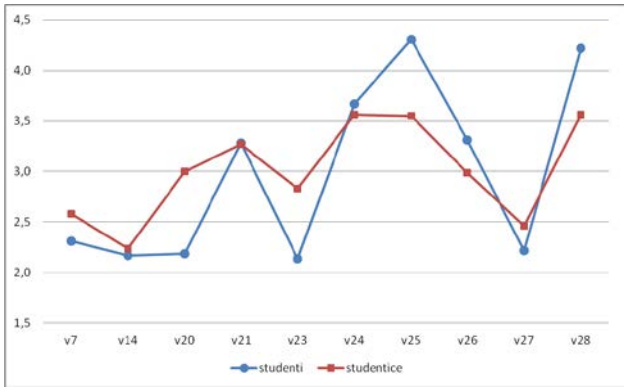
Od deset čestica, samo su u dvije utvrđene statistički značajne razlike (V7 i V14). Iz toga se može zaključiti da su studenti mlađe generacije statistički značajno manje uvjereni da se nastava neće odvijati u školi, nego u potpunosti preko interneta te da se o djeci kod kuće neće brinuti roboti-dadilje. Čini se da su studenti svjesni neminovnog prodora interneta kao obrazovnog medija te postupne robotizacije čak i u najosjetljivijim odgojno-obrazovnim segmentima društva kao što je primjerice odgoj djece (što je pak posredno povezano s profesionalnom karijerom budućih učitelja).

Na osnovi dobivenih rezultata možemo zaključiti da nema statistički značajnih razlika u stavu prema informacijsko-komunikacijskoj tehnologiji u budućnosti između starije i mlađe generacije studenata (razlika je dvije kalendarske godine).

C. *Razlike između spolova*

Iz psihologije i drugih društvenih znanosti poznato je da na formiranje i vrijednost stavova utječe spol. Do nedavno je prevladavalo mišljenje, gotovo na razini stereotipa, da osobe ženskog spola imaju pretežito negativan stav prema uporabi računala, internetu i tehnologiji općenito.

Kako bi se ispitalo aktualno stanje u budućih učitelja (zapravo učiteljica), u istraživanju su analizirane razlike u stavovima između studenata i studentica prema primjeni IK tehnologije u odgoju i obrazovanju u budućnosti. Razlike s obzirom na spol analizirane su samo u drugom ispitivanju, budući da je u prvom utvrđena velika heterogenost uzorka [4], dijelom uvjetovana i relativno malim brojem ispitanika (N=79), što a priori povećava pogrešku mjerenja.



Slika 3. Razlike između spolova

Uvidom u razlike između prosječnih vrijednosti muških i ženskih ispitanika prikazane na Slici 3 može se kazati da postoje kvalitativne i kvantitativne razlike između stavova studenata i studentica, iako se generalno mogu uočiti i određene sličnosti u strukturi njihovih stavova.

Kako bi se utvrdilo postoje li statistički značajne razlike u stavovima studenata s obzirom na spol, korišten je neparametrijski test Mann-Whitney za testiranje razlika između prosječnih vrijednosti dva nezavisna uzoraka (Tablica III).

TABLICA III. TESTIRANJE ZNAČAJNOSTI RAZLIKA IZMEĐU SPOLOVA

Mann-Whitney test

	vrijednost	p
V7	4572,5	0,201
V14	5175,0	0,902
V20	3633,5	**0,002**
V21	5229,0	0,986
V23	3762,0	**0,005**
V24	4832,5	0,454
V25	3650,0	**0,002**
V26	4352,5	0,091
V27	4875,5	0,485
V28	3776,0	**0,005**

U četiri od ukupno deset čestica utvrđene su statistički značajne razlike (V20, V23, V25 i V28), što čini 40% svih čestica.

Ispitanici muškog spola se statistički značajno manje slažu s tvrdnjom da *učitelji uopće neće upoznati svoje učenike budući da će se sva nastava odvijati preko interneta* (V20). Isto tako se manje slažu s izjavom *ne želim biti sudionik budućih promjena u obrazovanju jer je prevelik naglasak na ulozi i razvoju tehnologije* (V23).

S druge strane, studenti se statistički značajno više slažu s tvrdnjama da *će tiskani udžbenici nestati, a koristit će se samo digitalni multimedijski sadržaji* (V25) te da *će umjesto bilježnica i pisala, učenici koristiti razne digitalne uređaje za pisanje i pohranjivanje sadržaja* (V28).

Dobivene statistički značajne razlike upućuju na zaključak da muški studenti statistički značajno pozitivnije gledaju na buduće promjene u odgoju i obrazovanju uvjetovane primjenom informacijsko-komunikacijske tehnologije nego njihove kolegice te statistički značajno više anticipiraju prelazak na digitalne medije u obrazovanju.

Ako usporedimo profil odgovora studentica na Slici 3 s oba profila na Slici 2 (profil odgovora cijelog uzorak u prvom i u drugom ispitivanju), razvidno je da u iskazanim odgovorima prevladava obrazac odgovora studentica. To je i očekivano, budući da su studentice u oba ispitivanja daleko više zastupljenije u uzorku. Ujedno, ova spoznaja navodi na pretpostavku da bi uočene razlike između spolova bile statistički značajnije da je uzorak ispitanika izjednačen po spolu.

IV. ZAKLJUČAK

U deset čestica koje ispituju stav prema primjeni informacijsko-komunikacijske tehnologije u odgoju i obrazovanju u budućnosti, prevladava neutralan stav ispitanika – pretežito se niti slažu niti se ne slažu s predloženim tvrdnjama. Ipak, kroz nekoliko čestica koje odstupaju od neutralne točke na skali procjene moguće je procijeniti narav njihovih stavova. Ispitanici kao cjelina anticipiraju moderne digitalne medije koji će zamijeniti klasičnu nastavu i didaktička sredstva, ali imaju negativan stav prema korisnosti uporabe moderne tehnologije u nastavi, smatrajući da ona ne podiže kvalitetu prije svega odgojne komponente odgojno-obrazovnog procesa.

Kada je riječ o razlikama u stavovima prema informacijsko-komunikacijske tehnologiji s obzirom na spol, istraživači su podijeljenih mišljenja. Dok jedni ističu kako ova tehnologija omogućava gradnju identiteta odvojeno od spolnih određenja jer npr. omogućava korisnicima interneta da budu percipirani poglavito prema onome što kažu [5], drugi upozoravaju da postoji velika razlika između spolova u percepciji računalne kompetencije kao i preferenciji uporabe informacijsko-komunikacijske tehnologije [6], [7].

Rezultati ovog istraživanja upućuju da postoje određene razlike u stavovima prema informacijsko-komunikacijskoj tehnologiji u budućnosti s obzirom na spol, ali da su vjerojatno prikrivene zbog neizjednačenosti uzorka po zastupljenosti spolova, na što se autor odlučio kako bi u uzorku zadržao omjer koji postoji u populaciji studenata Učiteljskog fakulteta.

Prema rezultatima nekih dosadašnjih istraživanja za očekivati je da nove generacije studenata u odnosu na starije generacije imaju pozitivniji stav prema primjerice učenju putem interneta [8], što se u našem istraživanju nije pokazalo. Osim u dvije čestice, u ostalih osam nisu dobivene statistički značajne razlike između prvog i drugog ispitivanja.

Razlog tomu vjerojatno treba tražiti u premalom vremenskom razmaku između dva ispitivanja (vremenski razmak između dviju generacija studenata "samo" je dvije godine, pri čemu samo stavljamo pod navodnike jer su u informatičko svijetu dvije godine respektabilan period za potencijalne promjene). Osim toga, razlog možda treba tražiti i u činjenici da ispitanici iz oba mjerenja (već) pripadaju generaciji "digitalnih urođenika" (*digital natives*), tako da nije za očekivati veće razlike između njihovog informatičkog (pred)iskustva.

U pokušaju analiziranja i interpretiranja dobivenih rezultata potrebno se zapitati temeljno pitanje, specifično za studente buduće učitelje - je li dobro ili loše što u ispitanika nisu dobivene značajne razlike između prvog i drugog ispitivanja? Iako je očito da studentice kao prevladavajući dio uzorka ispitanika imaju negativan stav prema uvođenju i primjeni informacijsko-komunikacijske tehnologije u budućnosti smatrajući da se time ne podiže kvaliteta odgojno-obrazovnog procesa s naglaskom na odgojnu komponentu, takav rezultat ne treba nas iznenaditi. Također, niti navesti na pretjeranu kritičnost. Po autorovom mišljenu, dobro je da budući učitelji, odnosno učiteljice visoko vrednuju odgojnu komponentu i smatraju da ju informacijsko-komunikacijska tehnologija ne može (i ne smije) zamijeniti.

Dobiveni rezultati također upućuju na zaključak da bi tijekom studija budućih učitelja u nastavnim sadržajima i metodičkim vježbama trebalo posvetiti više pozornosti upoznavanju studenata s mogućnostima i prednostima informacijsko-komunikacijske tehnologije i njenoj humanoj primjeni u odgojno-obrazovnom procesu. Tek će temeljito i kritično poznavanje njenih mogućnosti osigurati budućim učiteljima suvereno korištenje te tehnologije u (budućoj) nastavi.

LITERATURA

[1] Pastuović N. (1999). Edukologija: Integrativna znanost o sustavu cjeloživotnog obrazovanja i odgoja, Zagreb: Znamen.

[2] Suzić, N. (2004). *Futurologija obrazovanja*. u: Jugoslovenska pedagogija druge polovine 20. veka, Beograd: Institut za pedagogiju i andragogiju Filozofskog fakulteta u Beogradu,. 125–143.

[3] Milutinović J. (2008). *Futurološka dimenzija ciljeva obrazovanja i učenja*. Pedagogija, 63(1): 41-49.

[4] Galešev V., "Stavovi budućih učitelja prema odgoju i obrazovanju u budućnosti, u: Računala u obrazovanju (zbornik radova), 34. međunarodni skup MIPRO 2011, 23.-27. svibnja 2011., Opatija

[5] Hakken D., "Cyborgs@Cyberspace?: An Etnographer Looks to the Future", Routlege, 1999, pp. 88.

[6] Busch T.,"Gender differences in self-efficacy and attitudes toward computers", Journal of Educational Computing Research, 1995, vol. 12, 147-158.

[7] Burger T., "Percepcija i stavovi studenata prema 'učenju putem Interneta'", ocjenski rad, Filozofski fakultet u Zagrebu, 2003.

[8] Valenta, A., Therriault, D., Dieter, M. and Mrtek, R., "Identifying Student Attitudes and Learning Styles in Distance Education", Journal of Asynchronous Learning Networks; Volume 5; Issue 2- September 2001.

Primjena web sadržaja i web alata na području filozofije znanosti

Martina Sertić
Industrijsko-obrtnička škola Slavonski Brod, Republika Hrvatska
Studentica PDS Informacijskih i komunikacijskih znanosti Filozofskog fakulteta u Zagrebu, Republika Hrvatska
sertic_79@yahoo.com

Sažetak – Ovaj rad daje osnovni uvid u sadašnje društvo znanja i napretka informacijsko-komunikacijske tehnologije te iznosi neke od principa i organizacije upravljanja informacijama i znanjem, i to ponajprije kroz internet i mrežne tehnologije. U radu sam navela nekoliko web sadržaja i web alata. Područje istraživanja i primjeri u radu vezani su uz filozofiju znanosti. Posebno sam istaknula važnost prenošenja, prikupljanja i obrade informacija kroz sustave za upravljanje sadržajem. Tom prilikom navela sam dva najpoznatija i najkorištenija CMS alata *Joomlu* i *Drupal*. U sadašnjem društvu znanja razvojem informacijske tehnologije pristup podacima postao je lakši, jednostavniji i efikasniji. Društvene mreže i sam CMS sustav dali su današnjem svijetu priliku da se globalno poveže i organizira, ali kultura organizacije informacija i kvaliteta informacija putem modernih tehnologija još uvijek ostaje upitna. Može se pretpostaviti da će budućnost donijeti nešto novo, neki organiziraniji sustav upravljanja sadržajem na globalnoj razini, koji neće biti tako lako izloženi manipulacijama i nedorčenosti sadržaja.

I. UVOD

Web sadržaji i web alati sve više postaju sastavni element obrazovnog procesa. Učenici korištenjem informacijske tehnologije u nastavi postaju aktivniji sudionici, a ne samo pasivni promatrači obrazovnog sustava. U radu sam navela nekoliko primjera, tj. web sadržaja i web alata, koje bi učenici svakako trebali koristiti tijekom obrazovnog procesa i to ponajprije na područjima prirodnih znanosti.

Osnovna ideja ovoga teksta zasniva se na istraživanju i pojašnjenju naravi i temelja prostora i vremena korištenjem web sadržaja i web alata. Kako bih zadovoljila ciljeve ovoga izlaganja ponajviše ću se posvetiti prikazu korištenja web sadržaja vezanih uz područje fizike s posebnim osvrtom na filozofiju prirode. Predstavit ću i program *Orbit Xplorer* kao dostojnu virtualnu simulaciju kretanja planetarnih tijela.

Na kraju rada nalazi se i anketni upitnik koji je sastavljen s namjerom provjeravanja zadovoljstva učenika korištenjem informacijskih tehnologija tj. web sadržaja i web alata u nastavi prirodnih znanosti. Zatim su prikazani podaci analize anketnog upitnika, a neki, važniji za ovaj rad, i grafički su istaknuti.

II. DRUŠTVO ZNANJA

Terminom "društvo znanja" nazivamo društveni i gospodarski poredak u visoko razvijenim zemljama. Termin se odnosi na ekonomiju koja je utemeljena i građena na znanju koje se koristi za proizvodnju i upravljanje. U društvu znanja znanje se koristi kao alat, a u gospodarstvu znanja znanje je proizvod. Kao ključni pojam tog sektora gospodarske aktivnosti često se spominju i internetske tehnologije. [7]

Koncept društva znanja je, između ostalih, 1966. godine koristio i američki sociolog Robert E. Lane. Američki sociolog Daniel Bell popularizirao je koncepciju društva znanja 1973. godine u svojoj studiji "The Coming of Post-Industrial Society". Po nekim teorijama znanje je postalo strateški resurs u proizvodima i uslugama, a u budućnosti znanje će biti umreženo, decentralizirano i interdisciplinarno.

A. *Društvo znanja i informacijska tehnologija*

Internet i digitalna tehnologija pružaju nove prilike za učenje, omogućavaju pojedincima direktnu prodaju ideja, usluga ili proizvoda, pristup informacijama te zadovoljavanje emocionalnih i psiholoških potreba. Javljaju se tzv. "virtualne zajednice" utemeljene na zajedničkim interesima. S druge strane, vidljiva je i pojava otuđenja, najčešće unutar institucije obitelji, te ovisnosti o novim tehnologija – o *Facebooku*, igricama, filmovima i drugim interaktivnim sadržajima. [8]

Globalizacija podrazumijeva brisanje umjetnih granica slobodnim protokom dobara, usluga, kapitala, znanja i ljudi. Omogućile su ju niske cijene prijevoza i komunikacija, a sve to dovodi do povezanosti globalnog društva i društva znanja s novonastalom ICT revolucijom. Unatoč globalizaciji, i dalje svi nemaju jednaku mogućnost pristupa informacijskoj tehnologiji i internetu. Zapadne zemlje (Europa i Sjeverna Amerika) uvelike prednjače u napretku i razvoju ICT sektora, za razliku od zemalja Istoka (Azija, Afrika), izuzev Japana i Južne Koreje te, u novije vrijeme, i Kine. Taj razdor između informacijski bogatih i informacijski siromašnih dovodi do neravnopravne dostupnosti informacijskim sadržajima

i u neku ruku pomalo i usporava pokrenuti globalizacijski proces. [6]

U informacijskoj ekonomiji produktivnost i konkurentnost tvrtki ovise o njihovoj mogućnosti stvaranja, obrade i primjene informacija utemeljenih na znanju, a oslanjaju se na intenzivnu uporabu informacijske tehnologije. Informacijska je ekonomija globalna, visoko produktivna i uključuje profit koji proizlazi iz brzine inovacija. Utjecaj ICT revolucije očituje se u pojavi e-trgovine, tj. poslovnim transakcijama koje koriste elektroničku komunikaciju i digitalnu tehnologiju. S novom tehnološkom revolucijom došlo je do promjena na tržištu rada. Naime, sve su traženiji stručnjaci na području ICT usluga, tehnologije i informacijske ekonomije, ali i niz pružatelja intelektualnih usluga, kao što su odvjetnici, znanstvenici, profesori, novinari i sl. U ovom će novom okružju neophodno biti cjeloživotno učenje, koje će omogućiti upravo informacijska tehnologija. [3] [8]

III. WEB SADRŽAJI I WEB ALATI

Web sadržajima i web alatima najlakše pristupamo preko mreže putem interneta. Ovaj termin također može značiti primjenu računala na web tehnologijama čiji se sadržaji prikazuju u web preglednicima tj. primjenu računala koje se kodira u pregledniku podržanog jezika, kao što su *JavaScript* u kombinaciji s HTML-om i oslanja se na zajedničke web preglednike kako bi bili omogućeni zahtjevi određenog softvera. [4]

Web sadržaji i web alati popularni su zbog korištenja web preglednika, sposobnosti brzog ažuriranja i održavanje web aplikacije i to bez dodatne distribucije i instaliranja softvera na računalo. Ključni je razlog njihove popularnosti, kao što je već gore navedeno, velika kompatibilnosti sa svim platformama za rad unutar nekog sustava. Uobičajene web aplikacije uključuju webmail, online trgovinu, online aukcije, online bankarstvo, *e-learning*, sustave za upravljanje sadržajem i mnoge druge funkcije. [2]

A. *Sustav za upravljanje sadržajem (CMS)*

Content management system (CMS), tj. sustav za upravljanje sadržajem, računalni je sustav koji omogućuje objavljivanje, uređivanje i modificiranje sadržaja, kao i održavanja web lokacije iz središnjeg web preglednika na računalu. On pruža mogućnost upravljanja sadržajem u kratkom vremenu na vrlo jednostavan, praktičan i efikasan način.

Temeljne značajke sustava za upravljanje sadržajem variraju od sustava do sustava. Mnogi jednostavniji sustavi posjeduju samo osnovne mogućnosti, dok oni složeniji pružaju pregršt mogućnosti po kojima CMS postaje moćan alat za upravljanje sadržajem.

Neke su od osobina sustava za upravljanje sadržajem sljedeće: sustav posjeduje kontrolni pristup podacima na temelju uloge korisnika, jednostavno pisanje izvještaja, unošenje i izmjene podataka, olakšavanje i poboljšavanje komunikacije među korisnicima, definiranje podataka kao što su dokumenti, filmovi, slike, članci i sl. [6]

B. *Primjeri primjene CMS sustava*

U ovom poglavlju bit će riječi o trenutačno jednom od najkvalitetnijih i najkorištenijih alata za upravljanje web sadržajem – *Joomli i Drupalu*.

Joomla je CMS sustav za upravljanje web sadržajem otvorenog koda (*open source*) koji služi objavljivanju sadržaja na World Wide Webu, besplatan je za korištenje. Pisan je u PHP-u, a za spremanje podataka koristi bazu podataka MySQL. *Joomla* je jedan od ponajboljih sustava za upravljanje sadržajem koji omogućuje izradu web stranica i moćnih online alata za sve namjene. [1]

Joomla dolazi s opcijama za upravljanje tekstom i slikama, prikazom sadržaja u RSS-u, PDF-u i verziji za printanje, raznim modulima kao što su "posljednje novosti i popularni sadržaji", upravljanje *bannerima* i anketama, tražilicom i kontakt-formom. Osim navedenih opcija dostupne su tisuće specijaliziranih dodataka koji mogu *Joomla* web stranicu pretvoriti npr. u web trgovinu, multimedijalnu galeriju ili online zajednicu. Još je jedna odlika ovoga sustava i podrška za strane jezike pa je tako moguće preuzeti hrvatski jezični paket koji će stranicu u potpunosti prevesti na hrvatski jezik. *Joomla* je izdana pod GNU GPL v2 licencom. [1]

Slika 1. Prikaz online ponuđenih besplatnih uzoraka CSS-a za izradu weba u *Joomli*

Sadržaj *Joomlinih* web stranica može biti raznolik: tekst, video, fotografije, glazba, dokumenti ili nešto sasvim drugo. Glavna prednost svakog CMS sustava, pa i ovog *Joomlinog*, jest to što za izradu weba nije potrebno veliko znanje o programiranju. Zato se *Joomla* koristi širom svijeta za izradu web stranica raznih oblika i veličina, i to ponajviše: poslovnih web stranica i portala, online časopisa, novina i publikacija, e-rezervacija i e-prodaje, e-bankarstva, državnih i komercijalnih aplikacija, obrazovnih i vjerskih portala, društvenih i turističkih portala, osobnih i obiteljskih web stranica. [5]

Drupal je besplatan sustav za upravljanje sadržajem (CMS), otvorenog koda (*open-source*), pisan u PHP jeziku. D*rupal* ima široku primjenu – od malih osobnih prezentacija, blogova, pa do robusnih velikih poslovnih web stranica za e-trgovinu, e-obrazovanje, e-bankarstvo i sl. Postoji podrška i za My SQL i za PostgreSQL. *Drupal* je najstariji i jedan od najmoćnijih CMS-a. Na tržište je izašao kao alat otvorenog koda 2001. godine i s vremenom se razvio u izuzetno napredan i skalabilan sustav pogodan i za najveće projekte. Virtualna aplikacija za instalaciju *Drupala* može se pronaći na stranicama Bitnamija www.bitnami.com, ali svi sadržaji na ovim i sličnim stranicama na kojima se nalaze potrebne aplikacije za rad u *Drupalu* nisu uvijek besplatne kako što je to slučaj s *Joomlom*. [1]

Drupal je CMS za najzahtjevnije jer je, za razliku od Joomle izuzetno kompliciran za implementaciju i administraciju, a u neku ruku i za korištenje s čudnim rasporedom grafičkog sučelja: npr. Drupal nema *rich text editor* dok ga ne dodate u obliku *plug-ina*.

Slika 2. *Drupal* sučelje alata za upravljanje sadržajem

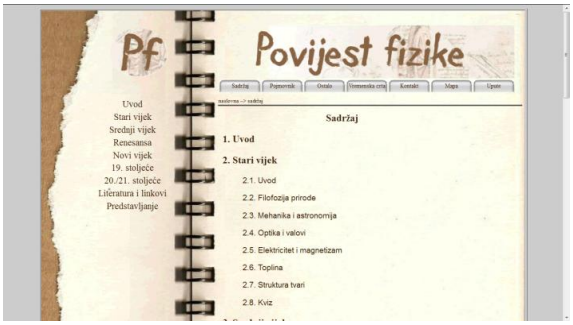
Ovo je CMS velikih mogućnosti, izuzetno kompleksan za učenje, za potpuno iskorištavanje svih njegovih mogućnosti potrebno je dobro poznavanje ostalih web aplikacija, programskih jezika, rada na bazama podataka. Nije prikladan za manje projekte ili osobne web stranice i za organizacije s ograničenim IT resursima.
Najpoznatije web stranice izrađene u *Drupalu* su: www.whitehouse.gov i www.economist.com. [5]

IV. WEB SADRŽAJI NA PODRUČJU FILOZOFIJE ZNANOSTI

U ovom poglavlju navodi se nekoliko online sadržaja čija je svrha učiniti nastavu fizike i filozofije znanosti zanimljivom, jednostavnom i interesantnom polaznicima, tj. učenicima ili studentima.

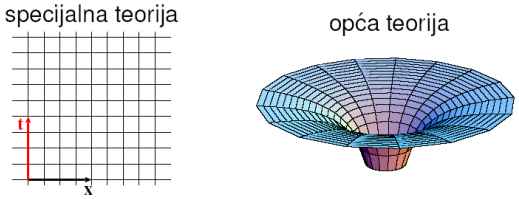
A. *Fizika i filozofija znanosti kroz web sadržaje*

Web stranica koja nudi virtualni pogled u povijest fizike zanimljiv je i privlačan interaktivni sadržaj koji treba omogućiti učenicima lakše savladavanje gradiva iz fizike i filozofije vezanog uz područje filozofije znanosti i prirode. Web adresa na kojoj možemo naći navedeni sadržaj je:
http://eskola.hfd.hr/inter_fizika/inter_fiz.htm. Stranica nosi ime *Povijest fizike* i sastoji se od nekoliko glavnih linkova koji su podijeljenih u nekoliko kategorija. Na stranici se mogu pronaći i razne simulacije eksperimenata, slikovni i video zapisi s područja prirodnih znanosti.



Slika 3. Prikaz web stranice *Povijest fizike* s web sadržajem

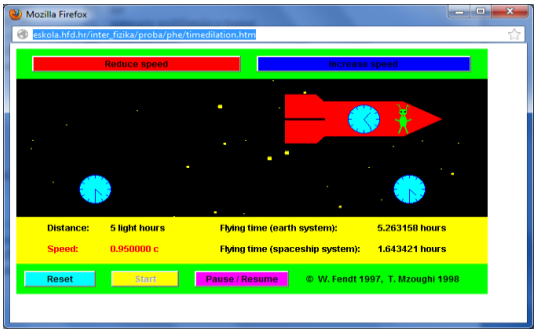
Grafički primjer prikazan na *Slici 4.* trebao bi pojednostaviti shvaćanje pojma zakrivljenog prostora. Kao što je poznato, ravan prostor je onaj u kojem vrijedi Euklidska geometrija npr. Pitagorin poučak, a kao primjer može poslužiti ravna površina stola ili papira. Zakrivljeni prostora je sfera tj. površina lopte. Slikoviti prikaz kako materija zakrivljuje prostor i vrijeme u ovom slučaju izgledao bi kao na *Slici 4.*



Slika 4. Geometrijski prikaz specijalne i opće teorije relativnosti

Ovakve pojednostavljene slikovne simulacije mogu se izraditi jedino korištenjem modernih web alata. Neke od takvih alata možemo pronaći na web stranicama koje se bave problematikom fizike, filozofije prirode tj. filozofije znanosti i astronomije. Stranica na kojoj se nalaze programi potrebni za izradu sličnog primjera zove se *Interaktivna fizika*, a web adresa joj je:
http://eskola.hfd.hr/inter_fizika/inter_fiz.htm.

Na navedenoj se stranici mogu pronaći interaktivne animacije iz svijeta fizike, astronomije, razni pokusi, besplatni alati, virtualni laboratoriji, zadaci i testovi i sl. (Slika 5.).



Slika 5. Prikaz dilatacije vremena kroz web aplikaciju

B. *Program OrbitXplorer*

OrbitXplorer je program pogodan za korištenje u nastavi srednjih škola i fakulteta, prikladan je za studente, nastavnike i svakoga tko voli fiziku i astronomiju. Korištenjem ovoga programa možemo izrađivati i simulirati orbite svih poznatih planeta, izračunati gravitacijska svojstva, definirati položaje zvijezda i ostalih svemirskih tijela. Sve simulacije popraćene su uzbudljivim i dinamičkim grafičkim prikazima stvarnog stanja svemirskih tijela.

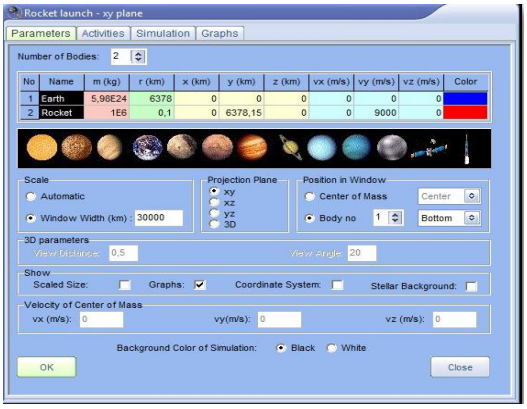
OrbitXplorer dolazi s 30 posebnih simulacija koje se mogu mijenjati i prilagođavati. Svaka simulacija uključuje jednu ili više predloženih aktivnosti i nudi rješavanje različitih poteškoća. Program dopušta i izgradnju kao i dizajniranje vlastite simulacije.

Neke su od važnijih značajki programa:

- izračunavanje orbita planeta, utjecaja gravitacijske sile na masu,
- jednostavno stvaranje i spremanje vlastitih simulacija,
- korištenje stvarnih slika planeta i svemira,
- projekti orbita svemirskih tijela u 3D pogledu,
- mogućnost pokretanja do četiri simulacije istovremeno,
- stvaranje grafova, dijagrama udaljenosti, postignute brzine, potrošene energije i dr.

Neke su od simulacija koje nudi program:

- Keplerov prvi, drugi i treći zakon planetarnih gibanja,
- simulacije slijetanja na Mars i Mjesec,
- lansiranje raketa u Zemljinu orbitu,
- putanje planeta,
- gravitacijski utjecaji na svemirska tijela,
- simulacije putanja meteorita i kometa i dr.



Slika 6. Prikaz izrade simulacije u web aplikaciji *OrbitXplorer*

V. PRIMJENA WEB SADRŽAJA I WEB ALATA NA PODRUČJU FILOZOFIJE ZNANOSTI

Primjetno je da se sve više profesora i predavača u nastavi koristi web sadržajima i web alatima kako bi učenicima i slušačima svoja predavanja učinili zanimljivijima, prihvatljivijima i primjenjivijima u današnjem svijetu moderne tehnologije. Kako bih dokazala da je gornja tvrdnja istinita, provela sam među učenicima istraživanje u obliku ankete.

Hipoteza 1: Učenici će pozitivno reagirati na primjenu web sadržaja i web alata u nastavi (dokazati kroz anketni upitnik).

Hipoteza 2: Učenici četvrtih razreda bit će puno više zainteresirani za korištenje web sadržaja i web alata u nastavi nego li učenici drugih razreda.

A. *Anketni upitnik*

ANKETNI UPITNIK

Škola:_____
Razred:_____
Spol: _____ M/Ž

Cilj je ovoga istraživanja utvrditi mišljenje učenika o internetu i web sadržajima i web alatima na području fizike i filozofije znanosti.

1. Koliko vremena na dan koristite internet?
   a) manje od dva sat na dan
   b) četiri sata
   c) više od šest sati_____ (napiši koliko)
   d) ne koristim ga uopće
2. Internet koristim:
   a) za rješavanja svakodnevnih školskih obaveza
   b) u edukativne svrhe
   c) korištenje društvenih mreža i igranje igrica
   d) _____

3. U školi bismo na nastavi trebali koristiti internet kako bi predavanja bila interaktivna i zanimljivija:
   a) da
   b) ne

c) ponekad (napiši zašto)_____

4. Mislite li da su web sadržaji koje možete naći na internetu korisni za vaše obrazovanje?

a) da
b) ne
c) ne znam

5. Smatrate li da su web sadržaji ono što vas je potaknulo da zavolite neki od predmeta u školi?

a) da
b) ne
c) ne znam

6. U kojim je predmetima, po vašem mišljenju, neophodno korištenje web sadržaja i web alata u nastavi:

a) fizici
b) filozofiji
c) geografiji
d) biologiji
e) nekom drugom predmetu (navedi kojem) _____

7. Navedi dva predmeta u kojima bi se trebali koristiti web sadržaji i web alati u nastavi, i napiši zašto:

_____
_____
_____

8. Smatrate li da će primjena web sadržaja i web alata poboljšati vaše znanje iz nekog od predmeta:

a) da
b) ne
c) ne znam

9. Po vašem su mišljenju web sadržaji u nastavi filozofije na području filozofije znanosti bili:

a) zanimljivi i poučni
b) nezanimljivi i nepotrebni
c) vrlo učinkoviti, potaknuli su me na daljnje istraživanje sličnih web sadržaja i iz drugih predmeta

10. Što vam se najviše svidjelo prilikom primjene web sadržaja u nastavi

a) eksperimenti i simulacije stvarnih događanja
b) interaktivnost sadržaja i povratnih informacija
c) ništa od navedenog
d) posebno su mi bili zanimljivi sljedeći sadržaji (napiši koji)

_____

11. Napišite koji biste web sadržaj ili web alat voljeli koristiti u nastavi:

_____

12. Navedite u kojim predmetima biste koristili navedeni web sadržaj ili web alat

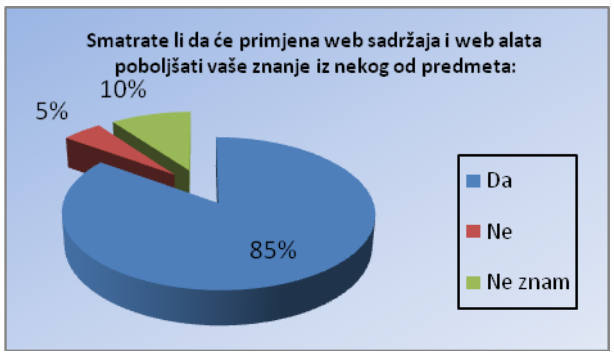
_____

Zahvaljujem.

B. Rezultati anketnog upitnika

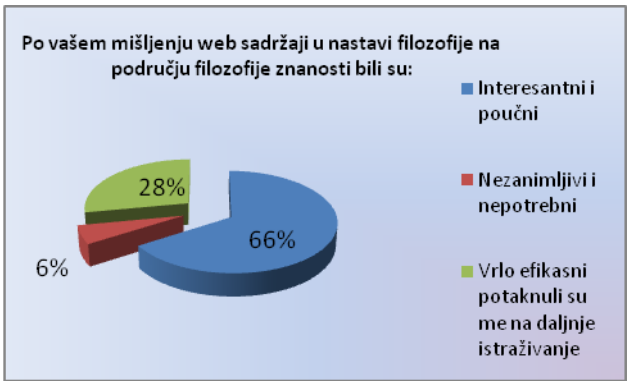
Koristeći anketni upitnik provela sam istraživanje na uzorku od 122 učenika. Ženskih je ispitanika bilo 84, a muških 38. Broj ženskih ispitanika u četvrtim razredima bio je 40, a broj muških ispitanika 20. Ukupan broj ispitanika za oba spola u četvrtim razredima iznosio je 60. Broj ženskih ispitanika u drugim razredima iznosio je 44, a broj muških ispitanika bio je 18. Ukupan broj ispitanika oba spola u drugim razredima je 62. (Slika 9.).

Obradom podataka i posebnim osvrtom na neka od postavljenih pitanja dobiveni su idući rezultati (Slike 7.). Većina je ispitanika na pitanje "Smatrate li da će web sadržaji i web alati poboljšati vaše znanje iz nekog predmeta" odgovorila potvrdno, njih 85%, 10% učenika nema nikakvo mišljenje, a samo 5% misli da im web sadržaji i web alati nisu potrebni u nastavi radi usavršavanja njihovoga znanja.



Slika 7. Prikaz rezultat anketnog upitnika (pitanje broj 8)

Sljedeći rezultati koje sam izdvojila i grafički prikazala vezani su uz pitanje broj 9. "Po vašem su mišljenju web sadržaji u nastavi filozofije na području filozofije znanosti bili /.../" (na ovo su pitanje odgovarali samo učenici četvrtih razreda): cca 66% učenika ogovorilo je da su im web sadržaji u nastavi filozofije na području filozofije znanosti bili interesantni i poučni. Ostatak učenika, njih oko 28%, odgovorilo je da su im web sadržaji bili vrlo zanimljivi i potaknuli su ih na daljnje istraživanje, a samo je 6% posto učenika odgovorilo da su im web sadržaji bili nezanimljivi i nepotrebni (Slika 8).

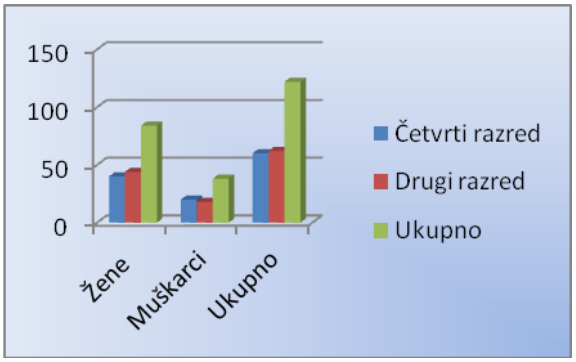


Slika 8. Prikaz rezultata anketnog upitnika (pitanje broj 9)

Usporedbom navedenih podataka (Tablica 1.) i (Slike 9.) dobili smo i potvrdni odgovor na postavljenu prvu hipotezu da će učenici i četvrtih i drugih razreda pozitivno reagirati na primjenu web sadržaja i web alata u nastavi. Slika 9. grafički prikazuje odnos ispitanika po spolovima u četvrtim i drugim razredima.

Tablica I. Prikaz broja ispitanika po razredima i spolu

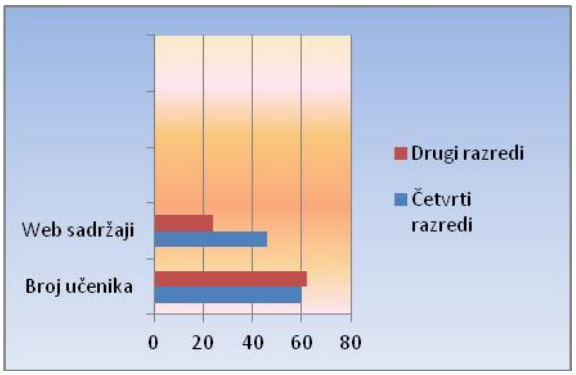
	Žene	Muškarci	Ukupno
Četvrti razred	**40**	**20**	**60**
Drugi razred	**44**	**18**	**62**
Ukupno	**84**	**38**	**122**



Slika 9. Grafički prikaz broja ispitanika po razredima i spolu

Iz prikazanoga je vidljivo da su učenici četvrtih razreda, u skladu s očekivanjima, bili puno više zainteresirani za primjenu web sadržaja i web alata u nastavi, te daljnje proučavanje sadržaja. Ovim je zaključkom dokazana i vjerodostojnost druge hipoteze .

Broj učenika u četvrtim razredima zainteresiranih za web sadržaje iznosi 46 (od ukupno 60 ispitanika), a broj učenika u drugim razredima zainteresiranih za web sadržaje iznosi 24 (od ukupno 62 ispitanika). (Slika 10.).



Slika 10. Zainteresirani učenici za web sadržaje po razredima

VI. ZAKLJUČAK

Upotreba web alata i web sadržaja u obrazovnom procesu predstavlja efikasan i vrlo efektivan način učenja. Ovaj način učenja potaknuo je ljude na stvaranje novih kreativnih ideja i to ponajviše korištenjem informacijske i komunikacijske tehnologije koja je postala sastavni dio društvenih zbivanja na globalnoj razini. Pojava CMS sustava pridonijela je razvoju kvalitetnijeg weba i stvorila nov način prezentiranja određenih obrazovnih sadržaja.

Einstein je jednom prilikom izjavio: *".... Pitati za napredak je besmisleno, znati odgovor na to znači biti religiozan ... "*. Tražiti nova znanja i usavršavati ih za stvaralački nastrojenog pojedinca znači nadograditi staro i otkriti novo. Einstein je vjerovao da filozofija može reći nešto o znanosti samo onda kada uspije doći do spoznaje osnovnih zakona koji vladaju prirodnim pojavama. Prema tome, nema apsolutne istine, savršeno ispravne metode, a naša spoznaja vanjskih predmeta, stvari, svakodnevno doseže višu razinu. Od istine koja ima za cilj samu sebe nema koristi. Ona mora biti definirana nekim zakonom, teorijom, kako bi bila stvarna. Ograničiti spoznaju nekim apsolutnim zakonom bilo bi nemoguće.

Filozof Spinoza je jednom prilikom rekao: *" .... Dok mislim, osjećam se slobodnim ... "*. Vođen ovom izjavom novovjekovnog filozofa današnji čovjek trebao bi i morao znati iskoristiti svoja znanja i mogućnosti koje mu pruža moderno informacijsko doba.

LITERATURA

[1] *Mreža*. Informatički časopis za IT profesionalce, broj 5, godina XVII. svibanj 2012. Tema broja: OpenSource CMS.

[2] *Mreža.* Informatički časopis za IT profesionalce, broj 10, godina XVII. listopad 2012. Tema broja: IT tržište.

[3] Mateljan, Širanović. Načela oblikovanja edukativnog multimedijskog sadržaja u online sinkronom Web okruženju. INFuture2007., Zagreb. Članak preuzet s web stranice: http://infoz.ffzg.hr/INFuture/2007/pdf/7-03%20Mateljan%20&%20Siranovic%20&%20Siranovic,%20Nacela%20oblikovanja%20edukativnog%20multimedijskog%20sadrzaja.pdf; (9.1.2013.).

[4] N. Selwyn. Education 2.0? Designing the web for teaching and learning. 2008. Članak preuzet s web stranice :www.tlrp.org/pub/documents/TELcomm.pdf; (21.1.2013.).

[5] Open Source sustava za upravljanje web sadržajima. Sadržaj preuzet s web stranice: http://www.crojoomla.com/; (24.1.2013.).

[6] Orehovački, Tihomir; Konecki, Mario; Stapić, Zlatko. Primjena Web 2.0 tehnologije u poslovanju. Članak preuzet s web stranice: http://student.foi.hr/~zstapic/dwnld/radovi/strucni/2008/Web%202.0.pdf; (18.1.2013.).

[7] Slavić, A. Semantički Web, sustavi za organizaciju znanja i mrežni standardi. U: Informacijske znanosti u procesu promjena (ur. J. Lasić-Lazić). Zagreb: Zavod za informacijske studije, 2005.

[8] Uzelac, Aleksandar. Utjecaj novih informacijskih tehnologija na kulturni razvoj: uloga virtualnih mreža. Sveučilište u Zagrebu, doktorska disertacija, Zagreb 2003. Sadržaj preuzet s web stranice: http://www.imo.hr/~sandra/docs/uzelac_doktorat.pdf ; (4.1.2013.).

Information literacy instruction: self-education through e-tutorials

Anita Papić*
* Faculty of humanities and social sciences/Department of information sciences, Osijek, Croatia
apapic@ffos.hr

Abstract - This paper gives theoretical insight into importance of information literacy in formal and informal education and self-education of individual in information society. Libraries traditionally have great role in rising information literacy and one way of rising information literacy could be through their e-tutorials about information literacy. Working hypothesis in this paper is that level of presence of e-tutorials about information literacy in Croatian language is not satisfying. This hypothesis is confirmed namely unsatisfying level of presence of e-tutorials about information literacy in Croatian language is determined. Since mentioned hypothesis is confirmed five e-tutorials about information literacy in English were chosen according to Primo (Peer-Reviewed Instructional Materials Online) criteria for identification and comparison of their elements with method of qualitative content analysis with main purpose to give directions for development of e-tutorials about information literacy in Croatian language.

I. INTRODUCTION

Information literacy is in strong connection to lifelong learning which is necessary in todays' information society. Information literacy is beacon of knowledge society and is recognized as necessary skill for life and work in 21. century. According to American Library Association - ALA „information literate person must be capable to recognize information need, to find, evaluate and effective use needed information "[1].

Information literacy is very important in information society where technology develops rapidly. Information environment develops rapidly and that's why individuals are confronted with information overload at work and at home. Information are available from different media and come quickly to individual in unfiltered form and that's why their authenticity, reliability and validity must be check out. Information literacy is common to all disciplines and all educational settings. Namely, information literacy helps individual to expand viewpoints and to take over control for learning.

Information literate person is capable to: access necessary information effectively and efficiently; evaluate critically information and its source; select information and transform them into his/her knowledge base; use information effectively in specific purpose and understand economic, legal and social problems that implicate use of information [1].

Models of information literacy instruction can be divided into metamodels and contextual models. Some of famous metamodels of information literacy instruction are: Kuhlthau model, model of collecting berries and relational model. Some of famous contextual models of information literacy instruction are: Big six skills, SCONUL model and TFPL model for information literacy instruction at work. In this paper special attention is given to Kuhlthau model because of its popularity and presence in online tutorials about information literacy.

Kuhlthau model is well known in libraries across the world and in information sciences. Kuhlthau model is based on theories from psychology, sociology and similar fields. Namely, it is based on many studies about information search process among different users especially students. Conclusions drawn from these studies describe cognitive, affective and psychomotor users' behaviors in every stage of information search process. That's why Kuhlthau model is important for understanding information needs and information behavior of users and it gives directions for information literacy instruction which should be implemented into online tutorials about information literacy [5]. Kuhlthau model consists of following six stages of information search process: introduction into task, topic selection, exploring wider field and research direction, defining focus, information gathering, search conclusion and starting writing. In these first three stages user is confused and uncertain and needs help. In next stages can be seen increment of user's self-confidence.

Information literacy is key for lifelong learning [7]. Namely, information literacy enables learning outside the classroom and that's why it is very important to develop information literacy instruction programs for example through online tutorials about information literacy. Libraries traditionally have great role in rising information literacy and one way of rising information literacy could be through their e-tutorials about information literacy.

Review of web sites of academic libraries in Croatia shown that level of presence of e-tutorials about information literacy in Croatian language is not satisfying. Namely, only web site of National and university library (NSK) consists one e-tutorial about information literacy named Virtual classroom [10]. Instructions of e-tutorial about information literacy in

Croatian language – Virtual classroom NSK (National and university library) are divided into eight parts: defining information problem, types and usage of information sources, finding information sources in libraries, searching strategies, online library catalog of NSK (National and university library), information from World Wide Web, information evaluation and citing literature. In first step instructions for defining fields and topics are given and in this step should be selected keywords which would help in next stages. In second step instructions about ways of searching, in regards to form of research for example retrospective, historical or actual, are given. Then information sources should be selected for example encyclopedia, peer review scientific journals etc. Third step of this e-tutorial deals with library literacy namely instructions how to find needed materials in library according to place in library or other criteria are given. In step about search strategies: basic and advanced searching, searching according to keywords, descriptors and thesauri are described. In fifth step instructions for use of online catalog NSK (National and university library) are given. Sixth step information evaluation encompasses evaluation of printed and electronic sources according to following criteria: purpose, authorship, content, coverage, actuality and usage. Last step points out to consistent styles of citing literature and avoiding plagiarism. This e-tutorial about information literacy is adequate for different users of different ages. Every step of e-tutorial is followed by short online quiz for users to check their information literacy.

Since the level of presence of e-tutorials about information literacy in Croatian language is not satisfying, in this paper five e-tutorials about information literacy in English were selected according to PRIMO (Peer-Reviewed Instructional Materials Online) criteria for the purpose of content analysis of their elements to draw directions about construction of similar e-tutorials in Croatian language.

For the purpose of conducting content analysis of e-tutorials about information literacy in English following research question were stated:

(Q1) Who are target users of e-tutorial?
(Q2) What is thought in e-tutorial (content)?
(Q3) What are the goals of learning through e-tutorial?
(Q4) How is information literacy defined in e-tutorial?
(Q5) How is information literacy evaluated in e-tutorial?

II. METHOD

For the purpose of conducting research in this paper scientific method of content analysis was selected. Content analysis is flexible research method and it can be qualitative or quantitative but also within it can be used combination of these two methods. Namely, in this paper was used qualitative content analysis of text within selected e-tutorials about information literacy to give answers to research questions stated in this paper and to put conclusions in wider context dealing information literacy instruction.

Content analysis is research method which objectively and systematically describes written, verbal or visual communication [2]. Krippendorff describes content analysis as research method for derivation of repeatable and valid conclusions from text within its context [4]. Content analysis helps to determine presence of certain words, phrases or concepts within text or collections of texts. Texts can be from books, speeches, interviews, essays, historical documents, scientific articles, pictures, web pages etc.

Content analysis can be implemented to any written or other form of human communication and that's why in recent decade it is often used in marketing, cultural studies, journalism, sociology, politics, phycology, cognitive science, economy, information sciences etc.

III. INTERPRETATION OF RESULTS

Five e-tutorials about information literacy in English are selected according to PRIMO (Peer-Reviewed Instructional Materials Online) criteria for the purpose of content analysis of their elements.

A. Mc Henry county college tutorial [6]

Students and university teachers are target users of this e-tutorial. First step is library map in which is very detail and virtual described position of materials, collections etc. Second step points out importance of determination of concrete information need of individual. Namely, it is important to have a plan and some questions for determining plan are suggested: what are you looking for?; which source would you use?; where would you search?; in which why would you search?; is needed information or source found? In next step importance of databases is described. Namely, database is defined as organized collection of computer records which have standard format designed for effective information retrieval. Databases have three main purposes: to organize, to archive and to retrieve information in effective way. Databases consist of archives which have their own fields and according to these fields information are searched. The most used search method is method according keywords. Furthermore, it is described how to create search strategy: identification of sources, determination of location of sources, transformation of topic into research questions, identification of keywords, identification of synonyms. Search techniques like limitation techniques, Boolean operators and phrase searching are described. Instructions for evaluation of search results are given according to following criteria: content, intellectual property, correctness, importance, updated information and bias.

At the end, term of plagiarism is described and the best ways for avoiding plagiarism are suggested namely the best ways of citing literature are suggested.

The goal of this e-tutorial is to guide users step by step through process of exploring library and to introduce users with possibilities that library provides.

In this e-tutorial information literate person is described as person who possesses knowledge and capabilities to cope with all life challenges.

Information literacy in this e-tutorial can be checked through short quiz similar to famous „Who wants to be a millionaire? ".

B. Washington state university: teaching strategies [11]

This e-tutorial does not consist of same concepts like other e-tutorials and its content is adequate for wider population. This e-tutorial describes reasons of plagiarism among students and special attention is given to teachers who have a great role in prevention of plagiarism. Content of this e-tutorial is oriented to citing literature and negativity of plagiarism that is usually the last step in other e-tutorials. Although this e-tutorial does not consist definition of information literacy its content stimulates elements of information literacy. At the end there is a quiz which covers content of e-tutorial.

C. TILT - the information literacy tutorial [8]

Students are target population of this e-tutorial. E-tutorial is divided into three parts: introduction and three modules. Every of these three modules consist of short overview, list of key concepts and skills for users, main text and short quiz. Information are available from books and journals as well as from library online databases. First module discusses differences between information sources including instructions where to find them and in which way to select the best sources for specific search. Second module is focused to performances of effective searching of library online databases. Third module is focused to criteria of evaluation of credibility of sources and how to cite printed or online literature. E-tutorial is designed to introduce users with possibilities and performances of searching. Within e-tutorial information literacy is defined. At the end of e-tutorial there is an online quiz.

D. Griffin tutorial [3]

This e-tutorial is created for students. This e-tutorial differs from other e-tutorials because it does not consist words namely for every part of this e-tutorial is recorded video with instructions. For each format in which we can find information search criteria is given for example according to: author, title, keywords, topic etc. Griffin tutorial points out how to search books, journals, newspapers, audio and video data, archives, special collections, manuscripts etc. There is no explicit definition of information literacy. At the end quiz for knowledge check and information literacy check is given. Quiz consists of classical theoretical questions but also demands implementation of knowledge through examples.

E. UCSC net trail: information literacy tutorial [9]

Target population is students. This e-tutorial consists of six parts. In first part is given introduction into different types of information sources and in second part instructions how to select topic and which are advantages of spreading or summarizing the topic. Third part is focused to Internet and libraries and their comparison in regards to information sources. Fourth part is about CRUZCAT, online catalog of their library where instructions for its use are given. Fifth part is dedicated to knowledge transfer. Last part is focused to information ethics and citing literature. This e-tutorial helps to learn how to find needed information, how to evaluate and use information. There is no definition of information literacy. At the end quiz which covers content of e-tutorial is given.

Tab.1 shows comparison of e-tutorials about information literacy: similarities and differences according to conducted content analysis of e-tutorials about information literacy in English. It can be concluded that following common elements dominated across e-tutorials about information literacy: (a) defining topic, (b) creating query, (c) evaluation of results, (d) correct citation of literature, (e) avoiding plagiarism and (f) information literacy check.

TABLE 1. COMPARISON OF E-TUTORIALS ABOUT INFORMATION LITERACY: SIMILARITIES AND DIFFERENCES

E-TUTORIAL	defining topic	creating query	evaluation of results	correct citation of literature	avoiding plagiarism	information literacy check
MC HENRY COUNTY COLLEGE	+	+	+	+	+	+
WASHINGTON STATE UNIVERSITY: TEACHING STRATEGIES	-	-	-	-	+	+
TILT- THE INFORMATION LITERACY TUTORIAL	+	+	+	+	+	+
GRIFFIN TUTORIAL	+	+	+	+	-	+
UCSC NET TRAIL: INFORMATION LITERACY TUTORIAL	+	+	+	+	-	+

Defining topic is usually first step of e-tutorials about information literacy, creating query is usually second step of e-tutorials about information literacy, evaluation of results is usually third step of e-tutorials about information literacy and correct citation of literature is usually fourth step of e-tutorials about information literacy. Namely, Tab. 1 shows that these four steps contain almost all e-tutorials about information literacy explored in this research except one e-tutorial - Washington state university: teaching strategies.

Very important step of e- tutorials about information literacy named avoiding plagiarism consist more than half explored e-tutorials about information literacy in this research. Namely, two of five e-tutorials about information literacy do not contain this step about negativity of plagiarism and these e-tutorials are Griffin tutorial and UCSC net trail: information literacy tutorial.

Information literacy check is usually the last step of e-tutorials about information literacy and Tab. 1 shows that all e-tutorials explored in this research contain this step.

IV. CONCLUSION

Information literacy is key for lifelong learning. Namely, information literacy enables learning outside the classroom and that's why it is very important to develop information literacy instruction programs for example through online tutorials about information literacy.

Libraries traditionally have great role in rising information literacy and one way of rising information literacy could be through their e-tutorials about information literacy.

Since the level of presence of e-tutorials about information literacy in Croatian language is not satisfying five e-tutorials about information literacy in English were chosen according to Primo (Peer-Reviewed Instructional Materials Online) criteria for identification and comparison of their elements with method of qualitative content analysis with main purpose to give directions for development of e-tutorials about information literacy in Croatian language.

According to conducted content analysis of e-tutorials about information literacy in English it can be concluded that following common elements dominated across e-tutorials about information literacy: (a) defining topic, (b) creating query, (c) evaluation of results, (d) correct citation of literature, (e) avoiding plagiarism and (f) information literacy check. These elements should contain all e-tutorials developed in purpose of information literacy instruction.

REFERENCES

[1] American Library Association. President's Committee on Information Literacy report, 1989. http://www.ala.org/ala/mgrps/divs/acrl/publications/whitepapers/presidential.cfm (2013-02-02)

[2] B. Barelson, Content analysis in communication research. New York : Free Press, 1952.

[3] Griffin tutorial. http://www.wsulibs.wsu.edu/electric/trainingmods/new_griffin_tutorial/ (2013-02-02)

[4] K. Krippendorff, Content analysis: an introduction to its methodology. Thousand Oaks : London : New Delhi: Sage Publications, 2004.

[5] C. C. Kuhlthau, Seeking meaning: A process approach to library and information services, Westport: Libraries Unlimited, 2004.

[6] Mc Henry county college. http://www.mchenry.edu/library/tutorial/starttutorial.htm (2013-02-02)

[7] S. Špiranec, Informacijska pismenost – ključ za cjeloživotno učenje, 2003. http://edupoint.carnet.hr/casopis/17/clanci/1 (2013-02-02)

[8] Tilt- The information literacy tutorial. http://www.wsulibs.wsu.edu/electric/trainingmods/tilt/nf/ (2013-02-02)

[9] UCSC Net trial: Information literacy tutorial. http://nettrail.ucsc.edu/ (2013-02-02)

[10] Virtualna učionica Nacionalne i sveučilišne knjižnice. http://www.nsk.hr/usluge/edukacija_korisnika/virtualna-ucionica/index.html (2013-02-02)

[11] Washington state university: Teaching strategies. http://www.wsulibs.wsu.edu/plagiarism/teaching1.html (2013-02-02)

ZAŠTITA PODATAKA OD PLAGIRANJA

prof. dr. sc. Slavko Šimundić, Danijel Barbarić, mag. iur.
Sveučilište u Splitu, Pravni fakultet Split, Katedra za informacijske znanosti i statistiku, Split, Croatia
slavko.simundic@pravst.hr, danijel.barbaric@pravst.hr

Sažetak - U današnje vrijeme neophodno je posebnu pažnju posvetiti tehnologiji pomoću koje se možemo boriti protiv varanja i zaštiti od plagiranja. Novim tehnologijama se možemo boriti od plagiranja na razne načine. S obzirom da je podatak kojeg smo samostalno „napravili“ zaštićen kaznenim zakonom kao naše autorsko djelo s jedne strana, isti bi trebalo zaštiti i kriptozaštitom s druge strane. Sigurno je da se opasnosti koje prijete od varanja i plagiranja, kao i težina njihovih posljedica, mijenjaju od slučaja do slučaja i od vremena do vremena, jer se i okolnosti vremenom mijenjaju. S obzirom na veoma mali postotak diverzija u odnosu na sve druge elemente, nije potrebno znatnije razvijati postojeće fizičke i organizacione mjere zaštite, već pažnju treba posvetiti grješkama i neovlaštenom pristupu podatcima. Na samom početku važno je ustvrditi gdje je nastao problem, a potom koji je najbolji način za rješavanje istog. Istraživanje valja usmjeriti na hardwer/softwer i na komunikacijsku zaštitu, a posebnu pažnju treba posvetiti uspješnosti te zaštite.

1. UVOD

Podatke je moguće zaštititi od plagiranja na razne načine. Klasični način čuvanja podataka po arhivima, kartotekama, dosjeima i sl., vezan je uglavnom za fizičku zaštitu prostora u kojem se ti nosioci podataka nalaze. Taj način zaštite podataka star je koliko i povijest njihova prikupljanja, zbog čega nam je pojmovno blizak i poznat. U takvom tzv. klasičnom tj. starom načinu čuvanja podataka,

podatke smo štitili od plagiranja uglavnom organizacijskom, administrativnom i fizičkom zaštititom. Svakako da nije primjerno dopuštati bilo kojoj osobi da se samostalno koristi nekavnim podatcima bez službenog nadzora. Takva osoba lako može izmijeniti, uništiti ili otuđiti bitne podatke.

Podatci se štite i zakonom. Zaštita autorskog prava i srodnih prava osigurava se pomoću Zakona o autorskom pravu i srodnim pravima (Narodne novine br. 167/03, 79/07 i 80/11) te pomoću odredaba Kaznenog zakona (Narodne novine br. 125/11 i 144/12). Kazneni zakon je posvetio cijeli jedan svoj dio (glava XXVII.) zaštiti intelektualnog vlasništva. Za povredu autorskih prava predviđene su čak i zatvorske kazne.

Novo vrijeme donosi i nove tehnologije. Tako nam elektronička obrada donosi sa sobom mnogo novog, što bitno mijenja ili proširuje tradicionalni pojam zaštite podataka od plagiranja. Mogućnost neovlaštenog korištenja, izmjene sadržaja, krađe ili uništenja podataka, poprima novu dimenziju, jer više ne zahtijeva fizički pristup do prostora gdje se oni čuvaju. Iz tih razloga razvijene su brojne i učinkovite metode zaštite podataka, koje sustavu za elektroničku obradu daju veći stupanj pouzdanosti i zaštite podataka nego što su ih imali u klasičnim načinima.

Svaki naš podatak koji se nalazi bilo gdje, potrebno je zaštititi; u protivnom sami sebi dopuštamo da naš podatak bude plagiran od strane drugih osoba.

2. METODE I POSTUPCI ZAŠTITE PODATAKA OD PLAGIRANJA

Danas je već u zadovoljajućoj mjeri razvijena mogućnost za zaštitu podataka od plagiranja. Pažnju prije svega treba posvetiti pogreškama koje se (eventualno) događaju, te neovlaštenom pristupu podatcima. Potrebno je utvrditi koja vrsta zaštite je najuspješnija.

Kod zaštite podataka od plagiranja veliku pažnju trebamo posvetiti komunikacijskoj zaštiti tj. zaštiti koja se koristi za zaštitu podataka pri prijenosu komunikacijskom linijom tj. spojnim putem. Osnovu komunikacijske zaštite čini kriptografija. Kriptografija je postupak zaštite tajnosti informacije primjenom algoritama koji transformiraju razumljivi (čisti) oblik teksta u šifrirani tekst, nerazumljiv osobama koje ne poznaju algoritam transformacije. Primarna namjena kriptografije je očuvanje tajnosti informacija koje se prenose između računala i terminala. Kriptografija također osigurava i autentičnost poruka. Šifrirana poruka može uključivati informaciju koja pokazuje da je ona stigla od navedenog pošiljaoca i da nije izmijenjena. Kriptografija se danas sve više koristi i za zaštitu podataka u datotekama i bazama podataka. Kriptografija ima povijesni razvoj i praksu. Svaki put kad ljudi nisu bili sigurni da važna (i tajna) poruka neće pasti u pogrješne ruke, koristili bi kriptografiju. Danas se koriste različite kriptografske metode. Nezavisno od pojedine metode u osnovi se radi o sljedećem: koriste se dva čitača papirne trake, jedan čita papirnu traku s “čistim” podatcima, a drugi čita papirnu traku s “krivim” podatcima. “Krivi podatci” se “dodaju” u “čiste” podatke i tako nastali podatci šalju se direktno preko telegrafske linije. Onaj koji prima poruku ima traku s istim “krivim” podatcima, koje “oduzme” od primljene poruke i dobije “čiste” podatke. Takvu vrstu zaštite moguće je ostvariti pomoću hardwera i softwera, izbor ostaje korisniku

koji sam treba odrediti koja razina zaštite mu je potrebna, a to će najčešće ovisiti o važnosti podataka i informacija koje treba zaštititi.

Potrebno je postaviti razliku između pojma zaštita podataka i pojma sigurnost podataka. Pod zaštitu podataka spadalo bi sprječavanje pristupa i/ili korištenja podataka od neovlaštenih osoba. Pod pojmom sigurnost podataka podrazumjevamo postizanje same sigurnosti podataka, tako da isti ne budu oštećeni, promjenjeni ili nepovratno izgubljeni uslijed problema u radu računalnog sustava, elementarnih nepogoda, propusta u radu i sl.
Metode zaštite podataka trebalo bi primjenjivati na svim područjima, na kojima se proteže mreža računala uključenih u cjelovit sustav i na svim mjestima gdje taj sustav može biti "provaljen". Gruba shema povezivanja računala s terminalnim stanicama (ili s drugim računalom) dana je na slici.

Slika 1. Metode zaštite podataka

2.1. SOFTWERSKA ZAŠTITA PODATAKA OD VARANJA U OBRAZOVANJU

Ubrzani razvoj tehnologije donio je i softwerska rješenja uz pomoć kojih se možemo zaštiti od varanja (prvenstveno pri provjeri znanja) u obrazovanju. Računala u obrazovanju su naša svakodnevnica. S obzirom da današnje masovno obrazovanje i ispitivanje uz pomoć računala ovlaštenom ispitivaču ili nastavniku naprosto nedaje mogućnost da fizički ima kontolu nad svim računalima u predavaonici., u nastavni proces neophodno je uvesti softwer kako bi olakšalo samo obrazovanje, podigla kvalitata i napravilla kvalitetna zaštita od varanja. Taj problem rješava upravo softwer. Umrežavanjem računala i korištenjem odgovarajućeg softwera nastavnik može imati kontolu nad svim računalima u informatičkoj učionici. Može s jednog računala upravljati svim ostalim računalima, s jednog računala pratiti rad na svim ostalim računalima, dopuštati i/ili uskraćivati korištenje aplikacija nad drugim računalima i sl.
Kada nastavnik ima određenu razinu kontrole nad drugim računima točno može znati što se na kojem računalu radi i tko vrši radnje. Softwerska rješenja u jednom velikom dijelu pomažu ublažavanju posljedice varanja u obrazovanju. No, to nije dovoljno. Uvijek postoji mogućnost da određeni pojedinac ima velika znanja na području informacijske tehnologije, te može pronaći način kako onesposobiti softwer i uspješno varati na području obrazovanja. Zato je potrebno zajedničkom suradnjom stručnjaka iz obrazovanja, informacijskih znanosti i prava naći odgovarajuće rješenje kako bi se uspješno doskočilo rješavanju problematike varanja u obrazovanju.

3. PRAVNA ZAŠTITA PODATAKA

Masovno korištenje suvremene informacijske tehnologije dovelo je do toga da podatci postanu dostupni velikom broju osoba, što za sobom povlači niz opasnosti po sigurnost, pouzdanost i potpunost podataka i informacija pohranjenih u takvim sustavima. Stoga i ne začuđuje da su zakonodavne aktivnosti na tom planu prvenstveno bile usmjerene na pitanja pravne zaštite podataka. Pitanje zaštite podataka u uskoj je svezi s naglim razvojem informatičke tehnologije i komunikacija, te stvaranjem velikog broja različitih podataka koji sadrže čitav niz informacija o pojedincima, kao i informacija s različitih područja ljudske djelatnosti (trgovine, prava, prometa, financija, i dr.).

Povezivanje informatičke tehnologije i komunikacija dovelo je do naglog porasta međunarodnog prometa podataka. Konkurencija između kompanija koje proizvode računalnu opremu ili pružaju razne druge informatičke usluge učinila je državne granice gotovo nevažnim. Podatci i informacije mogu se prenositi brzo i jeftino putem telefonskih i satelitskih veza bez obzira na udaljenost.

Opasnost da pojedinci ili privatne, odnosno javne institucije zlouporabe te informacije je velika, tim više što ne postoji apsolutna zaštita takvih podataka. Brzi razvoj informatičke tehnologije samo uvećava takvu opasnost, jer ona sniženjem cijena postaje dostupna velikom broju ljudi.

Opasnost od narušavanja privatnosti podataka može doći sa svih strana (privatnih osoba, konkurencije, tajnih službi, inozemstva itd.). Privatnost naših podataka mogu narušiti stručnjaci, ali i učenici ili studenti koji se slučajno i iz puke znatiželje mogu naći u sustavu i doći do povjerljivih informacija.

3.1. VRSTE ZAŠTITE PODATAKA I INFORMACIJA

Zaštita podataka i informacija koje se obrađuju primjenom informatičke tehnologije, ovisi o vrsti zloporabe, odnosno nepravilnosti koja se, ako se ne osigura zaštita, može dogoditi. Tako se, mora osigurati fizička zaštita sustava. Tu se podrazumijeva čuvanje opreme od krađe i uništenja, osiguranje stalnih stabilnih izvora električne energije. Predviđaju se mjere u slučaju elementarnih nepogoda (potresa, poplava, požara), ratnog stanja i drugih situacija, u kojima postoji mogućnost da se neovlaštena osoba koristi podatcima,

odnosno informacijama. Sustav fizičke zaštite mora biti osiguran i od špijunaže, i od krađe podataka i sl. Drugačija je zaštita PC-a, drugačija velikih sustava. Fizička zaštita kombinira se s administrativno-organizacijskom zaštitom, tako da se propisuju norme ponašanja osoblja koje radi na računalnom sustavu i oko njega. Rad na sustavu valja organizirati tako da se u svakom trenutku zna tko je odgovorna osoba, a za svako ključno mjesto u sustavu mora biti organizirana moguća zamjena, kako izvođenje rada na sustavu ne bi došlo u pitanje.

3. 1. 1. ZAŠTITA PODATAKA U SUSTAVU GDJE SE NALAZE PODACI

Prije nego korisnik počne unositi podatke valja izgraditi sustav identifikacije korisnika, odnosno, utvrđivanje ima li korisnik ovlaštenje da pristupi tim podatcima. Na taj način možemo neovlaštene osobe spriječiti da plagiraju naše podatke. Također je potrebno paziti da se prilikom izgradnje sustava vodilo računa o fazi programiranja, fazi kreiranja sustava, fazama obrade i sl.

3. 1. 2. KOMUNIKACIJSKA ZAŠTITA

Za zaštitu LAN, WAN i velikih sustava koji imaju terminale i računala na različitim lokacijama, koriste se sredstva komunikacijske zaštite. Postoje različite metode zaštite, od kojih su najčešće: kriptografska zaštita, korištenje specijalnih uređaja za promjenu signala (scrambling), korištenje pametnih kartica i dr.

3. 1. 4. ZAŠTITA REZULTATA OBRADE PODATAKA I INFORMACIJA

U toj fazi razvijene su metode zaštite, koje onemogućuju neovlašteno korištenje na temelju osiguranja tajnosti podataka, ako podatci predstavljaju službenu tajnu. Međutim, u toj fazi moguće su i druge zloporabe. Tu se često puta radi o falsificiranju dokumenata pomoću suvremene informatičke opreme. Takav slučaj poznat je iz naše prakse (prije desetak godina), kada je službenica banke krivotvorila listiće za isplatu osobnih dohodaka u banci. Drugi je slučaj kad je službenica lažno povećavala radnicima svog poduzeća osobne dohotke. Razliku svote uplaćivala je u korist poduzeća ili banaka kod kojih je uzela kredit. Otkrivena je slučajno.

3. 1. 5. ZAŠTITA VEĆIH SUSTAVA – VATRENI ZIDOVI I DR.

Računala koja imaju pristup internetu potrebno je posebno zaštititi. Zaštita od neprijateljskih računalnih mreža ili pojedinaca vrlo je važna i predstavlja najveći izazov osobama koje štite sustav i osobama koje ga napadaju. Jedan od najboljih današnjih sustava za zaštitu računalnih mreža naziva se vatreni zid (Firewall). Koristi se kako bi sudionicima omogućio pristup internetu bez opasnosti za podatke na lokalnoj računalnoj mreži s koje se pristupa, kao i kada se želi zaštititi veza između javnog Web poslužitelja organizacije, dostupnog svima, i interne računalne mreže s kojom je povezan. Najjednostavniji način zaštite mreže organizacije je fizička izolacija ili cijele mreže ili pojedinog računala u mreži. Fizičkom izolacijom mreže onemogućuje se napadačima izvan organizacije pristup mreži, dok je za počinitelje unutar organizacije on i dalje otvoren. Ovo je vrlo jednostavan način zaštite, a koristi ga većina organizacija već godinama. U mnogim je okruženjima ovaj način još uvijek najbolji način mrežne sigurnosti, jer se malo može dobiti od veze izvan mreže, a može se jako puno izgubiti. Rast mreže svih mreža Interneta otežao je fizičku izolaciju organizacija. Naime, sve veći broj zaposlenika u organizacijama mora za svoj rad koristiti e-mail adrese, pretraživati WWW i sl.

Kako bi dopustili djelomično spajanje i pristup Internetu, a i zadržali i izolirali pojedine dijelove svoje organizacije, većina se korisnika odlučuje za Firewall kako bi zaštitila svoju sigurnost, odnosno kako bi osigurala podatke koji se nalaze na korisničkim računalima. Firewall predstavlja vrlo moćan alat kojega nikada ne valja koristiti umjesto nekih drugih sigurnosnih mjera, već ga valja koristiti kao dodatak tim sigurnosnim mjerama. Firewall organizacijama daje način da stvore sredinu između računalnih mreža koje su u potpunosti izolirane od eksternih računalnih mreža, kao što su Internet, i onih koje su u potpunosti spojene. Smješten između unutarnje i vanjske računalne mreže Firewall omogućuje jednostavan način kontrole količine i vrste podataka, koji će biti propušteni između tih mreža. Sam izraz Firewall je iz engleskog jezika, a upotrebljava se u građevinarstvu. Naime, kada se grade zgrade, one su često opremljene Firewall-ovima, specijalno konstruiranim zidovima otpornim na vatru. Ako dođe do požara u zgradi, on nekontrolirano može gorjeti u jednom dijelu, a Firewall zaustavlja ili usporava požar dok ne dođe pomoć. Ista filozofija se ne može upotrijebiti za zaštitu lokalnih računalnih mreža od vanjskih napada. Firewall može ograničiti količinu štete: uljez može provaliti u jedan skup računala, ali će Firewall zaštititi druga računala. Postavljen između računalne mreže organizacije i Interneta u cijelosti, Firewall sprječava napadača, koji je stekao kontrolu nad računalima izvan zidova organizacije, da se zadrži unutra. Osnovna funkcija Firewall-a jest ograničeni protok informacija između dviju računalnih mreža. Za konfiguraciju Firewall-a moraju se odrediti vrste podataka kojima je prolaz dopušten te one kojima nije.

Razlikujemo dva osnovna načina konfiguriranja Firewall:

1. Određena dozvola - postavlja se skupina uvjeta koji će rezultirati blokiranjem podataka. Svaki posjetitelj koji nije pokriven policom, bit će pušten.

2. Određena zabrana - upisuje se određeni protokol koji omogućuje prolaz, određeni posjetitelji mogu proći i biti u vezi, ostali su odbijeni.

Svaki od ovih načina posjeduje i prednosti i nedostatke. Osnovna prednost određene dozvole je to što je lakša za konfiguriranje: jednostavno se blokiraju protokoli koji su "preopasni". Kod određene zabrane: omogućavaju se protokoli

koje korisnici ili rukovoditelji traže. Svaki protokol, koji organizacija ne koristi, bit će blokiran. S obje police, moguće je stvoriti dva tipa Firewalla, siguran i nesiguran.

4. ZAKLJUČAK

Razvoj informacijske tehnologije uvelike je utjecao i na promjene ponašanja. Opasnost od plagiranja postoji svugdje. Danas je svaki podatak potrebno štititi. Ako podatke ne štitimo, ostvaljamo mogućnost da će isti biti plagirani. Potrebno je uvesti najmoderniju hardwersku i softwersku tehnologiju, te obavljati kontrolu i testiranje nad istom, kako bi se otklonili svi mogući pokušaji plagiranja. Također, uvijek je potrebno ići dalje tj. težiti boljim, uspješnijim i sigurnijim načinom zaštite podataka od plagiranja.

LITERATURA:

[1] Šimundić, Slavko; Pravna informatika, Sveučilište u Splitu, Pravni fakultet, Split, Split. 2007. god.

[2] Šimundić,Slavko; Žužuj Josip; Upravna informatika, Sveučilište u Splitu, Pravni fakultet, Split, Split. 2010. god.

[3] Šimundić,Slavko; Franjić, Siniša; Sveučilište u Splitu, Pravni fakultet, Split, Split. 2009. god.

[4] Bača, Miroslav; Uvod u računalnu sigurnost, Narodne novine, Zagreb, 2004. god.

[5] Zakon o autorskom pravu i srodnim pravima (Narodne novine br. 167/03, 79/07 i 80/11)

[6] Kazneni zakon (Narodne novine br. 125/11 i 144/1

Primjer uporabe dinamičkih struktura podataka u programskom jeziku Python

P. Brođanac *, L. Budin**, Z. Markučič *** , S. Perić****

*V. gimnazija, Zagreb, Hrvatska

** Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, Zagreb, Hrvatska

***XV. gimnazija, Zagreb, Hrvatska

****II. gimnazija, Zagreb, Hrvatska

{ predrag.brodjanac1@zg.t-com.hr; leo.budin@fer.hr; zmark@mioc.hr; smiljana.peric@gmail.com}

Sažetak - Ostvarenja apstraktnih struktura podataka kao što su: liste, redovi, stogovi, stabla i grafovi određuju efikasnost algoritama. Iako načini ostvarenja struktura ne određuju vremensku složenost, trajanje izvođenja algoritama jednake složenosti bitno je određeno trajanjem osnovnih operacija nad strukturama. Isto tako, način ostvarenja apstraktnih struktura može utjecati na prostornu složenost algoritama. U ovom radu prikazana je jedna od mogućih uporaba Pythonove strukture podataka rječnika (engl. Dictionary). Prikazani primjer je zanimljiv jer standardnu strukturu kao što je matrica implementira na jedan posve drugačiji, nestandardni način. Ugrađena struktura Dictionary ostvarena je vrlo efikasnim raspršenim adresiranjem (engl. hashing). Ovakav način implementacije struktura nije proširen jer u većini ostalih programskih jezika uporaba raspršenog adresiranja zahtijeva vrlo temeljito poznavanje navedene tehnike programiranja. Struktura Dictionary u programskom jeziku Python s prikladnim operacijama i metodama čini dobro okruženje za pripremu klasa kojima se ostvaruju spomenute strukture podataka.

I. UVOD

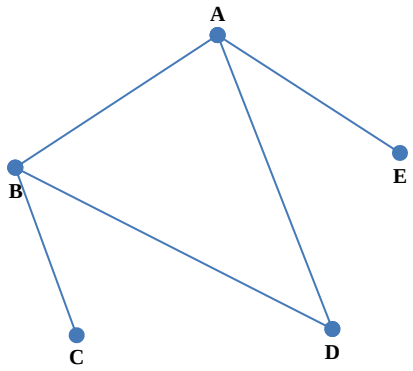
Pojavom programskog jezika Python mijenjaju se i neki pogledi na podučavanje programiranja. Jedna od velikih novina je interaktivno sučelje IDLE koji omogućava testiranje naredbi prije njihove implementacije u sam program. Među mnoštvom inovacija su i neke strukture podataka koje nisu postojale u nekim drugim jezicima koji su se standardno koristili za podučavanje programiranja (Pascal, BASIC, C,…). Jedna od takvih struktura jest rječnik (engl. Dictionary).

Nastavnici koji su programiranje podučavali u nekim ranijim programskim jezicima neće često upotrebljavati ovu strukturu jer će probleme moći rješavati koristeći njima poznatije strukture (primjerice: liste) koje su im nešto bliže, jer postoje i u ostalim programskim jezicima. Pokazalo se, međutim, da učenici, neopterećeni poznatim strukturama pristupaju rješavanju problema uporabom ove strukture mnogo slobodnije. Štoviše, oni njihovom uporabom demonstriraju i stanovite inovativne pristupe u rješavanju problema.

Za ilustraciju ovih navoda u nastavku ćemo uvesti klasu za rijetko popunjene matrice prikladnu za implementaciju matrice incidencije grafova.

II. TIP PODATAKA MATRICA

Matrica je kod programiranja vrlo česta struktura podataka, a najčešće implementira kao dvodimenzionalno polje. Jedna od čestih upotreba matrice je kod teorije



SLIKA 1 PRIMJER GRAFA

grafova [3]. Prisjetimo se, graf definiramo kao uređeni par G = (V, E), pri čemu je V skup vrhova dok je E skup bridova.

Na slici 1 dan je primjer grafa čiji je skup vrhova (V) jednak: {A, B, C, D, E}, dok je skup bridova (E) jednak {AB, AD, AE, BC, BD}. Jedan od načina prikaza grafa u računalu jest matrica incidencije. Matrica incidencije je kvadratna matrica dimenzije |V|, pri čemu je |V| broj vrhova grafa G. Primjerice dimenzija grafa prikazanog na slici 1 je 5 (graf ima 5 vrhova). Elemente matrice incidencije (a) definiramo na sljedeći način:

$$a_{i,j} = \begin{cases} 1, ako\ postoji\ brid\ između\ čvorova\ i\ i\ \ j \\ 0, ako\ ne\ postoji\ brid\ između\ čvorova\ i\ i\ j \end{cases}$$

Primjerice za graf sa slike 1 pripadna matrica incidencije imat će sljedeći oblik:

$$a = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Kao što možemo primijetiti u matrici je veliki broj elemenata prazan (elementi na kojima se nalaze nule). Što je graf veći to je broj praznih mjesta u matrici incidencije veći, a time je i veličina prostora koji matrica zauzima u memoriji veća. Ovo ne bi bio tako veliki problem da većina tog prostora nije neiskorišteno. Jedan od rješenja ovog problema jest prikazati graf u memoriji računala na neki drugi način, primjerice listom susjedstva. Takav način

prikaza nije praktičan za implementaciju nekih standardnih algoritama nad grafovima (broj putova duljine n između neka dva vrha, BFS, DFS i sl.) Drugi pristup jest da matricu incidencije implementiramo na neki drugi način tako da ona sadrži sve potrebne informacije a da zauzima minimalnu količinu prostora u memoriji računala. Uporabom rječnika dobit ćemo upravo takvu matricu.

III. RJEČNIK (DICTIONARY)

Za razliku od liste gdje su elementi indeksirani rednom brojevima: 0, 1, 2, … n – 1. Rječnik je struktura kod koje indeksi mogu biti bilo kojeg tipa i mogu ići bilo kojim redom. Indekse kod rječnika zovemo ključevima dok vrijednosti koje se nalaze na pripadnim mjestima u rječniku kao i kod liste zovemo jednostavno vrijednostima. Rječnik je struktura koja opisuje upravo ono što nam prvo padne na pamet kada kažemo rječnik, a to je rječnik nekog stranog jezika, primjerice englesko – hrvatski rječnik. Taj rječnik je najčešće koncipiran na način da je prvo naveden neki izraz na engleskom jeziku, a zatim jedan ili više pripadnih izraza na hrvatskom jeziku, primjerice: conjunction – spona, veza, spoj. U ovom slučaju bi conjunction bio ključ dok bi prijevodi te riječi na hrvatski jezik (spona, veza, spoj) predstavljali pripadnu vrijednost, koja bi u ovom slučaju mogla biti lista riječi (['spona', 'veza', 'spoj']).

Objekt (d) tipa rječnik kreirat ćemo naredbom r = dict() ili da jednostavno napišemo r = {}. Prilikom definicije moguće je inicijalizirati rječnik na neku početnu vrijednost. U tom slučaju elementi rječnika se međusobno odvajaju znakom zarez (,) te se ključ od vrijednosti odvaja znakom dvotočka (:).

A. *PRIMJER 1.*

Definirajmo rječnik koji će opisivati mjesece u godini, gdje će nazivi mjeseci biti ključevi dok će odgovarajuće vrijednosti biti broj dana za pripadni mjesec (pretpostavka je da se radi o godini koja nije prijestupna).

RJEŠENJE:

```
>>> mjeseci = {'siječanj' : 31, 'veljača' : 28, 'ožujak' : 31, 'travanj' : 30, 'svibanj' : 31, 'lipanj' : 30, 'srpanj' : 31, 'kolovoz' : 31, 'rujan' : 30, 'listopad' : 31, 'studeni' : 30, 'prosinac' : 31}
```

Vrijednost pojedinog elementa rječnika možemo dobiti na način da napišemo ime rječnika te unutar uglatih zagrada ključ. Primjerice broj dana u *ožujku* dobit ćemo naredbom:

```
>>> mjeseci['ožujak']
 31
```

Broj dana u veljači promijenit ćemo naredbom:

```
>>> mjeseci['veljača'] += 1
>>> mjeseci['veljača']
 29
```

Do problema će doći ukoliko pokušamo pristupiti ključu koji ne postoji:

```
>>> mjeseci['august']
Traceback (most recent call last):
      File "<pyshell#4>", line 1, in <module>
            mjeseci['august']
 KeyError: 'august'
```

Rješenju ovog problema možemo doskočiti na način da prvo provjerimo postoji li odgovarajući ključ u rječniku. To ćemo napraviti metodom keys(), koja vraća vrijednosti svih ključeva rječnika. Slično ćemo metodom values() dobiti listu svih vrijednosti rječnika. Ilustrirajmo to na primjeru kreiranog rječnika:

```
>>> mjeseci.keys()
dict_keys(['svibanj', 'studeni', 'srpanj', 'prosinac', 'lipanj', 'travanj', 'rujan', 'listopad', 'kolovoz', 'siječanj', 'veljača', 'ožujak'])
>>> mjeseci.values()
dict_values([31, 30, 31, 31, 30, 30, 30, 31, 31, 31, 29, 31])
 >>>
```

Drugi način na koji možemo riješiti problem pristupa nepostojećem elementu rječnika je korištenje metode get(). Opći oblik ove metode je: r.get(ključ, p). Ukoliko u rječniku (r) postoji element s pripadnim ključem ova metoda će vratiti pripadnu vrijednost, a ukoliko ne postoji element s ovim ključem metoda će vratiti vrijednost drugog navedenog parametra p. Ilustrirajmo to primjerom:

```
>>> mjeseci.get('august', -1)
 -1
```

Element ćemo u rječnik dodati metodom update(d), pri čemu je d rječnik odnosno element rječnika koji navodimo unutar vitičastih zagrada, a vrijednost je od ključa odvojena dvotočkom. Isto tako ukoliko pokušamo promijeniti vrijednost nekom elementu čiji navedeni ključ ne postoji kreirat će se novi element s pripadnom vrijednošću i ključem:

```
>>> mjeseci['august'] = 31
>>> mjeseci.get('august', -1)
31
>>> mjeseci.update({'december' : 31})
>>> mjeseci.get('december', -1)
 31
```

Element rječnika r s ključem k obrisat ćemo funkcijom `del(r[k])`:

```
>>> del(mjeseci['august'])
>>> mjeseci.get('august', -1)
 -1
```

IV. IMPLEMENTACIJA MATRICE RJEČNIKOM (KLASA MATRIX)

Svojstva rječnika koja smo upravo naveli iskoristit ćemo za implementaciju matrice te operacije s matricama. Osnovna ideja za implementaciju je da u rječniku čuvamo indekse i pripadne vrijednosti matrice. Indeks nekog elementa matrice bit će ključ dok će pripadna vrijednost elementa matrice biti vrijednost tog elementa rječnika. Dakle, ključevi rječnika bit će parovi brojeva (i, j).

B. *PRIMJER 2.*

Matricu:

$$a = \begin{bmatrix} 1 & 0 & 3 \\ 2 & 0 & 0 \\ 0 & 1 & 2 \end{bmatrix}$$

prikažimo rječnikom na način da u rječnik spremamo samo elemente koji su različiti od 0.

RJEŠENJE:

```
>>> r = {(0, 0) : 1, (0, 2) : 3, (1, 0) : 2, (2, 1) : 1, (2, 2) : 2}
```

Matricu ćemo implementirati kao klasu (Matrix). Konstruktorom ćemo definirati dimenzije matrice, pri čemu ćemo podrazumijevati da ukoliko nije drugačije rečeno da je matrica kvadratna, dimenzije 1, a ukoliko se konstruktor poziva sa samo jednim parametrom onda će se raditi o kvadratnoj matrici dimenzije n x n, pri čemu je n vrijednost proslijeđenog parametra. Elemente matrice, u obliku rječnika čuvat ćemo u rječniku _data. Osim toga definirat ćemo svojstva kojima ćemo moći mijenjati broj redaka i stupaca matrice i nakon kreiranja objekta tipa `Matrix`:

```
class Matrix:
    def __init__(self, n = 1, m = 0):
        self._data = {}
        self._n = n
        if m == 0:
            self._m = n
        else:
            self._m = m
        return

    @property
    def rows(self):
        return self._n

    @rows.setter
    def rows(self, value):
        self._n = value

    @property
    def columns(self):
        return self._m

    @columns.setter
    def columns(self, value):
        return self._m = value
```

Python raspolaže specijalnim metodama kojima možemo predefinirati neke postojeće operacije (`+, -, *, **, …`). Isto tako moguće je definirati operaciju pridruživanja ili vraćanja vrijednosti neke kolekcije elemenata odnosno prikaz objekta u obliku za ispis [1]. Popis specijalnih metoda dan je u tablici 1.

Za početak ćemo implicirati operacije za postavljanje (`__setitem__()`) i vraćanje (`__getitem__()`) vrijednosti matrice te vraćanje matrice u obliku za ispis (`__repr__()`).

Metoda __setitem__(self, k, v) ima dva parametra. Prvi parametar k predstavlja ključ na koji ćemo u rječniku spremiti vrijednost v. Primijetimo da je k u ovom slučaju par brojeva (i, j). Prilikom postavljanja vrijednosti nekog elementa još ćemo samo provjeriti jesu li indeksi unutar dimenzija matrice:

```
def __setitem__(self, i, t):
    if i[0] < 0 or i[0] >= self._n or i[1]
      < 0 or i[1] >= self._n:
        raise Exception('Element ne
           postoji')
    else:
       self._data[i]  = t
    return
```

Naziv metode	*Opis*
`__add__(self, b)`	zbrajanje (+)
`__sub__(self, b)`	oduzimanje (-)
`__mul__(self, b)`	množenje (*)
`__floordiv__(self, b)`	cjelobrojno dijeljenje (//)
`__truediv__(self, b)`	dijeljenje (/)
`__mod__(self, b)`	ostatak cjelobrojnog dijeljenja (%)
`__pow__(self, n)`	potenciranje (**)
`__lt__(self, b)`	manje od (<)
`__le__(self, b)`	manje ili jednako od (<=)
`__gt__(self, b)`	veće od (>)
`__ge__(self, b)`	veće ili jednako od (>=)
`__eq__(self, b)`	jednako (==)
`__ne__(self, b)`	različito (!=)
`__setitem__(self, k, v)`	vrijednost elementa s indeksom (ključem) *k* postavlja na *v*
`__getitem__(self, k)`	vraća vrijednost elementa s indeksom (ključem) *k*
`__repr__(self)`	vraća objekt u obliku za ispis

TABLICA 1. Najčešće korištena predefinirane metode

Na sličan način ćemo definirati i metodu koja vraća vrijednost koja se u matrici nalazi na mjestu (i, j), odnosno vrijednost rječnika _data čiji je pripadni ključ jednak (i, j), odnosno 0 ako pripadni ključ ne postoji u rječniku:

```
def __getitem__(self, i):
    if i[0] < 0 or i[0] >= self._n or i[1]
      < 0 or i[1] >= self._n:
        raise Exception('Element ne
           postoji')
    else:
       return self._data.get(i, 0)
```

Na kraju ćemo definirati metodu koja će vraćati matricu u obliku za ispis:

```
def __repr__(self):
   s = ''
   for i in range(self._n):
      for j in range(self._m):
         s += '{0:3d}'.format(self [i, j])
      s += '\n'
   return s[:-1]
```

Primijetimo da smo kod ispisa elemenata matrice pisali `self[i, j]`, čime smo zapravo pozivali metodu `__getitem__()`.

Ilustrirajmo upotrebu kreirane klase na nekoliko primjera:

```
>>> a = Matrix(3)
>>> a
  0  0  0
  0  0  0
  0  0  0
>>> a[0, 0] = 1
>>> a[1, 1] = 2
```

```
>>> a[2, 2] = 4
>>> a
  1  0  0
  0  2  0
  0  0  4
>>> a[3, 2] = 5
Traceback (most recent call last):
  File "<pyshell#27>", line 1, in <module>
    a[3, 2] = 5
  File
"C:\Users\predrag\Desktop\matrica.py",
line 45, in __setitem__
    raise Exception('Element ne postoji')
Exception: Element ne postoji
>>>
```

Važno je primijetiti da je čitava matrica od 9 elemenata pohranjena u rječniku od 3 člana, što se može i provjeriti:

```
>>> a._data
{(0, 0): 1, (1, 1): 2, (2, 2): 4}
>>>
```

Ovime smo dobili upravo ono što smo htjeli, implementaciju tzv. rijetkih matrica, a takve su upravo one koje koristimo za zapisivanje grafova.

Definiranje operacija nad matricama (+, -, *, **, …) je poprilično jednostavno i ne razlikuje se od definicije u slučaju kada matrice implementiramo standardnim poljima. Ilustrirat ćemo to na primjeru množenja dviju matrica, pri čemu ćemo ponovo paziti da ukoliko je vrijednost nekog elementa 0 da ga ne spremamo u rječnik:

```
def __mul__(self, a):
    if self.columns != a.rows:
        raise Exception('Matrice nisu
            ulančane')
    else:
        _r = Matrix(self.rows, a.columns)
            for i in range(self.rows):
                for j in range(a.columns):
                    t = 0
                    for k in range(self._m):
                        t += self[i,k] * a[k,j]
                    if t != 0:
                        _r[i, j] = t
        return _r
```

Dok implementacija potenciranja može imati sljedeći oblik:

```
def __pow__(self, n):
    if self.rows != self.columns:
        raise Exception('Potencirati se mogu
            samo kvadratne matrice')
    else:
        _tmp = Matrix(self.rows)
        for i in range(self.rows):
            _tmp[i, i] = 1
        for i in range(n):
            _tmp = _tmp * self
    return _tmp
```

V. PRIMJER UPOTREBE KLASE MATRIX

Za kraj ćemo još ilustrirati upotrebu kreirane klase na primjeru računanja broja putova između vrhova *x* i *y* u zadanom grafu prikazanom slikom 1. Neka je *a* matrica susjedstva zadanog grafa *G* te neka su *x* i *y* dva vrha. Broj putova duljine *n* od vrha *x* do vrha *y* jednak je *b[x, y]*, pri čemu je $b = a^n$.

C. PRIMJER 3.

Odredimo broj putova duljine 10 između čvorova A i C u grafu na slici 1.

Rješenje:

Matrica susjedstva danog grafa je:

$$a = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Napišimo niz naredbi kojima ćemo kreirati objekt t koji će predstavljati matricu *a*:

```
t = Matrix(5)
t[0, 1] = 1
t[0, 3] = 1
t[0, 4] = 1
t[1, 0] = 1
t[1, 2] = 1
t[1, 3] = 1
t[2, 1] = 1
t[3, 0] = 1
t[3, 1] = 1
t[4, 0] = 1
```

Sada ćemo broj putova duljine 10 između vrhova A i C dobiti tako da pročitamo vrijednost elementa $t^{10}[0, 2]$, a što ćemo dobiti naredbom:

```
print((a ** 10)[0, 2])
```

te ćemo dobiti rezultat: 607, što znači da postoji ukupno 607 putova duljine 10 između čvorova A i C.

VI. ZAKLJUČAK

Rječnik (*Dictionary*) u jeziku Python je vrlo upotrebljiva, ali nedovoljno korištena struktura jednostavno zbog njezina nedovoljnog poznavanja. Autori ovog članka su u svojoj knjizi [2] pokušali opisati osnovna svojstva te strukture. U sljedećoj knjizi, koja je u pripremi, bit će pokazane mogućnosti njezine svestrane uporabe.

LITERATURA

[1] D. M. Beazley, Python Essential Reference, Forth Edition, Addison-Wesley, Upper Saddle River, NJ, 2009

[2] L. Budin, P. Brođanac, Z. Markučič, S. Perić, Rješavanje problema programiranjem u Pythonu, Element, Zagreb, 2012.

[3] R. Sedgewich, Algorithms in C, Addison-Wesley Publishing Company Inc, 1996.

Programiranje u općoj, jezičnoj i klasičnog gimnaziji

mr.sc.Milomir Menciger
Gimnazija Vladimira Nazora Zadar
Perivoj Vladimira Nazora 3/2, Zadar, Hrvatska
Telefon: 023-778 777 Fax: 023-315 311 E-mail: milomir.menciger@skole.hr

Sažetak - Informacijsko komunikacijska tehnologija omogućila je da računala postanu pomagala pri rješavanju brojnih problema u različitim strukama. U školama se ostvaruju preduvjeti za uspješno učenje na svim razinama i u svim tipovima obrazovanja. Informatička pismenost je dio opće kulture i treba postati sastavni dio nastavnih sadržaja. Računala u nastavi informatike trebala bi omogućiti učenicima da savladaju osnovna znanja o programiranju za lakše usvajanje gradiva iz informatike i ostalih predmeta koji se izučavaju u školi. Suvremena tehnologija zahtijeva nova znanja i vještine. S gledišta obrazovanja to znači da je potrebno prilagoditi i sadržaje obrazovnih programa i obrazovnu tehnologiju učenicima novog doba.

U ovom radu razmatraju se mogućnosti postupnog uvođenja programiranja u općoj, jezičnoj i klasičnoj gimnaziji. U radu se obrađuje: opis nastavnog plana i programa, potreba za programiranjem, odabir informacijsko komunikacijske tehnologije i programskog jezika. Rad se temelji na osobnom iskustvu autora.

I. UVOD

Za razliku od drugih strojeva koji su namijenjeni točno određenoj svrsi, računalo je univerzalni stroj, čija uporaba je ovisna o programima koje korisnik koristi. Informacijska znanost danas se naglo razvija kao posljedica stalnog širenja i mogućnosti njezine primjene u svim granama ljudske djelatnosti. Korištenjem ICT u obrazovnom procesu stvara se novi posrednik između učenika, nastavnika i nastavnog gradiva. Učenici su aktivni u stvaranju novih informacija. U školama se stvaraju uvjeti za individualni i timski rad. Suvremena ICT zahtijeva nova znanja i vještine, s gledišta obrazovanja, to znači da je tome potrebno prilagoditi i sadržaj obrazovnih programa i obrazovnu tehnologiju.

Programiranje obuhvaća proces rješavanja zadataka uz pomoć računala. Kao što je računalo beskorisno bez programa koje je napisao čovjek, tako i čovjek, da bi postigao veći učinak, ne može bez računala. Proces izrade programa odnosi se na stvaralački čin za rješavanje složenih problema, koji zahtijeva nijansirano prosuđivanje, dobro poznavane struke i prihvaćanje timskog rada, a računalo velikom brzinom i preciznošću izvršava upravo one naloge koje je dobilo u obliku programa, a pritom koristi samo one kontrole koje su također prethodno isprogramirane.

Učenike opće, jezične i klasične gimnazije treba upoznati s načinom programiranja i usvajanja jednog od programskih jezika. Savladavanjem tehnike programiranja učenici će uočiti da ona doprinosi boljem razumijevanju problema, što će doprinijeti i boljem usvajanju ostalih sadržaja iz drugih predmeta, a posebno iz matematike. Predstavljanjem apstraktnih pojmova pomoću programskog jezika vrši se njihova materijalizacija, što povećava i bolje razumijevanje naučenog gradiva.

Uvođenje nove obrazovne tehnologije traži još veću osposobljenost profesora, individualni pristup učenicima, timski rad i permanentno usavršavanje. Upravo zbog toga profesori ne smiju sebe ograničiti na primjenu gotovih obrazovnih programa. Oni moraju kreirati svoje obrazovne programe i poticati učenike na rješavanje problema pomoću računala.

Kao osnova i temelj primjene računala svakako bi trebalo biti e-učenje (učenje na daljinu) i programiranje u jednom od programskih jezika. Kroz ovakav način rada učenike uvodimo u svijet računala koja će oni kasnije upotrebljavati za svoje osobne potrebe i permanentno obrazovanje.

II. OPIS NASTAVNOG PLANA I PROGRAMA

Informatika u osnovnoj školi nije obvezatan predmet. Učenici koji se upisuju u opće, jezične i klasične gimnazije u većini slučajeva nisu imali informatiku kao izborni predmet. Većina takvih učenika posjeduje računalo i Internet. Njihova znanja su veoma različita, ali uglavnom koriste gotove programe za različite namjene.

Nastavnim planom i programom za Informatiku u općoj, jezičnoj i klasičnoj gimnaziji definirani su ciljevi i sadržaj koji učenici trebaju usvojiti u jednoj školskoj godini sa dva sata tjedno [1].

Cilj informatičke izobrazbe je stjecanje osnovne informatičke pismenosti, do razine rješavanja jednostavnih problemskih situacija, uz korištenje informatičke tehnologije.

Program informatike za izobrazbu u jezičnima, općima i klasičnim gimnazijama izrađen je od devet cjelina, od kojih su prve 4 i deveta (obvezatan je prvi broj sati) obvezatne, a ostale se biraju o mogućnostima škole i propisanoj satnici.

Navedeni program i broj sati koji su definirani predstavljaju problem u izvođenju nastave, a isto tako i u odabiru udžbenika koji su namijenjeni za izvođenje nastave.

Informatika je na državnoj maturi izborni predmet. Ispitni katalog za državnu maturu iz Informatike u gimnazijama temeljni je dokument ispita kojim se jasno opisuje što će se i kako ispitivati na državnoj maturi iz ovoga predmeta. Ispitni katalog sadrži sve potrebne

informacije i detaljna pojašnjenja o obliku i sadržaju ispita [2]. Njime se jasno određuje što se od pristupnika očekuje na ispitu. Ispitni katalog usklađen je s odobrenim jednogodišnjim nastavnim planom i programom za Informatiku u općim, jezičnim i klasičnim gimnazijama.

Uspoređujući sadržaje nastavnog plana i programa i ispitnog katalog za državnu maturu vidljivo je da se njihovi sadržaji znatno razlikuju. Iz navedenog razloga potrebno je donijeti novi nastavni plan i program koji će definirati sadržaj i približan broj sati za realizaciju takvog sadržaja.

Ispitnim katalogom za državnu maturu u školskoj godini 2012./2013. definirane su područja ispitivanja i njihov udio koji je prikazan u tablici 1.

TABLICA 1.

PODRUĆJE ISPITIVANJA	UDIO BROJA BODOVA
Osnove uporabe računala i primjenskih programa	20%
Poznavanje građe i načela funkcioniranja računala	40%
Rješavanja problema programiranjem	40%

Prvo područje je zastupljeno sa 20% udjela na državnoj maturi. Znanja koja bi učenici trebali usvoji iz ovog područja odnose se ne samo na poznavanje rada računala već i na poznavanje MS Office-a. Broj sati koji je predviđen za ovo područje prema važećem nastavnom programu iznosi 14 sati godišnje. U općim, jezičnim i klasičnim gimnazijama taj sadržaj se obrađuje gotovo sa 70 sati. Novim nastavnim programom sadržaji iz ovog područja trebali bi se odnositi na operacijski sustav i pomoćne programe, te CARNetov Moodle sustav za online učenje. Ostali sadržaji trebali bi biti sastavni dio izborne nastave.

Drugo područje je zastupljeno sa 40% udjela na državnoj maturi. Broj sati koji se odnosi na ovo područje je 28 godišnje. Navedeni sadržaji su važni za razumijevanje rada računala. Jedan dio sadržaja moguće je obraditi i u Python-u. Posebnu pozornost treba posvetiti logičkoj algebri i načinu rada računala.

Treće područje je zastupljeno sa 40% udjela na državnoj maturi. Broj sati koji se odnosi na ovo područje je 28 godišnje. U općim, jezičnim i klasičnim gimnazijama rješavanje problema programiranjem je zapostavljeno. Učenici ne dobivaju dovoljno znanja iz ovog područja što predstavlja problem pri polaganju državne mature, a isto tako i za nastavak studija. Uvođenjem programskog jezika Python učenicima je omogućeno da svoju pažnju usmjere na rješavanje problema, jer je sintaksa programskog jezika veoma jednostavna. Sučelje koje se koristi za pisanje programa je slično ostalim programima pod operacijskim sustavom Windows.

Na osnovu ispitnog kataloga potrebno je napraviti novi nastavni plan i program koji će se moći realizirati sa propisanom satnicom. Važeći nastavni programi informatike su neusklađeni. U osnovnim školama informatika je izborni predmet. Učenici se upisuju u srednje škole sa različitim predznanjem. Nastava iz informatike organizirana je sa razrednim odjelima koji imaju i preko trideset učenika. Udžbenici koji su na raspolaganju učenicima i nastavnicima ne sadrže sva područja koja su definirana ispitnim katalogom za državnu maturu. Sadržaj udžbenika veoma brzo zastarijeva, jer se pojavljuju nove verzije programa.

Nastava informatike mora se temeljiti na procesu poučavanja, a ne predavanja. Nastavnici trebaju uvažavati učenikove sposobnosti i razvijati potrebu za znanjem, kao i aktivnim sudjelovanjem na satu.

Novi nastavni plan i program treba sadržavati e-učenje (učenje na daljinu) i programiranje u jednom od programskih jezika.

Učenje na daljinu nije pokušaj zamjene klasične nastave nekom drugom nastavom već je samo sredstvo koje doprinosi uklanjanju nedostataka klasične nastave i daje doprinos reformi obrazovanja. To je interaktivan ili dvosmjeran proces između nastavnika i učenika uz pomoć elektroničkih medija: nastava koja je usmjerena na učenika (a ne na nastavnika), koja će poticati aktivno usvajanje i primjenu novih znanja, te međusobnu suradnju i učenika i nastavnika.

Programiranje u općoj, jezičnoj i klasičnoj gimnaziji nije stvaranje profesionalnih programera kao što ni pisanje školske zadaće iz hrvatskog jezika nije stvaranje književnika. Savladavanjem tehnike programiranja i načinom pretvorbe u računalni program učenici stječu vještine, a njihova znanja postaju konkretna.

III.ODABIR INFORMACIJSKO KOMUNIKACIJSKE TEHNOLOGIJE I PROGRAMSKOG JEZIKA

Stari koncept izvođenja nastave temeljio se na izvođenju predavanja i vježbi sa dva školska sata bez odvojenih grupa sa tridesetak učenika Zbog velikog broja izostanka i nedovoljne aktivnosti učenika otežano je usvajanje gradiva, vježbi i kontinuirano praćenje učenika.

Obrazovanje treba imati naglasak na učeniku kao aktivnom sudioniku u nastavnom procesu. Iz navedenih razloga potrebno je unaprijediti nastavu uz primjenu ICT-a, jer većina učenika raspolaže sa internetom i računalom kod kuće. Osnovni cilj uvođenja e-učenja je omogućiti učenicima što lakše savladavanje znanja i vještina iz informatike. Ovakvim načinom rada očekuje se da će učenici:

- biti više zainteresirani za nastavne predmete
- da će se nastava biti više okrenuta prema učeniku (individualni pristup)
- izostajanje iz škole bit će znatno manje
- usvojenost nastavnih sadržaja znatno veća
- bolji odnos između profesora i učenika
- traženje pomoći izvan škole znatno manja (smanjene repeticije)
- polaganje državne mature bez manje stresa

- lakši upis na željeni fakultet

Za odabir ICT autor predlaže otvaranje vlastitog tečaja u CARNetovom Moodle sustavu za online učenje pod nazivom Loomen. Svi članovi akademske i školske zajednice imaju pravo na potpuno besplatno korištenje i svu pomoć oko korištenja ovoga sustava. Za otvaranje tečaja potrebno je ispuniti Zahtjev za otvaranje novog tečaja koji se nalazi na adresi https://loomen.carnet.hr. Online tečaj „Izrada online tečaja pomoću Moodle-a 2" korisnici mogu pohađati uz podršku online mentora ili kao online tečaj za samostalno učenje.

Programiranje zahtijeva logičko mišljenje i precizno definiranje pojmova koji se obrađuju. Putovi rješavanja istih problema su različiti. Programi obrađuju iste podatke, daju iste rezultate, ali različitim postupcima obrade, to jest različitom brzinom izvršavanja programa. U tradicionalnoj nastavi zadaci se često rješavaju istim postupcima. Koristeći neki drugi način, učenik nije u stanju doći do konačnog rješenja, iako je, možda imao dobru ideju, ali mu vrijeme nije dopuštalo da ostvari cilj. Pri radu sa računalom kreativnost je prepuštena učeniku, a računalo veoma brzo i precizno izvršava sve rutinske poslove. Ispitni katalog za državnu maturu iz Informatike opisuje što će se i kako ispitivati na državnoj maturi iz cjeline koja se odnosi na rješavanje problema programiranjem. Algoritmi su opisani pseudojezikom. Detaljan opis korištenja pseudojezika i njegova usporedba s programskim jezikom Pascal, C/C++ i Pythonom nalazi se u ispitnom katalog. Znanje na državnoj maturi provjerava se na razini pseudojezika. Takav način učenja bez kodiranja u nekom od programskih jezika predstavljao bi veliki problem u usvajanju nastavnih sadržaja. Cilj obrazovanja koje koristi tehnologiju je da učenici ne upoznaju samo programiranje kao apstraktnu znanost, već da uoče njenu praktičnu primjenu u životu i grade svoju intuiciju kreativnošcu i stvaranjem. Kod pisanja programa potrebno je pseudojezik prevesti na programski kod kojeg će računalo razumjeti, a nakon toga provjeriti rezultate obrade sa poznatim podacima.

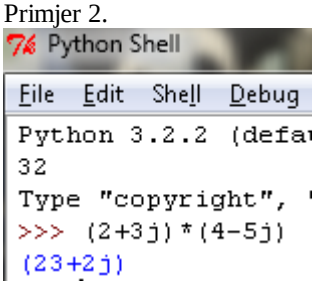
Za učenje programskog jezika autor predlaže programski jezik Python. Sintaksa pisanja koda je slična pseudojeziku, a isto tako je veoma razumljiva . Učenici gimnazije mogu početi učiti programiranje koristeći najjednostavniju sintaksu jezika da bi tijekom kasnijeg školovanja prešli na objektno orijentirano programiranje i paralelno programiranje, ali bez potrebe prelaska na druge programske jezike. Sintaksu i osnovne naredbe učenici gimnazije mogu savladati u jednom polugodištu. Pisanje programskog koda u programskom jeziku Pascal i C/C++ zahtijeva upravljanje memorijom što početnicima otežava prevođenje pseudojezika na programski kod. Python je više orijentiran na rješavanje konkretnog problema, pa se koristi za izradu prototipa, gdje je važnije dobiti funkcionalan algoritam nego optimizaciju koda, a jednom razvijen algoritma se može „preprogramirati" u nekom drugom jeziku koji daje brži kod [3].

Python je interaktivni programski jezik koji omogućuje zadavanje jedne po jedne naredbe koja se odmah izvršava. Takav način rada omogućuje učenicima savladavanje jednostavnih programskih koraka koje će koristit u pisanju programa. U interaktivnom sučelju moguće je izvoditi i aritmetičke operacije kao i na kalkulatoru bez poznavanja naredbi. Ilustracija je prikazana u sljedećim primjerima:

Primjer 1.

```
Python 3.2.2 (d
32
Type "copyright
>>> (3+4*2)*5
55
```

Primjer 2.



U primjeru 2. prikazano je množenje kompleksnih brojeva koji se izučavaju u drugom razredu jezične i klasične gimnazije, pa učenici mogu provjeriti svoje znanje iz matematike izvodeći one operacije koje su naučili rješavati na klasičan način.

Rješavanje složenijih problema izvodi se pomoću algoritama. Svaki dio programa naziva se modul. Modul na najvišoj razini naziva se glavni proces, a ostale module nazivamo podprocesima. Svaki sastavljeni proces može se opisati definiranjem sekvencije, selekcije i interacije. Proces može biti jednostavan ili složen. U jednostavne procese ubrzajmo one koji izvode operacije: unos, ispis, pridruživanje… i nizove takvih operacija koji tvore manje zaokružene cjeline. Sastavljeni procesi su oni koji se sastoje iz više jednostavnih procesa , odnosno bloka naredbi.

Sekvencija je niz podprocesa koji se odvijaju sekvencijalno, u nizu jedan za dugim, to jest od prvog do zadnjeg. Svaki podproces u sekvenciji može biti nova sekvencija, selekcija ili iteracija. Sekvencija kao budući modul u Pythonu predstavlja datoteku sa imenom i produžetkom (ime.py).

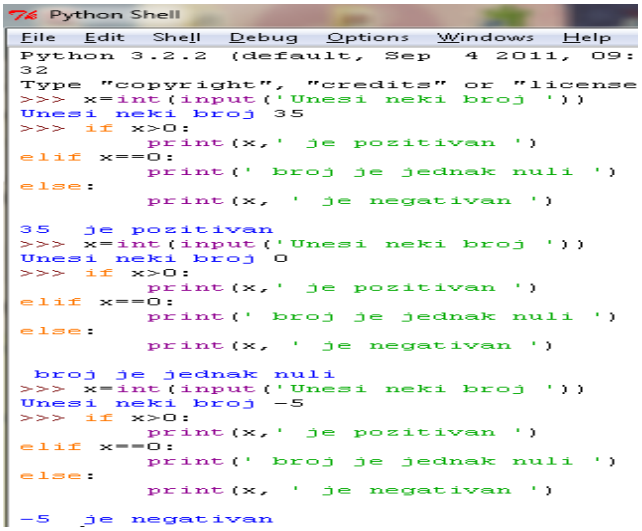
Selekcijom se odlučuje koji od dva procesa treba biti izvršen obzirom na postavljeni uvjet. Ako je uvjet ispunjen, onda se izvodi prvi proces, a ako uvjet nije ispunjen izvodi se drugi proces. Pored ove selekcije postoje i njegove modifikacije sa jednim ili više procesa. Iteracija je višestruko ponavljanje istog procesa, ali uvijek nad novim podacima. Iteracija se može izvoditi točno određen broj puta ili se može izvoditi sve dok je ispunjen postavljeni uvjet.

U Pythonu bi trebalo realizirati sadržaje koji su definirani ispitnim katalogom državne mature korištenjem Moodl sustava za online učenje pod nazivom Loomen. Takvim načinom rada tučenci će steći znanja koja su im potrebna za polaganje državne mature. Pomoću online učenja učenici će naučiti komunicirati na govornom jeziku. Koristeći Python moći će savladati matematiku. Za rješavanje zadataka znat će definirati problem i odrediti postupak (poštivajući pravila) kako bi ostvarili željeni cilj. U Pythonu dovoljno je pritisnuti tipku F1 da bi se dobila pomoć na engleskom jeziku koja opisuje način izvođenja

naredbi. Znanje engleskog jezika učenicima će olakšati rad sa Pythonom, a isto tako i sa drugim aplikacijama.

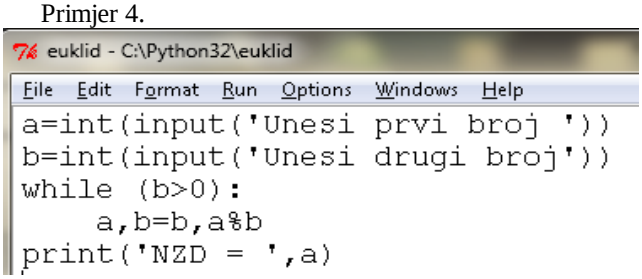
Za učenje je potrebno koristiti pseudojezik i sintaksu naredbi u Pytonu pomoću interaktivnog okruženja kako bi se bolje razumio način programiranja. Ilustracija je prikazana u sljedećom primjeru:

Primjer 3.



Pisanje programa izvodi se pomoću editora, a pokretanje programa se izvodi nakon snimanja pritiskom na tipku F5. Ilustracija je prikazana u sljedećem primjeru:

Primjer 4.



IV.ZAKLJUČAK

Osnovni cilj ovog rada bio je ukazati na potrebu uvođenja ICT tehnologije u opće, jezične i klasične gimnazije. Nastava informatike u navedenim gimnazijama izučava se jednu školsku godinu sa dva sata tjedno. Važeći nastavni plan i program je iz 1994. što za informatičku znanost predstavlja daleku prošlost. Iz navedenog razloga potrebno je donijeti novi nastavni plan i program. Za izvođenje nastave potrebno je koristi e-učenje i učenike osposobiti za izradu programa. Računala u nastavnom procesu zahtijevaju velika materijalna ulaganja, ali i određene pretpostavke kojima su ovladali i učenici i učitelji. Treba ići ukorak s novom tehnologijom, a kako je odgoj i obrazovanje djelatnost maksimalno okrenuta budućnosti nužno mu je posvetiti dužnu pozornost.

Programiranje predstavlja snažan alat kojim dobar učitelj može postići još bolje rezultate u odgojno obrazovnom procesu.

Škola treba postati tvornica znanja, a ne skladište izoliranih podataka koje učenik nije u stanju povezati u cjelinu prilikom rješavanja problema.

LITERATURA

[1] Glasnik Ministarstva, *„Nastavni programi za gimnazije“*, broj 1, Školske novine, Zagreb, 1944.

[2] Grupa autora, *„Ispitni katalog za državnu maturu u školskoj godini 2012./2013.“*

[3] http://www.monitor.hr/clanci/python-u-skole-1-5/152090

Poboljšanja u primjeni računalom vođenog demonstracijskog stola u nastavi

T.Horvat, Lj.Cvitaš
Tehničko veleučilište u Zagrebu, Zagreb, Hrvatska
thorvat3@tvz.hr; lcvitas@tvz.hr

Sažetak - U radu je opisano poboljšano rješenje za odvijanje nastave auditornih vježbi uz pomoć računalom vođenog demonstracijskog stola. Demonstracijski stol je primijenjen prvi puta u akademskoj godini 2011/12 na kolegiju „Materijali u elektrotehnici“, a u akademskoj godini 2012/13 provedena su poboljšanja opisana u ovom radu. Poboljšanja u opremi stola uključuju bolje mjerne instrumente i novu programsku podršku što omogućava kvalitetnije izvođenje pokusa i demonstracija uživo koje istovremeno prati velik broj studenata. Kako bi praćenje ovog oblika nastave bilo temeljitije, uvedene su i skripte u koje studenti zapisuju rezultate pokusa i svoje zaključke. U obje akademske godine provedene su ankete među studentima koje pokazuju opravdanost uložene investicije i uspješnost metode.

I. UVOD

U obrazovanju općenito, a posebno na stručnim studijima, bitno je teoretske postavke potkrijepiti praktičnim primjerima. U idealnim uvjetima praktični dio nastave izvodio bi se u laboratorijima sa specijaliziranom opremom i s malim brojem studenata.

Kolegij „Materijali u elektrotehnici“ izvodi se u prvom semestru stručnog studija elektrotehnike Tehničkog veleučilišta u Zagrebu te je zajednički za sve studente elektrotehničkog odjela [1]. Studenti dolaze iz različitih srednjih škola što donosi različita predznanja tj. opće poznavanje elektrotehničke struke. Također imaju i različite interese u pogledu daljnjeg obrazovanja i izbora između tri smjera (Energetska elektrotehnika, Automatizacija i procesno računarstvo ili Komunikacijska i računalna tehnika) na drugoj godini studija. Takav profil studenata zahtijeva prikaz širokog spektra različitih materijala, kako temeljnih materijala i njihovih općenitih primjena tako i specifičnih materijala koji se koriste u pojedinim segmentima elektrotehničke struke.

Satnica kolegija „Materijali u elektrotehnici“ ne predviđa laboratorijske, već samo auditorne vježbe uz predavanja. Veliki broj studenata koji pohađa kolegij ujedno znači i velike grupe (25 - 35 studenata) iz čega proizlaze razna ograničenja zbog kojih izvođenje nastave ovog kolegija u laboratoriju nije moguće. Najveće ograničenje je ono prostorno, u laboratorijima nije predviđeno izvođenje nastave za veliki broj studenata. Moguće rješenje bilo bi smanjiti grupe, no onda bi postojeći laboratoriji bili preopterećeni uzevši u obzir potrebe drugih kolegija. Dodatno ograničenje je količina opreme potrebne za pokuse. Bez obzira na ograničenja, cilj je studentima učiniti vježbe ovog kolegija zanimljivima, animirati ih, pobuditi im interes za materijale kao jednim od sastavnih dijelova elektrotehničke struke i pokazati im praktična svojstva materijala. Moguće rješenje bilo bi „donijeti studentima laboratorij u učionicu“, gdje pokuse vrše nastavnici ili asistenti uz studente kao publiku tj. promatrače.

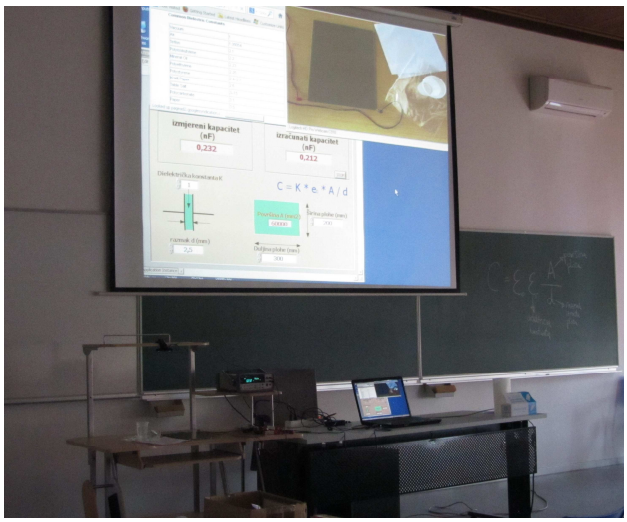
Uzimajući u obzir sve potrebe i želje te ograničenja kao rješenje se nameće računalom vođeni demonstracijski stol, koji preuzima na sebe funkciju koju inače ima laboratorijski prostor. Demonstracijski stol je primijenjen prvi puta u akademskoj godini 2011/12 na kolegiju „Materijali u elektrotehnici“ te je opisan u radu „Primjena računalom vođenog demonstracijskog stola u nastavi“ (MIPRO, 2012, CE) [2].

II. IZVEDBA

A. Informacijski okviri

Računalom vođeni demonstracijski stol treba većem broju studenata omogućiti praćenje auditornih vježbi sa praktičnim pokusima. Pod time se podrazumijeva kako svi studenti moraju moći dobro vidjeti postav i izvođenje pokusa, kao i rezultate mjerenja s instrumenata te ovisno o pokusu i čuti zvučne pojave. Kako bi to bilo moguće predviđeni su sljedeći informacijski okviri:

- Video slika površine stola sa ispitnim modelom i instrumentima te pomoćnim sredstvima uz model
- Video slika detalja na modelu (makro slika)
- Podaci sa dva digitalna multimetra, po izboru napon, struja, otpor, kapacitet ili temperatura
- Podaci sa dvokalnalnog digitalnog osciloskopa u obliku snimljene pojave koja se osvježava u sekundnom ritmu
- Podaci sa upravljivog izvora napajanja 0-30V, 3A ili video slika upravljačke ploče
- Prezentacijski podaci, tehnički podaci u pdf, jpeg ili drugom slikovnom formatu
- Rezultati mjerenja te grafovi prikazani u Labviewu



Slika 1 Prikaz informacijskih okvira

Dobra vidljivost osigurana je kamerom i projektorom priključenim na prijenosno računalo.

B. Struktura opreme i video podsustav

Struktura opreme korištene za ostvarivanje ranije navedenih informacijskih okvira u ak. godini 2011/12 temeljila se na istraživanju najpovoljnijeg izbora nužne opreme koje je izvršeno krajem 2011. godine [3]. Pri izboru prednost su imali postojeća instrumentacija u laboratorijima Zavoda za automatizaciju i procesna računala, neka od opreme koja je u tim laboratorijima stajala na raspolaganju te besplatna programska podrška.

Instrumentacija je morala zadovoljiti barem dva uvjeta:

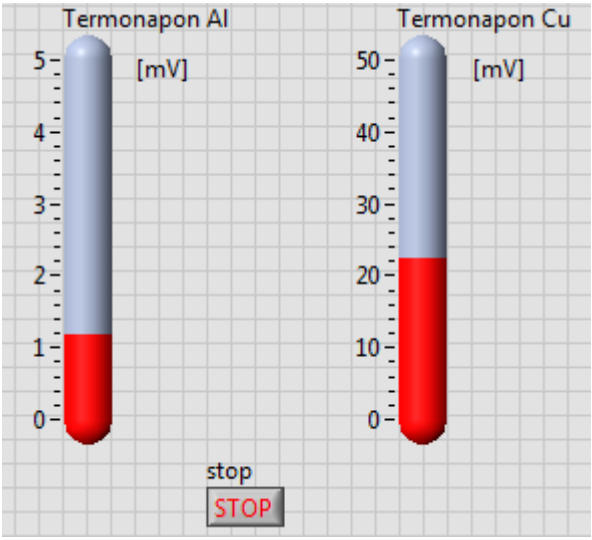
- Serijska komunikacija preko USB priključka
- Jednostavna progamska podrška te mogućnost otvaranja više instanci istog pogonskog programa u slučaju potrebe korištenja više istih instrumenata

U ak. godini 2011/12 korištena je sljedeća oprema:

- Prijenosno računalo manjih dimenzija kao vodeće računalo
- 7-kanalni USB razdjelnik tipa D-LINK DUB-H7 na koji se zvjezdasto priključuju ostali uređaji [4]
- Multimetri tipa PeakTech 3388 opremljeni originalno serijskom vezom RS232, ali se uz njih isporučuje i konvertor na USB serijski standard [5, str. 42-43]. Glavni nedostatak tih multimetara je nemogućnost daljinskog programiranja izbora načina mjerenja te izbora mjernog područja.
- Ugrađena programska podrška podržava jednosmjernu komunikaciju prema računalu, tj. multimetar PeakTech emitira blokove izmjerenih podataka u sekundnom ritmu.
- Dvokanalni digitalni osciloskop tipa Rigol DS1052E koji ima propusnost od 50MHz te frekvenciju uzorkovanja od 1000GS/s [6]. Opremljen je USB priključkom preko kojeg je moguće programiranje parametara osciloskopa kao i prijenos snimljene pojave na računalo.

Krajem akademske godine 2011/12 pribavljena su dva digitalna multimetra tipa Agilent 34405A [7], koja su već u prošlom radu spomenuta kao bolje tehničko rješenje od multimetara tipa PeakTech 3388. Pribavljen je i programski paket Labview (prvenstveno za potrebe Zavoda za komunikacijsku i računalnu tehniku, a dostupan svim odjelima Tehničkog veleučilišta u Zagrebu), koji je puno prikladnija programska podrška za ovaj tip vježbi [8]. Navedena oprema i programska podrška korištene su u auditornim vježbama „Materijala u elektrotehnici“ u akademskoj godini 2012/13.

Digitalni multimetri tipa Agilent 34405A omogućuju spajanje na računalo putem USB priključka, imaju mogućnost daljinskog programiranja izbora načina mjerenja i izbora mjernog područja. Navedeni multimetar ima i mogućnost povezivanja s programskim paketom Labview te je tako omogućeno prikazivanje podataka u živo s instrumenata. Moguće je istovremeno povezivanje više instrumenata s jednom aplikacijom u programskom paketu Labview, što eliminira potrebu za otvaranjem više instanci istog pogonskog programa u slučaju korištenja više istih instrumenata. Ta potreba je bila izražena kod korištenja multimetara tipa PeakTech i njihove ugrađene programske podrške, tj. za svaki priključeni instrument bilo je potrebno pokretati novu instancu programa kako bi se prikazivali podaci sa svih instrumenata. Grafički prikaz podataka je, kod korištenja dva PeakTech instrumenta, bio dugotrajan i nespretan za izvedbu na vježbama jer je za usporedbu prikupljenih podataka bilo potrebno podatke iz programske podrške spremiti u dvije Excel tablice, potom pravilno kopirati oba seta podatka u jednu Excel tablicu, ukloniti nepotrebne spremljene podatke i tek tada izraditi graf. Programski paket Labview podržava direktno iscrtavanje grafa, kao i druge načine grafičkog prikaza podataka (slika 2), bez korištenja dodatnih alata.



Slika 2 Grafički prikaz podataka prikupljenih sa dva instrumenta u Labviewu

Prijenos video informacija o pokusu koji se izvodi pred studentima realiziran je putem video podsustava. Živa slika se snima i prenosi koristeći web kameru male mase i prikladnih dimenzija, čime je omogućeno dinamičko prenošenje kamere na različita mjesta pokusnog modela te snimanje detalja zbivanja tijela ispitnog procesa. Kamera je spojena preko USB priključka na razdjelnik odnosno na vodeće računalo. Pomoćni pogonski program instaliran na računalu daje mogućnost pokretanja nekoliko instanci video okvira na monitoru računala u slučaju snimanja s više kamera [9].

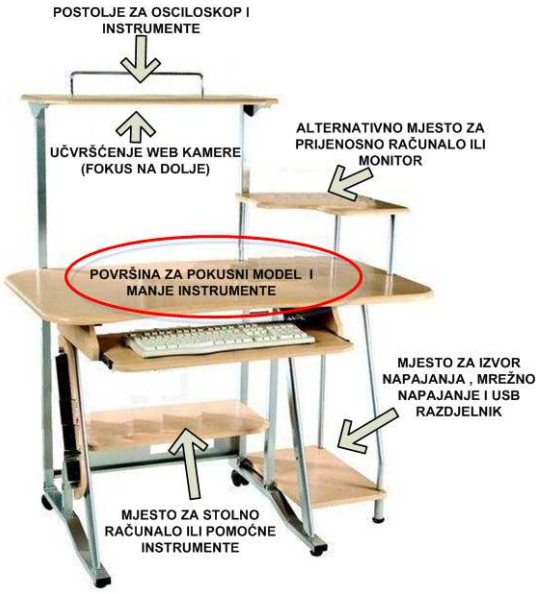
Izabran je tip web kamere Logitech HD Pro C910 zbog izvanrednih tehničkih performansi, a umjerene cijene [10]. Ta je kamera opremljena Carl Zeiss® optikom, snima u rezoluciji do 1920 x 1080 piksela te ima pogonski upravljački program kojim se može fokusirati određeni detalj, ručno ili automatski izoštriti slika. Ima ugrađena dva mikrofona za stereo snimanje tako da se može prenositi i zvučni kanal koji je od interesa u nekim pokusima. Zvuk se emitira ili preko ugrađenih zvučnika računala, preko vanjskih zvučnika ili kod bolje opremljenih predavaonica, preko zvučnog sustava koji postoji uz video projektor.

C. Konstrukcija

Za mehaničku izvedbu demonstracijskog stola pronađeno je najpovoljnije moguće rješenje; pretraživanjem ponude komercijalnih pisaćih odnosno stolova za računalo po hrvatskim trgovinama namještaja uočen je model stola koji originalnom konstrukcijom prednjači kao kandidat za alternativnu namjenu [11].

Konstrukcija stola ima sljedeće performanse:

- čelični noseći dijelovi kostura stola osiguravaju čvrstoću i nosivost drvenih horizontalnih ploha
- postolje stola opremljeno je kotačima na stražnjem dijelu što omogućava mobilnost, a u stacionarnom položaju stol ne kliže
- najveća površina (radna ploha) za pokusni model i manje instrumente je dimenzija 60x115 cm što je dovoljno veliko za organizaciju jednog pokusa, a opet dovoljno malo da stol bude mobilan
- gornje postolje nalazi se iznad modela pa ima dvostruku funkciju, na gornjoj plohi (30x60) ima dovoljno mjesta za npr. digitalni osciloskop i multimetar, a ispod plohe je pravi izbor za lokaciju i učvršćenje web kamere koja snima zbivanja na pokusnom modelu
- bočna polica može se iskoristiti za smještaj prijenosnog računala ili alternativno za manji monitor (do 14“) ako se umjesto prijenosnika koristi standardno stolno računalo. Umjesto za smještaj računala ova polica može se iskoristiti i za smještaj većeg multimetra ili generatora funkcija
- donje police (ispod radne površine i dolje bočno) predviđene su za smještaj teže opreme; programibilni izvor napajanja, stolno računalo i slično



Slika 3. Konstrukcija stola i predviđene pozicije za opremu

- lokalno osvjetljenje može se izvesti ugradnjom LED rasvjete na donju površinu gornjeg postolja

D. Skripta za vježbe

Kako bi praćenje ovog oblika nastave bilo temeljitije, uvedena je i skripta za auditorne vježbe [12].

U okviru auditornih vježbi na kolegiju Materijali u elektrotehnici predviđeno je 8 demonstracijskih auditornih vježbi. Skripta obuhvaća sve vježbe te je za svaku od vježbi podijeljena na dva dijela: pripreme za vježbu i rad na vježbi. Po završetku semestra studenti moraju predati barem 5 potpuno ispunjenih vježbi kao uvjet za potpis tj. kao uvjet za izlazak na ispit. Predaja 6 ili više vježbi nagrađuje se dodatnim bodovima.

U dijelu skripte koji se odnosi na pripreme za vježbu studenti odgovaraju na postavljena teoretska pitanja iz područja koje je ranije obrađeno na predavanjima. Izabrana su pitanja koja su relevantna za razumijevanje pokusa koji se izvode na vježbama.

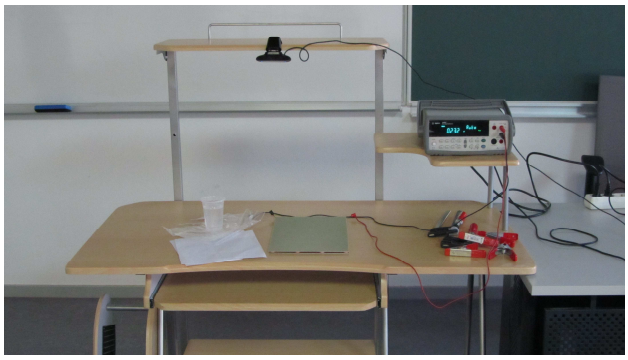
U dijelu skripte koji se odnosi na rad na vježbi, studenti, ovisno o vježbi, skiciraju postave mjerenja i sheme spoja kao i pojedine elemente, tablično prikazuju parametre kao i rezultate mjerenja i izračuna te upisuju svoja opažanja i zaključke.

Pripremu za vježbu studenti su dužni napraviti prije samih auditornih vježbi, a rad na vježbi ispunjavaju za vrijeme vježbi. Skripta za auditorne vježbe dostupna je studentima za preuzimanje na mrežnim stranicama kolegija Materijali u elektrotehnici.

III. IZVOĐENJE POKUSA

A. Priprema demonstracijskog stola

Prije početka auditornih vježbi potrebno je pripremiti demonstracijski stol za provedbu vježbi, kao što je prikazano na slici 4.



Slika 4. Model i postav mjerenja

Demonstracijski stol se za vrijeme pripreme auditornih vježbi oprema jednim ili više modela za izvođenje pokusa, odgovarajućim instrumentarijem za provedbu mjerenja te napajanjem. Model i manji instrumenti se smještaju na radnu plohu, a ako postoji potreba za vanjskim napajanjem ono se zbog težine smješta na neku od donjih ploha. Digitalni osciloskop se smješta na gornju plohu, a veći instrumenti uobičajeno se postavljaju na bočnu plohu. Prijenosno računalo se obično smješta na nastavnički stol u učionici, ali može ga se postaviti i na demonstracijski stol ako je potrebno.

Prije svake vježbe bilo je potrebno izraditi aplikaciju u programskom paketu Labview koja služi za brojevni kao i grafički prikaz mjerenih veličina.

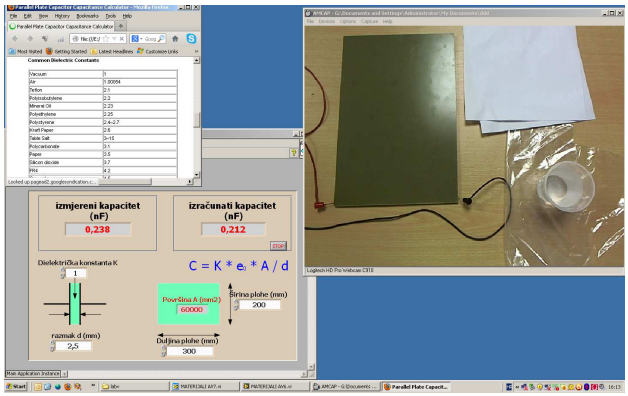
Na zaslonu računala trebaju biti prikazani informacijski okviri tako da su svi jasno vidljivi studentima za vrijeme izvođenja pokusa, kao što je prikazano na slici 5. Pojedinačni informacijski okvir može se za vrijeme vježbi privremeno povećati kako bi mu se posvetila veća pažnja.

B. Provedba pokusa: Izrada i ispitivanja na pločastim kondenzatorima

Na predavanjima su dane teoretske postavke o dielektričnim materijalima i pločastim kondenzatorima. Predstavljena je formula za kapacitet pločastog kondenzatora koja glasi

$$C = \varepsilon_0 * \varepsilon_r * A / d, \qquad (1)$$

gdje je ε_0 dielektrična permitivnost vakuuma, ε_r dielektrična konstanta materijala, A površina ploča i d razmak između ploča.



Slika 5 Informacijski okviri na zaslonu računala

Teorijske postavke biti će provjerene na vježbi gdje će se izraditi model pločastog kondenzatora koristeći jednostrane FR4 pločice i nekoliko različitih dielektričnih materijala.

Osnovni model se sastoji od dvije jednostrane FR4 pločice koje se inače koriste za izradu vodova. S jedne strane pločice je dielektrik FR4 (vitroplast), a s druge strane pločice tanki sloj bakra. Jedna od pločica, na bakrenoj strani, ima na rubovima zalijepljene odstojnike visine 2,5mm koji se koriste kako ne bi došlo do dodira dviju bakrenih pločica i drže udaljenost ploča fiksnom.

Kao dielektrici, tj. materijali koji će se umetnuti između ploča, koriste se zrak (uz pomoć odstojnika, ε_r=1), papir (ε_r=3,5), FR4 (ε_r=4,2) te voda (ε_r=80, pri 20°C). Korištenjem različitih dielektrika pokazuje se promjena kapaciteta ovisno o dielektričnoj konstanti. Pločice se mogu pomicati kako bi se pokazala promjena kapaciteta ovisno o površini ploča. Promjena kapaciteta ovisno o razmaku između ploča pokazuje se umetanjem prvo jednog pa više papira između ploča ili okretanjem pločica tako da su između bakrenih djelova prvo jedan sloj, a potom dva sloja FR4.

Tablica 1 preuzeta je iz skripte za vježbe te u istu studenti za vrijeme vježbi upisuju dielektrične konstante različitih dielektričnih materijala, površine ploča, razmake između ploča, rezultate izračuna te rezultate mjerenja kapaciteta. Tablica se po potrebi može proširiti dodatnim dielektricima.

TABLICA I. PARAMETARI I REZULTATI MJERENJA NA PLOČASTIM KONDENZATORIMA

Dielektrik	Dielektrična konstanta	Površina ploča [mm^2]	Razmak između ploča [mm]	Izmjereni kapacitet [nF]	Izračunati kapacitet [nF]
Zrak	1	60000	2,5	0,232	0,212
Papir	3,5	40000	0,15	6,98	8,26
Vitroplast	4,2	60000	1,5	1,418	1,487
Vitroplast	4,2	60000	3	0,709	0,743

IV. ANKETA

Pri kraju odvijanja nastave kolegija „Materijali u elektrotehnici“ provedena je dobrovoljna i anonimna anketa među studentima. Od 236 studenata koji su upisani na kolegij, njih 139 je ispunilo listić sa anketnim pitanjima. Anketa je sadržavala 9 pitanja vezanih uz tematiku provođenja auditornih vježbi po ovom modelu, a uz svako pitanje bilo je ponuđeno 5 mogućih odgovora. Deseto pitanje bilo je traženje prijedloga za poboljšanje auditornih vježbi.

Detaljni rezultati ankete prikazani su u tablici 2. Prema analizi odgovora većina studenata odabrala je odgovore:

- Smatraju auditorne vježbe pretežno do vrlo korisnima
- Kako treba i više praktičnih pokusa na vježbama
- Sviđa im se kada demonstrator na vježbama više izvodi pokus u živo
- Auditorne vježbe bez pokusa bile bi dosadne
- Prikazivanje pokusa u živo kamerama i instrumentima prilika je da svi dobro vide detalje, upoznaju praktična svojsta te potvrde teoretska znanja
- Vježbe su im zanimljive u grupama do 30 studenata
- Pokusi u živo zanimljivi su zbog zanimljivih ishoda, tehničkog sadržaja pokusa kao i podataka u živo s instrumenata
- Zadovoljni su razinom komunikacije između demonstratora i studenata na vježbama
- Smatraju pitanja za pripremu vježbe u skripti korisnima te putem njih uče i za kolokvije i ispite

Najistaknutiji i najčešći prijedlozi za poboljšanje auditornih vježbi su:

- Više pokusa
- Mogućnost sudjelovanja studenata u pokusima
- Samostalno izvođenje pokusa

TABLICA II. REZULTATI ANKETE PROVEDENE MEĐU STUDENTIMA

Pitanje	1	2	3	4	5
Auditorne vježbe proširuju opseg znanja dobivenih na predavanjima.	ne 1,44%	djelomično 15,83%	nisam siguran/sigurna 9,35%	mislim da su pretežno korisne 42,45%	vrlo su korisne 30,94%
Bilo bi dobro da se na vježbama ima i više praktičnih pokusa	sada ih ima previše 2,86%	treba ih malo smanjiti 0,00%	sada ih ima dovoljno 40,71%	slažem se 48,57%	treba ih puno više 7,86%
Sviđa mi se kada nastavnik na vježbama	više priča 10,16%	više crta po ploči 4,28%	više izvodi pokus u živo 49,20%	više razgovara s studentima 18,18%	više pokazuje detalje kamerom 18,18%
Auditorne vježbe bez pokusa u živo bile bi	dosadne 41,32%	ne bi ništa naučili 20,36%	iste kao predavanja 35,33%	bolje 1,80%	sadržajnije znanjem 1,20%
Prikazivanje pokusa uživo kamerama i instrumentima	zbunjuje me 0,00%	ne pruža ništa što već ne znam 1,14%	prilika je dobiti potvrdu teoretskih znanja 25,14%	prilika je upoznati praktična svojstva 30,86%	prilika je da svi dobro vide detalje 42,86%
Auditorne vježbe s pokusima zanimljive su u	velikim grupama studenata (>50) 4,90%	grupama do 50 osoba 9,79%	srednjim grupama (do 30 osoba) 63,64%	grupama do 15 osoba 16,78%	malim grupama (do 10 osoba) 4,90%
Pokusi u živo interesantni su radi	video prikaza detalja 18,22%	podataka u živo sa instrumenata 20,34%	tehničkog sadržaja pokusa 22,46%	zanimljivih ishoda pokusa 25,85%	primjenjene tehnike prikaza 13,14%
Treba više komunikacije između demonstratora i studenata	komunikacija nije potrebna 8,45%	i sada je previše 1,41%	sasvim je u redu 71,83%	da, treba malo više 14,08%	da, treba više 4,23%
Smatram kako pitanja za pripremu vježbe u skripti	ima previše 13,29%	ima premalo 0,58%	su beskorisna 4,62%	su korisna 43,35%	putem njih učim i za kolokvij / ispit 38,15%

V. ZAKLJUČAK

U ovom članku prikazan je alternativni način izvođenja auditornih vježbi kolegija „Materijali u elektrotehnici“ na Tehničkom veleučilištu u Zagrebu. Klasične teorijske auditorne vježbe zamijenjene su demonstracijskim vježbama u kojima se studentima „donosi laboratorij u učionicu“. Koristeći računalom vođeni demonstracijski stol studentima se na zanimljivi način prikazuju praktična svojstva materijala.

Računalom vođeni demonstracijski stol pokazao se kao jednostavno, ali efektno rješenje za odvijanje auditornih vježbi. Demonstracijski stol omogućava izvođenje praktičnih pokusa, uz dobru vidljivost detalja, pred većim brojem studenata. Na taj način radi se prezentacija mogućih primjena materijala, proširenje opsega te potvrda teoretskih znanja stečenih na predavanjima. Rezultati mjerenja prikazuju se u živo sa instrumenata što doprinosi laboratorijskom dojmu.

Ostvarena su neka od poboljšanja predloženih u prijašnjem radu, pribavljeni su novi instrumenti i prikladnija programska podrška čime se povećava kvaliteta vježbi. Programski paket Labview znatno olakšava prikaz podataka s instrumenata i čini ga jasnijim koristeći grafičko sučelje. Dodatno, moguće su brze i jednostavne izmjene u funkcionalnosti Labview aplikacije u slučaju proširenja opsega pojedine vježbe. Uvedena je i skripta za auditorne vježbe kako bi studenti kvalitetnije i temeljitije pratili nastavu. Naravno, postoji prostor i za daljnja poboljšanja. Primjerice postavljanje dodatnih kamera zbog još boljeg prikaza detalja te aktivnije uključivanje studenata u obavljanje pokusa.

Važan pokazatelj kvalitete ovakvog izvođenja nastave su pozitivne povratne informacije od studenata dobivene kroz izravnu komunikaciju, ali i kroz anketu.

Ovakav način izvođenja auditornih vježbi mogao bi se primjenjivati i na drugim kolegijima stručnog studija elektrotehnike na Tehničkom veleučilištu u Zagrebu, uz odgovarajuće izmjene ovisno o potrebama pojedinog kolegija. Naravno, ovaj model primjenjiv je i za izvođenje nastave u srednjim školama.

LITERATURA

[1] Lj. Cvitaš, “Zbirka predavanja iz kolegija Materijali u elektrotehnici“, Tehničko veleučilište u Zagrebu, repozitorij, 2012

[2] Lj. Cvitaš, T.Horvat, „Primjena računalom vođenog demonstracijskog stola u nastavi“, MIPRO 2012, prihvaćen za tisak

[3] H. Peharec,“Demonstracijski stol“, Završni rad br. 1104, Tehničko veleučilište u Zagrebu, 2011.

[4] D-LINK, DUB-H7, *www.**dlink**.com/products/?pid=149*

[5] PeakTech 3380 Digital Multimeter, Operation manual, CD priložen uz uređaj

[6] RIGOL, www.rigolna.com/products/digital-scilloscopes/ds1000e/

[7] Agilent, http://www.home.agilent.com/agilent

[8] Labview http://sine.ni.com/np/app/main/p/docid/nav-104/lang/hr/

[9] N. Danjou, “AMCap, video capture and priview software”, www.noeld.com

[10] Logitech, “HD Pro C910”, www.logitech.com/en-us/webcam

[11] KiKa Zagreb,” Stol za računalo Pamplona JN-9080", www.kika.com/hr

[12] Lj.Cvitaš, T.Horvat, “Materijali u elektrotehnici, Skripta za auditorne vježbe”, Tehničko veleučilište u Zagrebu, repozitorij, 2013

Metodologija izrade *E – learning* sadržaja u svrhu osposobljavanja za polaganje ECDL certifikata

S. Grgić* i A. Lacković *
* Visoko učilište Algebra – visoka škola za primijenjeno računarstvo, Zagreb, Hrvatska
silvija.grgic@racunarstvo.hr
andrej.lackovic@racunarstvo.hr

Sažetak - ECDL (engl. *European Computer Driving Licence*) predstavlja međunarodno priznat certifikat računalne pismenosti koji jamči vlasniku osnovna informatička znanja i vještine. Za polaganje certifikata osoba se može pripremati pohađanjem prilagođenih seminara ili samostalnim učenjem uz pomoć *E-learning* sadržaja. Rad opisuje metodologiju izrade *E–learning* sadržaja koji će na kvalitetan način pripremiti osobu za polaganje ECDL certifikata. Cilj rada je pokazati posebnosti različitih metodologija izrade *E–learning* sadržaja te razloge zašto je određena metodologija prikladna za izradu ove vrste sadržaja. Ujedno, ovaj rad daje pregled i organizaciju stvarnog *E–learning* sadržaja koji se koristi za osposobljavanje za polaganje ECDL certifikata. U završnom dijelu rada daje se osvrt na alate koji su korišteni za izradu ovog tipa *E–learning* sadržaja s naglaskom na njihovu kompatibilnost sa odabranom metodologijom izrade sadržaja.

I. UVOD

S obzirom da se nalazimo u vremenu velikih promjena, vremenu u kojem tehnologija u svim oblastima života svakodnevno napreduje, tako napreduje i područje obrazovnih tehnologija. Osim klasičnog obrazovanja, pojavljuje se potreba kvalitetnog, brzog i pravovremenog obrazovanja kako u školi, fakultetu, tako kasnije i na radnom mjestu. *E–learning* se nameće kao moćno rješenje za osiguranje brzog i pravovremenog obrazovanja. Da bi *E–learning* mogao konkurirati klasičnom obrazovanju u smislu kvalitete nastavnog sadržaja i kako bi mogao funkcionirati kao samostalni oblik edukacije potrebno je odabrati pravilnu metodologiju za izradu sadržaja.

Cilj ovog rada je opisati metodologiju koja stoji iz izrade *E-learning* sadržaja koji osposobljava polaznika za polaganje ECDL certifikata. ECDL certifikacijski program osnovan je 1996. godine na prijedlog Radne skupine Skupštine Europskog udruženja profesionalnih informatičara (CEPIS) s ciljem određivanja razine digitalne pismenosti. Osnovni ECDL predstavlja međunarodno priznat standard koji certificira osnovna informatička znanja. Sadržaj Osnovnog ECDL-a složen je tako da pokriva ključne koncepte računala i sastoji se od 7 modula:

1. modul 1: Osnovni koncepti informacijske tehnologije (engl. *Concepts of Information and Communication Tehnology* (ICT),
2. modul 2: Korištenje računala i upravljanje datotekama (engl. *Using the Computer and Managing Files*),
3. modul 3: Obrada teksta (engl. *Word Processing*),
4. modul 4: Tablične kalkulacije (engl. *Spreadsheets*),
5. modul 5: Baze podataka (engl. *Using Databases*),
6. modul 6: Prezentacije (engl. *Presentation*) i
7. modul 7: Rad u mreži i internet (engl. *Web Browsing and Communication*).

Za certifikaciju se osoba može pripremati na razne načine: samostalnim učenjem, korištenjem *E-learning* seminara ili pohađanjem prilagođenih seminara. Ovim radom daje se pregled i organizacija ECDL *E–learning* sadržaja i alata korištenih za izradu istog.

II. METODOLOGIJA IZRADE NASTAVNOG SADRŽAJA

E - learning seminar se sastoji od nastavnih jedinica, unutar kojih su poglavlja, a unutar njih teme kao najmanje jedinice odnosno nastavni objekti. Tema kao najmanji nezavisni edukativni element sadrži cilj nastavnog objekta u obliku izjave koja opisuje nova znanja i sposobnosti koje će kandidat posjedovati, nastavni objekt kojim se kandidata poučava i provjeru naučenih znanja i vještina koja kontrolira da li je kandidat postigao zadani cilj. Teme kao i poglavlja trebaju biti u skladu sa ciljevima nastavnih jedinica. Da bi se to osiguralo potrebno je odabrati pravilnu metodologiju za izradu određenog poglavlja odnosno teme.

Za izradu ECDL *E–learning* sadržaja kategorizacija očekivanih znanja i vještina radi se uz pomoć Bloomove taksonomije. Benjamin Bloom (1913. – 1999.) dijeli kognitivnu domenu učenja na 6 odvojenih razina usvojenosti znanja:

1. Znanje – prepoznavanje informacija
2. Razumijevanje – shvaćanje informacija
3. Primjena – primjena znanja u rješavanju problema
4. Analiza – razdvajanje informacija na sastavne dijelove
5. Sinteza – primjena informacija
6. Vrednovanje – prosuđivanje korisnosti

Prelazak u svaku višu razinu omogućen je isključivo ukoliko je zadovoljena razina znanja propisana svim prethodnim razinama.

Ova taksonomija prikladna je za ECDL *E–learning* sadržaj iz razloga što potiče izgrađivanje znanja više razine na temeljima postojećih znanja nižih razina. Svaka razina taksonomije povezana je sa ključnim glagolima koji omogućuju definiranje kvalitativnih i kvantitativnih ishoda učenja na osnovu kojih polaznici mogu pokazati usvojena znanja i vještine. Koristeći pravilan glagol povezan sa određenom razinom Bloomove taksonomije kreatori nastavnih sadržaja ostvaruju cilj svakog modula. Ako uzmemo za primjer modul 3 koji za cilj postavlja dokazivanje sposobnost za korištenje aplikacija za obradu riječi tada tema u kojoj se polaznika traži da imenuje ili nabroji sve alate za oblikovanje teksta što je razina 2 (razumijevanje) prema Bloomovoj taksonomiji ne bi u potpunosti dovela do ispunjenja zadanog cilja. Ako bi se polazniku predočio komad neformatiranog teksta i tražilo da primjeni alat za oblikovanje teksta što je razina 3 (znanje) prema Bloomovoj taksonomiji polaznik to ne bi znao napraviti bez savladane razine 2 (razumijevanje) i znanja koji su alati za oblikovanje teksta. Ovaj primjer dobro ocrtava zašto je pravilna upotreba Bloomove taksonomije važna kod izrade nastavnog sadržaja.

Osim pravilne kategorizacije znanja i vještina unutar svakog modula potrebno je i pravilno definirati metode koje će se koristiti za svaki modul. Robert Gange (1916.-2002.) definira instrukcijski dizajn koji postavlja teoriju da postoji nekoliko različitih vrsta ili razina učenja. Značaj njegove klasifikacije je da svaki drugačiji cilj zahtjeva drugačiju vrstu nastave. Osim toga teorija ocrtava devet nastavnih događaja, prikazanih na slici 1, a važnih za odabir odgovarajuće metodike i načina poučavanja.

Skretanje pažnje važno je kako bi se polaznik usredotočio na ono što će učiti. Skretanje pažnje može biti zanimljiva priča, demonstracija mogućnosti, prezentiranje problema koji treba riješiti, učiniti nešto na krivi način namjerno (kasnije će instrukcija pokazati pravilan način) ili samo naglasiti značaj gradiva koje slijedi. Upoznavanje s ciljem pomaže polazniku da organizira svoje misli i pripremi se na ono što slijedi, što će vidjeti, čuti i činiti. Prisjećanje na prethodno naučeno pomaže polazniku da gradi svoje znanje na prethodno naučeno znanje i vještinu. Informacije koje se prezentiraju potrebno je rascjepkati kako ne bi prenatrpali kratkoročno pamćenje. Korištenje nekoliko kanala prenošenja informacija odnosno prenošenje iste informacije na nekoliko različitih načina poboljšava pamćenje. Davanjem referenci za učenje olakšava se također pamćenje informacija jer se poziva se na poznate činjenice.

Aktivno sudjelovanje veoma je važan događaj i bitan za učenje. Dozvoliti polazniku da aktivno sudjeluje u učenju, tj. da napravi neke potrebne akcije svakako da poboljšava njegovu vještinu.

Slika 1. Nastavni događaji

Povratna informacija omogućuje analizu polaznikove uspješnosti. Provjera naučenog je testiranje znanje kako bi se utvrdilo da je cilj savladan. Transfer znanja je također bitan jer informira polaznika o sličnim problemima, situacijama, mogućnosti primjene naučenog gradiva.

Edukativna aktivnost ne samo da mora biti prikladna postavljenom cilju već mora i angažirati polaznika. Vrlo je važno da polaznik vidi direktnu i osobnu vrijednost u onome čemu se poučava. To se postiže upotrebom realnih primjera koji se susreću u svakodnevnom životu, ali i upotrebom nastavnih događaja. U protivnom polaznik brzo gubi interes jer smatra da mu predočen sadržaj nije potreban ili ne ispuni svoj cilj u potpunosti.

A. Organizacija ECDL E-learning sadržaja

Kod ECDL *E–learning* seminara nastavne jedinice su propisane *Syllabusom* odnosno one su moduli ECDL-a koji imaju točno definiran cilj:

- modul 1 - razumijevanje osnovnih pojmova informacijske i komunikacijske tehnologije (ICT) te poznavanje dijelova računala;
- modul 2 - dokazivanje sposobnosti u radu i korištenju osobnog računala;
- modul 3 - dokazivanje sposobnost za korištenje aplikacija za obradu riječi – teksta za kreiranje svakodnevnih pisama i dokumenata;
- modul 4 - razumijevanje osnova proračunskih tablica i demonstracija sposobnosti korištenja proračunskih tablica za preciznu izradu izlaznih rezultata;
- modul 5 - razumijevanje osnovnih pojmova baza podataka i demonstracija sposobnosti korištenja baza podataka;
- modul 6 - dokazivanje sposobnosti korištenja prezentacijskog softvera;
- modul 7 - poznavanje interneta i korištenje programa za pregledavanje te razumijevanje pojmova povezanih s elektroničkom poštom (*e-mail*) i poznavanje drugih mogućnosti komunikacije.

Poglavlja unutar nastavnih jedinica određuje sam kreator nastavnog sadržaja vodeći se navedenim ciljevima svake nastavne jedinice odnosno modula zadanog *Syllabusom.*

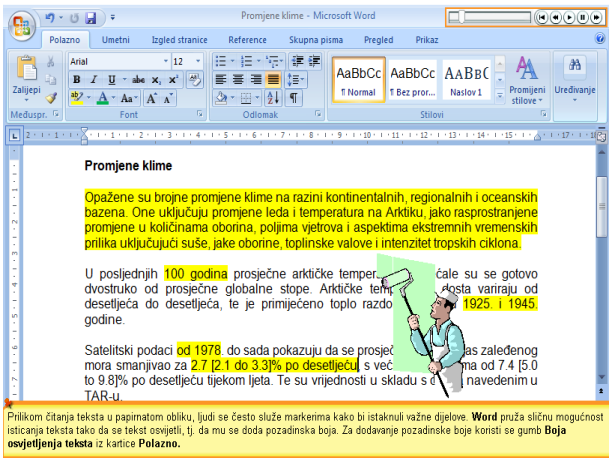
Cilj modula 1 je razumijevanje osnovnih pojmova informacijske i komunikacijske tehnologije (ICT) te poznavanje dijelova računala što je dosta teoretsko gradivo u kojem je cilj ovladati definicijama i opisima više nego praktičnom primjenom. Osim modula 1 kod kojeg su razina 1 (znanje) i razina 2 (razumijevanje) najviše izražene i u drugim modulima postoje poglavlja koja također traže samo znanje i razumijevanje pojmova. U svima takvim poglavljima sadržaj je organiziran tako da se svodi se na prikaz pojmova i definicija koje bi polaznik trebao upamtiti. Prikaz informacija treba biti sustavan,

Slika 2. Primjer teoretskog sadržaja

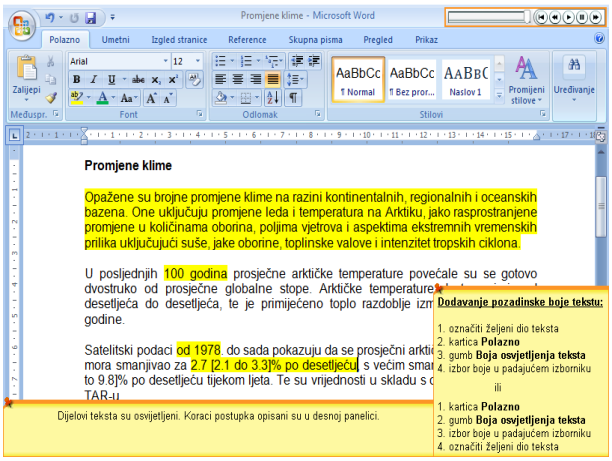
logički jasan, zanimljiv te na standardnom hrvatskom jeziku. Slika 2. prikazuje primjer *E-learning* sadržaja jedne takve teme.

Najzastupljenija razina znanja koja je potrebna za savladavanje ECDL certifikata je razina 3 (primjena). Da bi polaznici postigli razinu 3 (primjena) važno je da osim što im se demonstrira kako se primjenjuje određen alat da oni i aktivno sudjeluju odnosno da sadržaj prati svih devet koraka instrukcijskog dizajna. Pravilno kreiran sadržaj osim što pozitivno utječe na usvajanje znanja, doprinosi i razvitku polaznikovih sposobnosti.

Na slici 3. prikazan je sadržaj kojem je cilj postizanje razine 3 (primjena) odnosno konkretno u ovom primjeru za cilj postavlja primjenu alata za dodavanje pozadinske boje. Kao što i već iz same slike možemo vidjeti taj sadržaj prati svih devet koraka instrukcijskog dizajna koji preporučuje Gange. Skretanje pažnje postignuto je grafičkim prikazom čovjeka koji boji površinu. Upoznavanje s ciljem i prisjećanje na prethodno naučeno osigurano je uvodnim tekstom u temu. Informacije koje se prezentiraju su rascjepkane na nekoliko slajdova te se osim u tekstualnom obliku prenose i zvučnim putem. Referenciranje je postignuto analogijom s obilježavanjem markerom u tekstu. Slika 4 prikazuje aktivno sudjelovanje polaznika koji treba demonstrirati postupak prema navedenim koracima koji ujedno predstavljaju i provjeru naučenog.



Slika 3. Primjer praktičnog sadržaja



Slika 4. Primjer praktičnog sadržaja

Sustav sam bilježi napredovanje polaznika i ispravnost njegove aktivnosti tako da se kasnije to može analizirati. Kao što je već rečeno, bitna je sistematičnost i postupnost u objašnjavanju alata.

III. ALATI KORIŠTENI ZA IZRADU I ODRŽAVANJE E – LEARNING SADRŽAJA

Trenutno postoji nekoliko različitih definicija alata za e-obrazovanje (engl. *Courseware tool*), no još uvijek ne postoji jedinstvena definicija. Bez obzira na definiciju opće prihvaćena činjenica je da se *E–learning* sustav sastoji od dva osnovna elementa:

- LMS-a (engl. *Learning Management System*) i
- nastavnog materijala

A. Alati za izradu nastavnog materijala

Za izradu nastavnog materijala koji u potpunosti zadovoljava metodologiju izrade *E–learning* sadržaja opisane u prethodnom poglavlju koriste se dva alata: *Articulate Engage* i *Macromedia Captivate.*

Articulate Engage je jednostavan alat za brzu i jednostavnu izradu interaktivnih *flash* animacija koje vode polaznika kroz teorijski dio nastavnog materijala. Upotrebom gotovih predložaka vrlo jednostavno se stvarju bogati sadržaji u samo nekoliko minuta. Nastavni sadržaj stvara se odabirom između različitih tipova interakcije uključujući vremenski oblik, piramidalni prikaz, grafički padajući izbornici i mnogi drugi oblici. Završni oblik izrade materijala je poznata *flash* animacija koju je moguće umetnuti u gotovo bilo koju vrstu *E–learning* sadržaja ili čak može sama stajat na nekoj od web stranica.

Proces interakcije omogućuje polaznicima da otkriju korake linearnog procesa. Pomoću ovog tipa sadržaja daje se naglasak na ilustraciju koraka procesa, prolazak polaznika kroz niz povezanih stavki te objašnjavanje tijeka procesa ili određenog postupka.

Grafički naslovi omogućuju korisnicima interaktivnu identifikaciju ključnih elemenata slike nastavnog sadržaja. Koriste se za isticanje bitnih elemenata slike, objašnjenje pojedinosti određenog prozora korisničkog programa te isticanje bitnih elemenata izbornika i dijaloških okvira

Interaktivne kartice omogućuju istraživanje grupe srodnih predmeta. Uglavnom se koriste za prolazak kroz popis srodnih pojmova i utvrđivanja članova grupe.

Vremenska linija omogućuje interaktivnu kretnju kroz određeni vremenski period događanja. Najčešće se koristi za prikazivanje prolaska kroz određeni vremenski period te objašnjenje određenih obveza i događaja. Izgled sadržaja stvorenog u ovom alatu prikazan je na slici 2.

Macromedia Captivate je alat za maksimalno automatizirano i brzo pripremanje animiranih sekvenci koje će polazniku pokazati kako se u bilo kojem softveru obavlja neka procedura. Za izradu samog sadržaja dovoljno je učitati zadani program, odrediti dio ekrana koji će se pratiti i odabrati tipku *Record*. Time *Captivate* snima sve pokrete mišem, klikove i rad na tipkovnici, kao i ponašanje aplikacije na kojoj se radi. Po završetku snimanja program automatski razdijeliti obavljeni posao na više slajdova, od kojih svaki ima animirani pokazivač i oblačiće koji se pojavljuju i govore što u danom trenutku treba obaviti.

Takav način izrade sam po sebi štedi dosta vremena i truda. Pored samog prikaza, program je u stanju napraviti i interaktivnu probu naučenog pri kojoj će se polazniku dati da obavi isti zadatak kao što mu je prethodno demonstrirano. Ukoliko nešto napravi netočno, automatski će na odgovarajućim mjestima prikazati oblačići s komentarima što treba i kako izvesti.

Osim ova dva načina postoji i treća mogućnost, a to je izrada ispita kod kojega je polazniku ostavljeno da bez asistencije obavi zadatak, a računalo prati njegov rad i bilježi eventualne greške. Rezultati tog ispita, zajedno s analizom grešaka pohranjuju se u bazu podataka i kasnije se mogu analizirati.

Glavna razlika u odnosu na tipične prezentacijske programe je prisutnost preciznog vremenskog dijagrama događaja, koji se nalazi iznad glavnog predloška. Ovdje se podešavaju sve tranzicije između slika, pojavljivanje i pomicanje elemenata, reprodukcija zvučnih zapisa i slično. Primjer sadržaja stvorenog u ovom alatu prikazan je na slici 3 i slici 4.

B. Alat za održavanje E – learning sadržaja

LMS je komplet standardiziranih komponenti za učenje sa ciljem povezivanja nastavnog sadržaja i informatičkog sustava. Jedna od takvih aplikacija, a koja u potpunosti zadovoljava potrebe organizacije i održavanja ECDL sadržaja je *Moodle* (engl. *Modular Object-Oriented Dynamic Learning Environment*).

Moodle je aplikacija otvorenog kôda (engl. *Open source*) za izradu i održavanje *online* kolegija putem Interneta. Projekt se kontinuirano razvija s namjenom potpore obrazovnom okruženju društvenog konstrukcionizma.

Moodle je Web aplikacija napisana u PHP-u, a podržava više vrsta baza podataka (posebno *MySQL* i *PostgreSQL*). Sučelje je prevedeno na 65 jezika, a *Moodle* se trenutno koristi u 163 zemlje. Kako je *Moodle* projekt otvorenog koda, ima brojnu zajednicu korisnika (oko 150 tisuća registriranih korisnika) te izvrsnu dokumentaciju i podršku.

Mnogi korisnici aktivno sudjeluju u stvaranju *Moodlea*, radom na postojećim dijelovima, izradom novih modula, testiranjem proizvoda ili pružanjem podrške korisnicima.

Moodle sustav za upravljanje učenjem pruža punu računalnu podršku pri organizaciji i izvođenju *online* sadržaja. Neke od važnijih mogućnosti *Moodlea* su:

- izrada velikog broja tečaja na jednom sustavu
- planiranje tečaja – raspored aktivnosti, kalendar
- upravljanje korisnicima, korisničkim ulogama i grupama polaznika na tečaju
- rad s već postojećim datotekama i obrazovnim sadržajima
- provjera znanja i ocjenjivanje polaznika
- praćenje aktivnosti polaznika
- mnogobrojni alati za komunikaciju i kolaboraciju među polaznika
- upravljanje sustavom – sigurnosne kopije, statistike, logovi
- opsežan sustav pomoći

IV. ZAKLJUČAK

ECDL *Syllabus* propisuje koja znanja i vještine osoba treba posjedovati da bi položila ECDL certifikat. Osoba se za polaganje ECDL ispita može pripremati pohađajući prilagođen *E–learning* sadržaj. Da bi takav *E–learning* sadržaju u potpunosti bio adekvatan za pripremu za polaganje certifikata kreatori sadržaja trebaju se voditi Bloomovom taksonomijom i poštivati Gangeovih devet koraka instrukcijskog dizajna. Nastavni sadržaj ne samo da mora biti prikladna postavljenom cilju već mora i angažirati polaznika te mu pokazati direktnu i osobnu vrijednost u onome čemu se poučava. Alati koji se koriste za izradu takvih sadržaja dijele se na alate za izradu nastavnog sadržaja i alate za održavanje nastavnog sadržaja. Adekvatan alat za izradu teoretskog dijela nastavnog materijala je *Articulate Engage* dok je za izradu nastavnog materijala kojem je u cilju postizanje Bloomove razine 3 (primjena) puno prikladniji alat *Macromedia Captivate*. Kako bi se povezao nastavni sadržaj i informatički sustav koristi se *Moodle*. On osim što omogućuje povezivanje sadržaja omogućuje i jednostavno upravljanje sustavom od strane administratora te praćenje uspješnosti polaznika što je važno kako bi se polazniku dala informacija da li je spreman za polaganje ECDL ispita.

LITERATURA

[1] S. Jurak, " Metodika nastave informatike u svrhu dobivanja ECDL certifikata, MIPRO 2010/CE, pp. 446–448, Svibanj, 2010

[2] Gagne, R. M, Wager, W. W., Golas, K. C. & Keller, J. M. (2005). *Principles of Instructional Design, 5th ed.* Belmont, CA: Wadsworth/ Thomas Learning

[3] Bloom, B.S., (1956) Taxonomy of educational objectives, Longman, NY

[4] ECDL, http://www.algebra.hr/ECDL_Certifikacija.asp, 31.1.2013

[5] ECDL, http://www.ecdl.org/publisher/index.jsp , 31.1.2013

[6] Algebra, http://lms.algebra.hr, 1.2.2013.

[7] Moodle, http://moodle.org/, 1.2.2013

Primjena testova s bodovanjem zasnovanim na sigurnosti ispitanika u ispravnost odgovora

B. Nožica, Ž. Širanović i M. Čarapina
Tehničko veleučilište u Zagrebu / Informatičko-računarski odjel, Zagreb, Hrvatska

bojan.nozica@tvz.hr, zeljko.siranovic@tvz.hr, mia.carapina@tvz.hr

Sažetak - Provjere znanja putem računalno podržanih testova postale su uobičajen način provođenja završnih pismenih ispita, na temelju kojih se sve češće, bez dodatnih ispitivanja, zaključuje i konačna ocjena. Znajući da se u takvim testovima uglavnom nude gotovi odgovori, mnogi studenti oslanjaju se na površno znanje birajući među najvjerojatnijim odgovorima i uzdajući se u dobru sreću. Nastojeći pridobiti studente na temeljitost u procesu učenja, nastavnici predmeta „Osnove programiranja" na Stručnom studiju informatike Tehničkog veleučilišta u Zagrebu primijenili su provjere znanja na početku i na kraju svake laboratorijske vježbe, koristeći u testovima opciju CBM (eng. Certainty Based Marking) u sustavu Moodle. Uključivanjem te opcije bodovi za odabrani odgovor izračunavaju se obzirom na ispravnost i razinu sigurnosti koju je pri tome student deklarirao (nisku, srednju ili visoku), pri čemu bodovi za točan odgovor rastu sa stupnjem sigurnosti u pozitivnom, dok za netočan rastu u negativnom smjeru. U radu se analiziraju prednosti i nedostaci primijenjene metode, te se predlažu i neke od mogućih modifikacija.

I. UVOD

Mjerenje znanja pomoću testova koji sadrže pitanja s više ponuđenih odgovora (eng. multiple choice question – MCQ) postaje sve raširenije, između ostalog i zbog toga što je prilagođeno računalnoj okolini. Postoje mnogi programski alati za izradu i obradu takvih testova, no u ovom članku ograničiti ćemo se na mogućnosti koje pruža Moodle, u svijetu vrlo rasprostranjen LMS (eng. Learning management system), sustav za upravljanje procesima i sadržajima vezanim uz učenje. Nakon kreiranja i podešavanja postavki testa kojima se upravlja distribucijom, izvođenjem i vrednovanjem rezultata, evidencija konačnih rezultata i razne statističke analize dostupne su trenutno po završetku provedenog testiranja. Na taj način otvara se, između ostalog, mogućnost realizacije vrlo kratkih testova u nastavi koja se izvodi u suvremenoj računalnoj učionici (realnoj ili virtualnoj), gdje svaki korisnik ima na raspolaganju umreženo računalo. Takvi kratki testovi (u školskom žargonu popularno zvani „blicevi") do sada su se i u klasičnoj nastavi znali koristiti, najčešće na kraju neke nastavne jedinice, uobičajeno kao metodički alat za utvrđivanje nedavno obrađenog gradiva. Nastavnik bi po obradi rezultata mogao npr. steći uvid u to koliko su studenti razumjeli novo gradivo. Ukoliko bi, međutim, nastavnik na početku slijedeće nastavne jedinice htio primijeniti sličan kratki test za uvid u to koliko su studenti usvojili prethodno obrađeno gradivo, nakon što su dodatno samostalno proučili preporučenu literaturu, najvjerojatnije bi zaključio kako je ukupno vrijeme potrebno za podjelu testova, vrijeme za rješavanje testa, te na kraju vrijeme za prikupljanje istih i barem letimični pregled i analizu danih odgovora, predugo. S druge strane, korištenjem Moodle-a ili sličnog alata, ukupno vrijeme potrebno za sve navedene radnje, uključivo i detaljnu statističku analizu rezultata, gotovo je jednako vremenu potrebnom za rješavanje testa, odnosno vremenu koje je nastavnik unaprijed mogao ograničiti u postavkama testa. Na primjer, za test koji sadrži pet pitanja, koji je zaštićen lozinkom i ima ograničeno vrijeme za rješavanje od tri minute, dovoljno je za realizaciju predvidjeti 4 minute – pola minute od objave lozinke dovoljno je da svi studenti učitaju na svoja računala test (koji opcionalno svakome od njih dolazi s izmiješanim redoslijedom pitanja i odgovora), tri minute za rješavanje testa, te otprilike još pola minute za uvid nastavnika u detaljnu statističku analizu danih odgovora. Ukoliko je nastavnik uz to postavio da se rezultati testa boduju, svakom studentu biti će automatski pridruženi ostvareni bodovi ili ocjena u kategoriji koju nastavnik također može sam odrediti. Može se reći da je nastavnik tako dobio vrlo efikasan alat kojeg može upotrijebiti u razne didaktičke svrhe, jednako kao i za formalno mjerenje znanja koje rezultira ocjenom.

Nažalost, nedostaci koji su već poznati kod testova s unaprijed ponuđenim odgovorima [2], povećavaju se obrnuto proporcionalno s brojem pitanja u testu. Naime, kratki testovi manje pouzdano mjere stvarno znanje jer vjerojatnost pogađanja točnog odgovora (koja se, ovisno o broju distraktora, uobičajeno kreće od 20% pa sve do 50%) značajnije utječe na konačni rezultat, nego kod testova s duljim nizom pitanja (vjerojatnost pogađanja svih točnih odgovora u testu jednaka je umnošku vjerojatnosti pogađanja na pojedinačno pitanje) . Treba imati na umu da je teško kreirati pitanja s idealnim distraktorima, odnosno netočnim odgovorima koji će se s jedne strane dovoljno bitno razlikovati od točnog odgovora, a s druge strane izgledati jednako vjerojatni kao i točan odgovor studentu koji nema dovoljno znanja da ih raspozna. To u praksi znači da je vjerojatnost pogađanja gotovo uvijek veća od nominalne (100/broj ponuđenih odgovora). Primjerice, ukoliko imamo pitanje s pet ponuđenih odgovora, no dva netočna odgovora nisu zavarala gotovo nikoga, tada vjerojatnost pogađanja točnog odgovora na to pitanje nije matematičkih 20% već je u stvarnosti bliskija 33%. Nije rijetkost ni da student s vrlo malim znanjem metodom eliminacije loših distraktora podigne vjerojatnost pogađanja sve do 50%. Naravno, treba imati u vidu da pitanja s odgovorima točno/netočno

po definiciji imaju vjerojatnost pogađanja 50%. Pri tome ne zabrinjava toliko činjenica da to utiče na statističko odstupanje rezultata od stvarnog znanja, koliko nekritičko ponašanje studenata koji nerijetko, nakon što pomoću sretno odabranih točnih odgovora postignu povoljan rezultat na testu, postaju uvjereni da su doista savladali predmetno gradivo.

Nastojeći odvratiti studente od pogađanja odgovora na testovima, nastavnici znaju poseći za negativnim bodovanjem netočnih odgovora, što je uglavnom nepopularno kod studenata ukoliko ne postoji postavljen sustav ocjenjivanja kojeg studenti doživljavaju pravednim. U konačnici testovi s negativnim bodovanjem postižu s istim brojem pitanja veći raspon bodova u grupi, a time i preciznije mjerenja znanja. Klasičan sustav „pravednog" negativnog bodovanja primjenjuje princip oduzimanja bodova u odnosu na broj distraktora (netočnih odgovora) među ponuđenim odgovorima. Tako se za pitanje s 4 distraktora oduzima 25% od ukupnog broja bodova na pitanju, za pitanje s 3 distraktora 33%, a za pitanja formata točno/netočno oduzima se 100%. Jedan od načina negativnog bodovanja, opisan u članku, oslanja se na samoprocjenu sigurnosti ispitanika u odabrani odgovor. U testovima gdje se želi primijeniti takav način bodovanja ispitanik je dužan, uz odgovor na osnovno pitanje, odgovoriti i na potpitanje tipa: „Koliko ste sigurni u to što tvrdite?".

To je pitanje inače uobičajeno postaviti u životnim situacijama kada je važno dobiti pouzdanu informaciju. Primjerice, liječnik koji priopćuje pacijentu neugodnu dijagnozu mora biti spreman odgovoriti i na pitanje koliko je u takvu dijagnozu siguran. Ljudi su u takvim situacijama često spremni zatražiti i mišljenje drugog neovisnog stručnjaka, stoga reputacija i uspjeh mogu bitno ovisiti o odgovoru na takvo pitanje. Tvrdnja za koju se ustanovi da je bila pogrešna, a da pri njenom izricanju nije ostavljen ni najmanji prostor za sumnju, može u stvarnom životu skupo stajati pojedinca, a ponekad i čitavu organizaciju koja stoji iza nje. S druge strane, tvrdnja koja je popraćena iskrenim priznanjem da su mogući i drugi ishodi, biti će općenito u društvu podložna blažim sankcijama.

Sličan princip nagrađivanja kritičkog razmišljanja, odnosno sankcioniranja uvjerenosti u pogrešno rješenje, primijenjen je u sustavu negativnog bodovanja koji je primijenjen na predmetima prve godine studija „Osnove programiranja" i „Programiranje" na Tehničkom veleučilištu u Zagrebu (TVZ). Takvo bodovanje koje se temelji na samoprocjenjenoj sigurnosti u dati odgovor poznato je pod engleskim nazivima Confidence Based Marking, odnosno Certainity Based Marking (CBM) [1].

II. TESTOVI S CBM OPCIJOM

A. Znanje i vjerovanje

Poznato je da znanje ne spada u kategoriju opipljivog, a pored toga ne postoji ni jedinstvena opće prihvaćena definicija znanja, pa je prema tome svaki pokušaj da se izmjeri nečije znanje unaprijed osuđeno na veće ili manje propuste. Ukoliko mjerimo samo točnost nečijeg odgovora na pitanje, bez poznavanja okolnosti pod kojima ga je ispitanik odabrao, može se reći da mjerimo samo jednu dimenziju znanja. Nažalost, većina testova znanja prilagođena računalnoj obradi ne može izmjeriti ništa više od toga. Međutim, uvođenjem potpitanja „Koliko ste uvjereni (sigurni) u svoj odgovor?" dobivamo, barem približno, uvid u još jednu dimenziju ispitanikovog znanja, čime se približavamo mjerenju znanja prema Platonu kao „istinitog, opravdanog vjerovanja". Analizirajući odgovor na pitanje i pripadajuću razinu uvjerenja (sigurnosti) koju je ispitanik deklarirao, možemo rezultat svrstati u neku od slijedećih kategorija [1]:

- Znanje
- Nesigurnost
- Neznanje
- Pogrešno shvaćanje
- Obmanutost

Pri tome pod znanjem priznajemo kombinaciju točnog odgovora uz iskazanu visoku sigurnost, u nesigurnost svrstavamo kombinaciju točnog odgovora i nižih razina sigurnosti, dok neznanjem možemo okarakterizirati i točan i netočan odgovor uz iskazanu razinu sigurnosti koja je bliska ili manja od matematičke vjerojatnosti za pogađanje točnog odgovora. Pod pogrešnim shvaćanjem kategoriziramo kombinaciju netočnog odgovora uz iskazanu sigurnost koja je veća od matematičke vjerojatnosti za pogađanje točnog odgovora, dok kombinaciju netočnog odgovora i pri tome iskazane visoke sigurnosti u odgovor možemo svrstati u kategoriju obmanutosti. Ovoj posljednjoj kategoriji često je potrebno posvetiti najviše pažnje, jer se radi o potencijalno opasnom znanju kada se primjenjuje u praksi.

B. Bodovna shema primijenjena na University College London

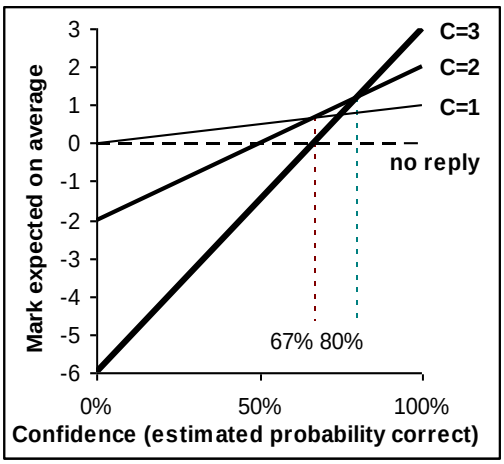
Kada se radi o mjerenju znanja pomoću testova s CBM opcijom, potrebno je razraditi bodovnu shemu koja će motivirati studente da kritički i iskreno deklariraju razinu sigurnosti koja prati njihov odgovor. U [1] je opisana bodovna shema koja se primjenjuje na Odjelu za fiziologiju pri University College London (UCL).

UCL Confidence-based scoring scheme

Confidence Level	**1**	**2**	**3**
Score if Correct	**1**	**2**	**3**
Score if incorrect	**0**	**-2**	**-6**
Probability correct	**< 67%**	**>67%**	**>80%**
Odds	**< 2:1**		**>2:1**

Slika 1. – Bodovna shema prema [1] koja se primjenjuje na UCL

.



Slika 2. – grafički prikaz očekivanog prosječnog broja bodova (po pitanju), za svaku od ponuđenih razina sigurnosti u odgovor, prema bodovnom sustavu na UCL [1].

Studenti pri odgovoru na pitanje rangiraju svoju razinu sigurnosti (eng. confidence ili certainty level) u ispravnost odgovora kao nisku (C=1), srednju (C=2) ili visoku (C=3) (slika 1.)

Da bi razumjeli razloge za ovako postavljenu bodovnu shemu, potrebno je istaknuti da se velika većina testova na kojima se ona pri UCL-u primjenjuje sastoji od pitanja s odgovorima točno/netočno, na kojima je vjerojatnost pogađanja točnog odgovora 50%. To u praksi znači da student koji pri odgovoru na pitanje takvog tipa deklarira razinu sigurnosti 1 (C=1, sigurnost <67%) u stvari priznaje da tek nagađa odgovor, te da bi s istom razinom sigurnosti mogao izabrati i dijametralno suprotni odgovor. Iz grafa na slici 2. je vidljivo da će upravo izbor te razine sigurnosti dati najbolji prosječni rezultat za raspon vjerojatnosti manji od 67%. Ukoliko student ne dvoji oko točnog odgovora, vidljivo je da je u tom slučaju najbolje odabrati C=3 (sigurnost > 80%), dok je u slučajevima kada ipak postoji određena nesigurnost, najbolje odabrati razinu sigurnosti 2 (C=2, sigurnost između 67% i 80%). U navedenoj bodovnoj shemi nikada nije optimalno izostaviti odgovor, budući razina sigurnosti C=1 nosi mogućnost osvajanja boda pri točnom odgovoru, dok za netočni nema negativnih bodova [1]. Prema [1], rezultati su pokazali da studenti mogu, nakon određenog vremena prilagođavanja, intuitivno birati upravo optimalnu razinu sigurnosti, dakle onu koja za procijenjenu vjerojatnost da će odgovor biti točan daje najbolji prosječni rezultat.

Prema [1], ovako postavljena jednostavna bodovna skala univerzalno je primjenjiva za sve vrste pitanja s odgovorima koji se mogu jednoznačno odrediti kao točni, odnosno netočni. Ipak, navodi se kako u testovima čiji rezultati ulaze u izračun konačne ocjene, na UCL-u koriste T/F pitanja (eng. True/False question - pitanje u obliku tvrdnje na koje su mogući odgovori isključivo točno/netočno).

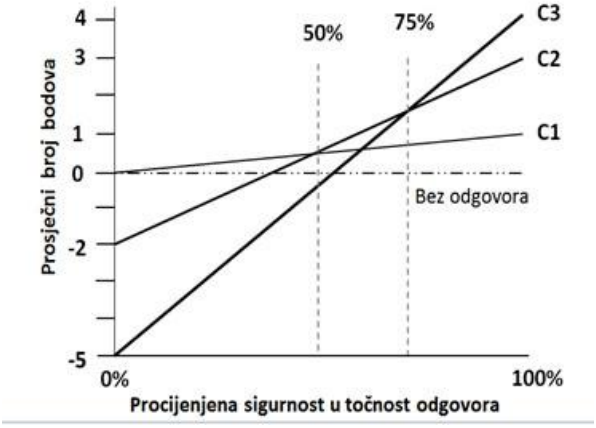
C. Bodovna shema primijenjena na Tehničkom veleučilištu u Zagrebu

Grupa nastavnika na TVZ-u koja je pomnije proučila UCL bodovnu shemu zaključila je da se, bez obzira na neospornu univerzalnost, radi o shemi koja je optimirana za testove s T/F pitanjima. Kako većina pitanja koja se koristi na testovima u kojima su htjeli uključiti CBM opciju ima oblik MCQ pitanja, smatrali su da postoji slična univerzalna shema koja bi bila optimirana za MCQ testove. Pored toga, odnos negativnih i pozitivnih bodova pri razini C=3 činio se neopravdano velik (2:1), što bi u kratkim testovima prouzročilo da već i samo jedan pogrešan odgovor kobno utječe na rezultat. Primjerice, ukoliko bi na testu s 5 pitanja student odgovorio na svih pet pitanja s najvišom razinom sigurnosti, samo jedan pogrešan odgovor značio bi gubitak polovice od ukupnog broja bodova na testu. Na testu s T/F pitanjima to možda ima opravdanja, no na testu s MCQ pitanjima vjerojatno bi izazvalo opravdano negodovanje studenata, što se htjelo izbjeći.

Poštujući 3 glavna principa na kojima se temelji originalna UCL bodovna shema (jednostavnost shvaćanja, ispravno motiviranje, fleksibilnost), na TVZ-u je postavljena bodovna shema kao na slici 3. Ispravno motiviranje i jednostavno shvaćanje sheme postignuto je time što student, na temelju vlastite procjene vjerojatnosti da je odabrani odgovor točan, lako izabire optimalnu razinu sigurnosti.

TVZ bodovna shema za testove s CBM opcijom

Razina sigurnosti	C1 <50%	C2 50-75%	C3 >75%
Bodovi za točan odgovor	**1**	**3**	**4**
Bodovi za netočan odgovor	**0**	**-2**	**-5**



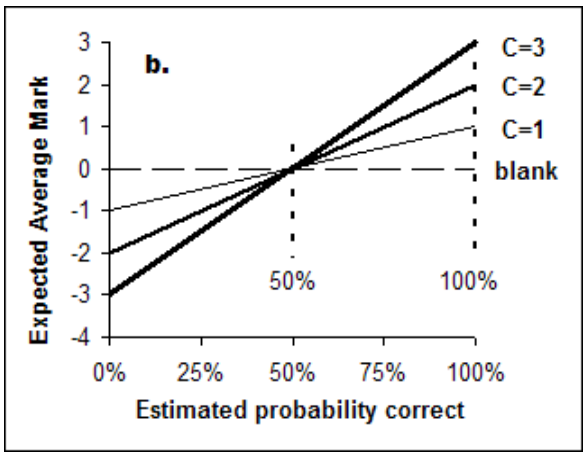
Slika 3. – Tabelarni i grafički prikaz modificirane bodovne sheme koja se koristi na Tehničkom veleučilištu u Zagrebu

Studentima je poželjno otvoreno objasniti i preporučiti strategiju biranja određene razine sigurnosti s ciljem poticanja kritičkog i iskrenog promišljanja o vlastitom znanju izraženog kroz odgovor na dodatno pitanje o sigurnosti u odgovor. Primjerice, studentima je moguće izložiti slučaj u kojem moraju odgovarati na pitanje s pet ponuđenih odgovora. Ukoliko se pri odabiranju odgovora uspije eliminirati dva odgovora, ali se među preostala tri svi čine jednako vjerojatni, tada je jasno da je vjerojatnost odabiranja točnog odgovora ispod 50%, stoga je, prema grafu na slici 3. optimalno odabrati razinu sigurnosti C1. U slučaju da je točan odgovor odmah prepoznat, ili se pak metodom eliminacije ostalih odgovora zaključilo koji je odgovor točan, tada nema dvojbe da treba odabrati razinu C3. U svim preostalim slučajevima srednja razina sigurnosti C2 predstavlja statistički najbolji izbor.

Fleksibilnost TVZ sheme očituje se u tome što je primjenjiva na sve vrste pitanja, a zbog blažeg omjera negativnih i pozitivnih bodova primjerena je i kratkim testovima. Istina, zbog toga, a i zbog činjenice da razine sigurnosti C1 i C2 kod vjerojatnosti od točno 50% daju statistički isti rezultat moglo bi se reći da se u testovima s T/F pitanjima otvorio prostor za nešto rizičnije ponašanje u odnosu na originalnu UCL shemu.

D. Ostale bodovne sheme

Nisu sve bodovne sheme koje se mogu naći u literaturi o CBM-u ispravno motivirajuće [3]. Primjer je takozvana balansirana shema, prikazana grafom na slici 3, u kojoj se za svaku razinu sigurnosti dodjeljuje simetričan broj pozitivnih i negativnih bodova, ovisno o tome da li je odgovor bio točan ili ne. Analiza grafa pokazuje da nikad nije racionalno za procijenjenu sigurnost u odgovor veću od 50% birati manju razinu od C=3, dok je za sigurnost manju od 50% optimalno ponašanje izbjeći davanje odgovora.



Slika 3. Primjer bodovne sheme koja ne motivira ispravno

III. REZULTATI PRIMJENE CBM OPCIJE U TESTOVIMA NA TEHNIČKOM VELEUČILIŠTU U ZAGREBU

S ciljem poboljšanja uspjeha studenata u savladavanju gradiva na predmetima iz programiranja u jeziku C, uvedeni su na početku i na kraju svakog bloka vježbi (u trajanju od 3 školska sata) kratki testovi koji su provođeni pomoću sustava Moodle. Također, odlučeno je da se ti testovi boduju i računaju pri izračunu završne ocjene na predmetu. Ukupno je na svim takvim testovima tijekom semestra bilo moguće ostvariti do 28% od maksimalnog broja bodova na predmetu, što je po svakom testu iznosilo tek nešto više od 1% ukupnog broja bodova.

Potrebno je napomenuti kako kratki testovi o kojima je ovdje riječ prije svega služe kao edukacijski alat. Analizirajući pred studentima njihove pogrešne odgovore s ulaznog testa (koji su u pravilu trajali 7 minuta), bilo je moguće identificirati do koje razine su studenti uspjeli savladati gradivo s prethodnog predavanja i pripremne interaktivne lekcije za vježbu, te shodno tome ponuditi dodatna pojašnjenja. Pred sam kraj vježbe, nakon praktičnog dijela u kojem su studenti samostalno rješavali zadatke iz istog gradiva, uslijedio bi izlazni test, pri čemu su pitanja bila iz istih kategorija kao i pitanja s ulaznog testa (sa stanovišta statističke analize, pitanja s ulaznog testa ponavljaju se na izlaznom, jedino je njihov podskup za svakog studenta drugačiji). Kako bi izlazni test dobio veći značaj, za njega je postavljen nešto veći maksimalni broj bodova. Pored toga, odlučeno je da na izlaznom testu bude uključena opcija CBM, kako bi se dobila i dodatna informacija o tome koliko su studenti sigurni u svoj odabir odgovora.

Rezultati izlaznih testova uz uključenu CBM opciju bili su očekivano nešto slabiji od ulaznih testova, iako su studenti nesumnjivo bili spremniji na kraju svakog bloka vježbi. Razlog su negativni bodovi za pogrešne odgovore, kao i smanjeni broj bodova za točne odgovore dane uz razinu sigurnosti ispod najviše. Unatoč tome, prosječna razlika na svim vježbama iznosila je svega 6,34 postotna poena u korist ulaznih testova, dok je ukupni prosjek uspješnosti na svim testovima iznosio 67,56%. Ako se uzme u obzir da su rezultati pojedinog testa bez uključene CBM opcije za istog studenta i do 20 postotnih poena veći, to bi značilo da bi se ukupni prosjek bez uključene CBM opcije popeo preko 80%, što bi bio nerealno visok rezultat. Zanimljiva je i korelacija između uspjeha na blicevima i ukupnog uspjeha studenata na predmetu. Iako su blicevi iznosili samo 28% ukupnog broja mogućih bodova, prosječno odstupanje od ukupnog prosjeka (72,60%) iznosilo je svega 5,14 postotna poena. Svi rezultati odnose se na 186 studenata koji su pohađali vježbe iz predmeta „Osnove programiranja“ u akademskoj godini 2012/2013 i pri tome riješili ukupno 18 kratkih testova. Detaljna analiza i neke druge korelacije biti će predmet istraživanja koje će uslijediti nakon prikupljanja dodatnih podataka na predmetu „Programiranje“, kojeg upisuje uglavnom ista grupa studenata.

Možda je ipak od svega najzanimljivije da nije zabilježena niti jedna ozbiljna primjedba od strane studenata zbog uključivanja CBM opcije, što znači da studenti takav sustav bodovanja osjećaju kao opravdan i pravedan. Nadalje, suprotno očekivanjima, nije bilo potrebno produljiti vrijeme za rješavanje testova s CBM opcijom u odnosu na testove bez te opcije – pokazalo se da studenti već nakon kratkog vremena prilagođavanja za unos razine sigurnosti trebaju vrlo malo vremena (2-3 sekunde).

IV. ZAKLJUČAK

Primjena testova s bodovanjem zasnovanim na sigurnosti ispitanika u ispravnost odgovora, tzv. CBM (Certainity Based Marking), opcija je koju je odnedavno moguće uključiti na testovima koji se provode putem Moodle-a. Bodovna shema koja se pri tome primjenjuje razvijena je na UCL-u (University College London) [4], gdje se primjenjuje već više od 10 godina, uključivo i na formalnim testovima, no koji se sastoje isključivo od T/F pitanja. Zaključivši da originalna bodovna shema zbog optimiziranosti prema T/F pitanjima suviše rigorozno sankcionira netočne odgovore sa najvišom razinom sigurnosti, na Tehničkom veleučilištu u Zagrebu upotrjebljena je modificirana bodovna shema. Ta shema jednako ispravno motivira studente da iskreno i kritički ocjenjuju vlastitu sigurnost u odabrani odgovor na testovima s prevladavajućim MCQ pitanjima, uz manji broj negativnih bodova. Na taj način izbjegnut je otpor studenata, koji su ovaj sustav bodovanja prihvatili kao opravdan i pravedan, te je postignut realniji rezultat na kratkim testovima kojima se na kraju nastavnog bloka željelo ustanoviti u kojoj su mjeri studenti usvojili predmetno gradivo.

LITERATURA

[1] A.R. Gardner-Medwin and M. Gahan - Formative and Summative Confidence-Based Assessment, www.ucl.ac.uk/lapt/tea/caa03.doc

[2] Davies P. (2002) There's no confidence in Multiple-Choice Testing, ... Proceedings of the 6th International CAA Conference, Loughborough, pp 119-130

[3] Gardner-Medwin, A.R. (1999) Rational and irrational marking schemes. Journal of Physiology, 515P: 48P

[4] Gardner-Medwin, A.R. (2003) LAPT: London Agreed Protocol for Teaching. <http://www.ucl.ac.uk/lapt

Informacijske tehnologije u funkciji ekološkog obrazovanja učenika strukovnih škola

dr. sc. Vladimir Matić
Strukovna škola Vukovar, Hrvatska
vladimir.matic@vu.t-com.hr

Sažetak - Ekološko obrazovanje mladih, naročito učenika strukovnih škola na razini srednjeg obrazovanja, uporabom računala i ostalih informacijskih tehnologija (IT), relativno je novo i još dovoljno neistraženo područje. Istraživanje koje je proveo autor ovoga rada pokazalo je da je uporabom upravo tih tehnologija moguće otkloniti cijeli niz slabosti koje karakteriziraju tradicionalno ekološko obrazovanje učenika strukovnih škola. Tradicionalno obrazovanje podrazumijeva oblik poučavanja koji se zasniva na klasičnom predavanju. Osnovni je cilj takvog obrazovanja prenošenje znanja od nastavnika do učenika, pri čemu je nastavnik izvor znanja, a učenik pasivan primatelj znanja. Suvremeno ekološko obrazovanje, međutim, ne podrazumijeva samo jednostavnu reprodukciju znanja, već stvara pretpostavke za aktivno uključivanje učenika u proces stjecanja i usvajanja znanja, za što se koriste različiti izvori. Uporabom suvremenih IT učenicima se omogućuje da sami izaberu izvor znanja, ali i da prilagode dinamiku usvajanja znanja vlastitim intelektualnim mogućnostima.

Ključne riječi: računalo; IT; nastava; metode

I. UVOD

Državnim pedagoškim standardom srednjoškolskog sustava odgoja i obrazovanja u Republici Hrvatskoj ("Narodne novine", br. 63/08. i br. 90/10.) uređuju se jedinstveni uvjeti za kvalitetan odgojno-obrazovni rad i uspješno ostvarivanje ciljeva i zadataka u srednjoškolskim ustanovama. Prostorni uvjeti za izvođenje strukovnih programa ne propisuju multimedijalne učionice, ali propisuju informatičke učionice (praktikume) opremljene računalnom mrežom sa serverom i petnaest računala.

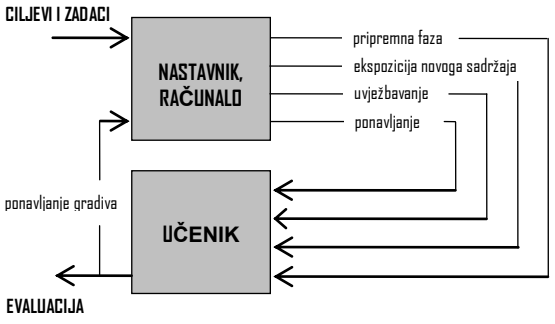
Računalna mreža predstavlja važan faktor informatičke podrške ekološkom obrazovanju učenika strukovnih škola i sastavni je dio ukupnog multimedijalnog informacijskog sustava. Suvremeno ekološko obrazovanje nezamislivo je bez Interneta. Ako se nastava ekologije izvodi u multimedijalnoj učionici, u kojoj postoji više međusobno umreženih računala, onda je omogućen istovremeni pristup Internetu svim učenicima. Ukoliko se u nastavi upotrebljavaju i obrazovni programi s ekološkim sadržajima primjerenim strukovnom obrazovanju, dovoljno ga je instalirati na samo jedno računalo (server), pa da mu istovremeno mogu pristupiti i koristiti ga svi učenici. Upravo ova dva primjera dovoljno govore o važnosti mrežnog okruženja u informatčkim učionicama.

Programska podrška primarno određuje okvire mogućnosti korištenja računala u odgojno-obrazovnom procesu. U odnosu na konvencionalna nastavna sredstva i pomagala, računala su inteligentna i programabilna. To znači da u njihovoj primjeni važnu ulogu imaju upravo obrazovni programi (software). Kvalitetan obrazovni software mora podržavati aktivno uključivanje učenika u nastavni proces, kako bi se računala mogla koristiti kao najbolja tehnika učenja. U tom smislu i software mora biti u funkciji primjene znanja, koji će razvijati mentalne kreativnosti učenika.

Slijedom navedenoga, moguće je zaključiti da kvalitetan, didaktički dobro osmišljen obrazovni software mora podržavati sljedeće razine učenja:

- razinu usvajanja novih nastavnih sadržaja (engl. Tutorial)
- razinu uvježbavanja rutinskih operacija (engl. Drill and Practice) i
- razinu nastavnog dijaloga (engl. Dialogue).

Dakle, software nedvojbeno određuje primjenu računala u nastavi i to u svim njenim fazama, kako je prikazano na sl. 1.



Slika 1. Faze nastave koje mora podržati obrazovni software

Nažalost, moramo konstatirati da kod nas gotovo i nema obrazovnog software-a namijenjenoga ekološkom obrazovnju, osobito ne ekološkom obrazovanju učenika strukovnih škola, pa su nastavnici prinuđeni isti razvijati sami. Druga je mogućnost nabaviti ga na svjetskom tržištu i prilagoditi vlastitim potrebama.

U kontekstu rečenoga, možda je najrealnije očekivati da će razvoj kvalitetnoga obrazovnog software-a, kojega je moguće implementirati u nastavu ekologije i ostalih predmeta u kojima se izučavaju ekološki sadržaji, morati inicirati sami nastavnici ekologije, pa i biti nositelji toga projekta, u kojega će uključiti i druge profile stručnjaka (informatičare, sociologe, pedagoge).

Učenje pomoću računala (e-learning) obuhvaća sve oblike elektronički podržanog **učenja** i **nastave**, koji su proceduralne prirode i imaju za cilj utjecati na izgradnju znanja u odnosu na individualno iskustvo, praksu i znanje učenika. Informacijsko-komunikacijski sustavi, bilo umreženi ili ne, služe kao specifični mediji u procesu učenja [9].

Osnovu komunikacije u tom sustavu omogućuje računalo s raznim servisima. Najpoznatiji među njima su servis za elektroničku poštu (e-mail) i software za pregledavanje i pretraživanje globalne računalne mreže Interneta (Microsoft Explorer, Mozila Firefox, Netscape).

E-learning omogućuje korisniku učenje na daljinu, pa ga u literaturi često nazivaju *učenje na daljinu* (engl. Distance Learning), a postao je oblik učenja korišten na najpoznatijim svjetskim sveučilištima (Harvard, Stenford). U odnosu na tradicionalni način učenja, ima mnoge prednosti:

- omogućuje korisniku izbor mjesta i vremena učenja
- omogućuje velike uštede u vremenu
- omogućuje korisniku pristup resursima znanja bez obzira na udaljenost (postoji mogućnost pristupa s različitih kontinenata)
- u znatnoj mjeri smanjuje troškove učenja, jer istovremeno može učiti velik broj korisnika koji su u odnosu na "izvor znanja" dislocirani, itd.

U ekološkom obrazovanju učenika strukovnih škola moguće je koristiti različite pojavne oblike učenja na daljinu, od kojih su najčešći: *e-mail učenje, računalna simulacija* i *elektronički udžbenici*.

Osnovna osobina *e-mail učenja* sadržana je u činjenici da se obrazovni sadržaji učeniku dostavljaju putem elektroničke pošte (e-maila). Dinamiku slanja dogovara nastavnik s učenikom, a ona ovisi o dinamici učenja pojedinaca. U ekološkom obrazovanju učenika strukovnih škola učestali je način komunikacije u obrazovanju odraslih. Komunikacija između nastavnika i polaznika odvija se putem tzv. *nastavnih pisama* koje izrađuju nastavnici i šalju ih e-mailom polaznicima. Budući da ta pisma sadrže i određen broj pitanja za ponavljanje gradiva, polaznici su u obvezi na njih odgovoriti i povratno ih dostaviti predmetnom nastavniku.

Nakon što je predmetni nastavnik ocijenio da je učenik savladao gradivo toga nastavnog pisma, šalje mu sljedeće nastavno pismo i tako, sekvencijski, dok polaznik potpuno ne savlada gradivo propisano nastavnim planom i okvirnim obrazovnim programom.

Računalna simulacija, u osnovi, podrazumijeva proces oponašanja (simuliranja) određenih zakonitosti, pojava ili događaja pomoću računala. Ona je u nastavi istovremeno i izvor i posrednik u prijenosu informacija, a njene didaktičke funkcije mogu biti različite, kao npr. vježbanje, motivacija itd.

U ekološkom obrazovanju učenika strukovnih škola računalna simulacija pruža posebne mogućnosti. Razlog je tome, između ostalog, što je sadržaje s područja ekologije lako moguće matematički modelirti (npr. simulacija onečišćenja rijeka ili jezera), a osnovu simulacije predstavlja određen, unaprijed odabran matematički i (ili) formalno-logički model, na temelju kojega se izrađuje algoritam simulacije [8].

Pri izboru sadržaja treba voditi računa da to budu sadržaji koji se mogu programirati, jer je programiranje sastavni dio računalne simulacije. Osim toga, dobro je da to budu sadržaji o kojima učenici imaju što manje predznanja, jer je jedino na taj način moguće utvrditi efikasnost napisanoga programa.

Mnoga istraživanja u svijetu i kod nas pokazuju da učenici sve više čitaju s ekrana računala zahvaljujući Internetu kao globalnoj mreži računala. Kao jedno od mogućih rješenja u sustavu e-learninga sve češće se spominju *elektronički udžbenici* za koje mnogi izdavači i institucije smatraju da im je naklonjena budućnost.

Postoji više razloga za ovakve procjene: elektronički udžbenici su mnogo jeftiniji, do njih je mnogo lakše doći (nema potrebe za odlaskom u knjižnicu ili knjižaru), ne postoji problem s posuđenim ili rasprodanim izdanjima. Jedna od značajnijih prednosti elektroničkih udžbenika u odnosu na konvencionalne je i činjenica da se tekst može obogatiti audio i video materijalima, kao i interaktivnim sadržajima. To, u mnogome, poboljšava komunikaciju učenika sa sadržajima koje čita.

Najčešće korišteni formati zapisa elektroničkih udžbenika su PDF, HTML i XML format, koje je moguće čitati pomoću računala i posebno dizajniranih čitača (npr. Adobe Reader). Elektronički format pruža i druge, značajne pogodnosti: udžbenik može pripremiti i sam nastavnik za internu uporabu, kombinirajući dijelove i poglavlja iz različitih udžbenika sa svojim osobnim zabilješkama.

Značajan i u današnje vrijeme u didaktičkoj praksi često korišten pojavni oblik daljinskoga učenja predstavljaju *streaming mediji*, jer učenje čine jedinstvenim i vrlo privlačnim. Sintezom audio i vizualne komunikacije omogućuje prezentaciju obrazovnih sadržaja na dimaničan i do detalja objašnjen način, a učenicima omogućuje praćenje predavanja i vježbi iako su dislocirani.

Naime, nove tehnologije za prijenos zvuka i slike u relnom vremenu omogućile su ostvarenje dugo prisutne ideje o stvaranju **virtualne učionice**[1]**,** koja predstavlja najsuvremeniji model učenja na daljinu. Uporabom videokonferencijskih sustava moguće je ostvariti gotovo identičan način obrazovanja kao u klasičnoj učionici, jer se radi o modelu u kome se komunikacija odvija potpuno sinkrono, odnosno u stvarnom vremenu (sl. 2.).

Slika 2. Sala za videokonferencije[2]

Streaming mediji daju jednu novu kvalitetu daljinskom učenju ali su, zbog prilično skupih tehnoloških rješenja i veoma sofisticiranih tehnologija, daleko manje zastupljeni u odnosu na druge pojavne oblike u obrazovanju učenika, pa tako i u ekološkom obrazovanju učenika strukovnih škola.

Implementiranjem obrazovnog software-a u daljinsko učenje, ono dobiva sasvim drugu dimenziju – dobiva nov i dinamičan oblik. Kvalitetan obrazovni software učenicima omogućuje interakciju s nastavnim sadržajima, što drugim riječima znači da u radu s takvim softwareom imaju potpunu kontrolu i potrebne povratne informacije koje znatno utječu na kvalitetu učenja, pa je upravo iz tih razloga vrlo pogodan za ekološko obrazovanje učenika strukovnih škola.

[1] U didaktici se pojam "virtualna učionica" (lat. virtus – snaga, moć) koristi kao sinonim za učionicu u kojoj je, uporabom novih tehnologija i videokonferencijskih sustava, moguće ostvariti prijenos zvuka i slike na daljinu u stvarnom (realnom) vremenu.

[2] Izvor: www.carnet.hr/**videokonferencije**/primjeri

Osim računala i Interneta, informatičku podršku u ekološkom obrazovanju učenika strukovnih škola čine i druge informacijsko-komunikacijske tehnologije, kao što su LCD projektor, pisač, scaner, web-kamera, digitalni foto-aparat, flekso-kamera, mikrofon, slušalice, itd. Kvalitetan model informatičke podrške ekološkom obrazovanju učenika strukovnih škola je onaj u kojem nastavnik(ci) sasvim precizno definira (ju) kako, kada i u kojem opsegu koristiti računala, obrazovni software, Internet i ostale informacijsko-komunikacijske tehnologije.

O tome koju će informatičku podršku koristiti odlučuju nastavnici na razini stručnoga aktiva dogovorno. Naravno da se ta podrška, od slučaja do slučaja, razlikuje, jer se i nastavni sadržaji koji se obrađuju međusobno razlikuju. Vrlo je bitno pronaći optimalan omjer između *klasičnoga predavanja* koje je svojstveno tradicionalnoj nastavi *i informatičke podrške*, što zapravo znači da treba kombinirati klasično predavanje s adekvatnom informatičkom podrškom.

U tablici 1. prikazan je pregled metoda distribucije nastavnih sadržaja kod raznih vrsta učenja na daljinu.

Tablica. 1. Pregled metoda distribucije nastavnih sadržaja kod raznih vrsta učenja na daljinu

Metoda distribucije	Osobine	Potrebe
E-mail	Brži način distribucije nastavnih sadržaja Ograničena samo na učenike koji imaju računalo i pristup Internetu Omogućuje slanje teksta i grafike Osigurava učestaliju komunikaciju između učenika i nastavnika.	Pristup računalu i Internetu
Internet stranice (WWW)	Korištenje velikog resursa podataka Podacima mogu pristupiti pojedinci ili skupina istovremeno Uključuje mogućnost korištenja linkova (veza) aktivne stranice s nekom drugom stranicom	Pristup računalu i Internetu
	Baza podataka se vrlo često obnavlja (ažurira) Učenicima omogućuje ograničen pristup podacima (sigurnost) Ograničena istovremena interakcija	Pristup računaru i Internetu
Telekonferencija	Omogućuje jednosmjernu ili dvosmjernu audio ili audio-video komunikaciju Mogućnost uključivanja pojedinca ili skupine istovremeno Vremenska i prostorna ograničenost Potpuni doživljaj nastavnih sadržaja	Tehničke pretpostavke za održavanje telekonferencije

Kako u domaćoj, tako i stranoj literaturi nije u dovoljnoj mjeri prezentirana mogućnost informatičke podrške ekološkom obrazovanju učenika, posebno na razini strukovnih škola, upravo je to motiviralo autora ovoga rada da provede istraživanje kako bi utvrdio u kojoj mjeri informatička podrška ekološkom obrazovanju učenika strukovnih škola povećava ukupnu efikasnost nastavnog procesa.

I. METODE, TEHNIKE I INSTRUMENTI ISTRAŽIVANJA

U radu su korištene metoda analize i sinteze, metoda kompilacije i metoda komparacije.

Korištene su osnovne istraživačke tehnike:

- analiza literature,
- testiranje i anketiranje učenika i
- obrada podataka.

Radi reprezentativnosti uzorka, istraživanje je vršeno slučajnim izborom trideset učenika iz deset razrednih odjela različitih zanimanja u Strukovnoj školi Vukovar. Od metoda za statističku obradu podataka korišteni su: IDA analiza (Initial Data Analysis), EDA analiza i sinteza.

II. CILJ ISTRAŽIVANJA

Osnovni je cilj provedenoga istraživanja da se, korištenjem suvremenih informacijskih tehnologija (IT), ukaže na statistčki značajnu mogućnost podizanja sveobuhvatne razine i kvalitete ekološkog obrazovanja učenika strukovnih škola.

Dakako, osim osnovnoga, postoje i sekundarni ciljevi, kao npr.:

- da se dođe do znanstvene spoznaje u kojoj mjeri informatička podrška ekološkom obrazovanju učenika strukovnih škola utječe na kvalitetu i kvantitetu usvojenih znanja
- da se dođe do egzaktnih pokazatelja i spoznaje o vremenskoj ekonomičnosti korištenjem informatičke podrške u ekološkom obrazovanju učenika strukovnih škola u odnosu na tradicionalnu nastavu
- da se ispita u kojoj mjeri informatička podrška ekološkom obrazovanju učenika strukovnih škola utječe na povećanje njihove motivacije u nastavnom procesu
- da se ustanovi u kojoj mjeri model informatičke podrške ekološkom obrazovanju učenika strukovnih škola respektira širok raspon individualnih sposobnosti učenika.

III. TIJEK ISTRAŽIVANJA

Istraživanje smo vršili na uzorku četiri zanimanja u četverogodišnjem i četiri zanimanja u trogodišnjem trajanju. Od zanimanja u četverogodišnjem trajanju istraživanjem su obuhvaćeni: *poljoprivredni tehničari fitofarmaceuti*, *tehničari nutricionisti*, *hotelijersko-turistički tehničari i turističko-hotelijerski komercijalisti*, a od zanimanja u trogodišnjem trajanju: *konobari JMO*[3], *kuhari JMO, frizeri JMO i kozmetičari JMO.*

Prvo smo izvršili analizu nastavnih planova i programa, kako bismo utvrdili da li, osim u nastavi ekologije, postoji još neki od predmeta u kojima se izučavaju ekološki sadržaji. Na temelju rezultata provedenog istraživanja smatramo da je informatičku podršku korištenu u nastavi ekologije, uz određene modifikacije, moguće koristiti i u nastavi tih predmeta. Odvojeno smo analizirali prirodnu i društvenu skupinu predmeta (tablica 2. i 3.).

Iako je daleko manji broj predmeta koji pripadaju društvenoj skupini čiji nastavni planovi i okvirni obrazovni programi sadrže ekološke sadržaje, ne treba ih zanemariti. Naprotiv, model informatičke podrške korišten u provedenom istraživanju, uz manje je modifikacije, moguće koristiti i u nastavi društvene skupine predmeta na način da se prilagodi programskim sadržajima svakog predmeta pojedinačno.

Tablica. 2. Pregled prirodne skupine predmeta u čijim su nastavnim planovima zastupljeni ekološki sadržaji

Red. broj	Nastavni predmet	Program – zanimanje							
		Poljopr. tehn. fitofarmaceut	Tehničar nutricionist	HTT	THK	Konobar JMO	Kuhar JMO	Frizer JMO	Kozmetičar JMO
1.	Biologija	+	+	–	–	–	–	–	–
2.	Biologija s ekologijom	–	–	+	–	–	–	–	–
3.	Biologija s higijenom i ekologijom	–	–	–	+	–	–	–	–
4.	Ekološka proizvodnja hrane	–	+	–	–	–	–	–	–
5.	Zaštita na radu, higijena i sanitacija	–	+	–	–	–	–	–	–

[3] JMO je skraćenica od *jedinstveni model obrazovanja*, u kojem učenici svih trogodišnjih programa (zanimanja) imaju iste opšteobrazovne predmete, zastupljene sa istim brojem nastavnih sati

Red. broj	Nastavni predmet	Program – zanimanje							
		Poljopr. tehn. fitofarmaceut	Tehničar nutricionist	HTT	THK	Konobar JMO	Kuhar JMO	Frizer JMO	Kozmetičar JMO
6.	Zaštita bilja	+	–	–	–	–	–	–	–
7.	Biologija i ekologija	–	–	–	–	+	+	–	–
8.	Praktična nastava	+	+	+	–	–	–	–	–

Tablica. 3. Pregled društvene skupine predmeta u čijim su nastavnim planovima zastupljeni ekološki sadržaji

Red. broj	Nastavni predmet	Program – zanimanje							
		Poljopr. tehn. fitofarmaceut	Tehničar nutricionist	HTT	THK	Konobar JMO	Kuhar JMO	Frizer JMO	Kozmetičar JMO
1.	Etika	+	+	+	+	+	+	+	+
2.	Geografija	+	+	+	–	–	–	–	–
3.	Turistička geografija	–	–	–	+	–	–	–	–
4.	Praktična nastava	+	+	+	–	–	–	–	–

Istraživanje se odvijalo u dva pravca. Jedan je dio ispitanika odslušao predavanje iz jedne nastavne cjeline (ONEČIŠĆENJE I ZAŠTITA OKOLIŠA) u klasičnoj učionici, gdje se nastava održavala na tradicionalan način. Drugi je dio ispitanika istu nastavnu cjelinu obradio samostalno, u multimedijalnom okruženju (informatička učionica) koristeći aplikaciju autora rada, koja u naravi predstavlja multimedijalni interaktivni udžbenik ekologije (sl. 3.) s mogućnošću pristupa Internetu i korištenja navedenih IT.

Za izradu modela multimedijalnog interaktivnog udžbenika ekologije korišten je alat Microsoft Visual Studio 2008 i ASP.NET platforma Framework 3.5. Aplikacija je razvijena kao web stranica, kako bismo je učinili dostupnom svim zainteresiranim korisnicima.

Bitna je osobina multimedijalnog udžbenika njegova interaktivnost. Zbog toga je u udžbenik "ugrađen" modul za samostalno testiranje učenika, s mogućnošću samokontrole dostignute razine znanja, pregleda točnih i netočnih odgovora i ponovnog samotestiranja.

Slika 3. Slika osnovnog zaslona

Nakon toga, svi su sudionici u istraživanju podvrgnuti testiranju, s ciljem da ustanovimo da li je informatički podržana nastava ekologije efikasnija od tradicionalnog načina učenja. Koristili smo test s nizom zadataka objektivnog tipa, a testiranje je trajalo zadnjih petnaest minuta nastavnoga sata.

Nakon obrade odgovora ispitanika, ustanovili smo da se obrađeni podaci bitno razlikuju, jer su učenici korištenjem informatičke podrške postigli daleko bolje rezultate u svladavanju novoga gradiva u odnosu na tradicionalan način učenja.

Kako bismo dokazali da uporaba suvremenih IT povećava ukupnu efikasnost nastavnog procesa, u sljedećoj tablici dajemo sintetički prikaz samo točnih odgovora učenika dobivenih tradicionalnim i informatički podržanim načinom učenja ekoloških sadržaja.

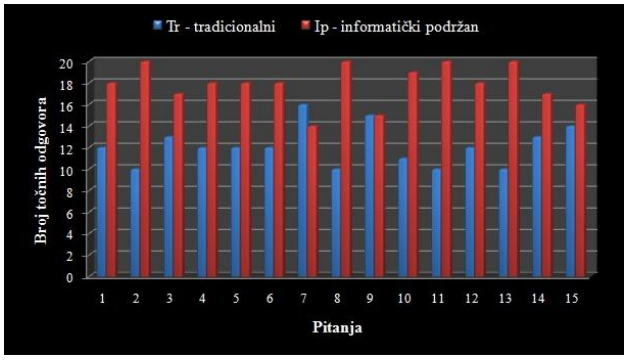
Tablica. 3. Sintetički prikaz točnih odgovora učenika

Pitanje	Tr	Ip	Tr [%]	Ip [%]
1.	12	18	40,00	60,00
2.	10	20	33,33	66,67
3.	13	17	43,33	56,67
4.	12	18	40,00	60,00
5.	12	18	40,00	60,00
6.	12	18	40,00	60,00
7.	16	14	53,33	46,67
8.	10	20	33,33	66,67

Pitanje	Tr	Ip	Tr [%]	Ip [%]
9.	15	15	50,00	50,00
10.	11	19	36,67	63,33
11.	10	20	33,33	66,67
12.	12	18	40,00	60,00
13.	10	20	33,33	66,67
14.	13	17	43,33	56,67
15.	14	16	46,67	53,33

Komparativna analiza podataka pokazuje da je uporabom IT učenje daleko efikasnije od tradicionalnog. Od ukupno 30 ispitanika na kojima je provedeno istraživanje, tradicionalnim je načinom učenja 11 učenika (40,44%) odgovorilo točno na postavljena pitanja.

U informatički podržanom načinu učenja taj je broj daleko veći. Na postavljena je pitanja odgovorilo 19 učenika (59,56%), što nas upućuje na zaključak da je *informatički podržan način učenja efikasniji od tradicionalnog načina učenja*, tj. da korištene IT u provedenom istraživanju povećavaju ukupnu efikasnost nastavnog procesa (sl. 4.).



Slika 4. Grafički prikaz točnih odgovora ispitanika tradicionalnim i informatički podržanim načinom učenja

IV. ZAKLJUČAK

Bitan utjecaj na razvoj odgoja i obrazovanja kroz povijest čovječanstva imaju tehnološka otkrića koja je moguće primijeniti u nastavnom procesu i tako ga učiniti efikasnijim. Početak XXI. stoljeća obilježava primjena suvremenih IT, pri čemu tradicionalna nastava postupno gubi svoje prvobitno značenje.

Osnovni problem koji je bio predmet provedenog istraživanja je ustanoviti kakva je uloga IT u ekološkom obrazovanju učenika strukovnih škola, tj. dokazati da uporaba IT u nastavi povećava ukupnu efikasnost nastavnog procesa.

U okviru istraživanja izvršeno je testiranje i mjerenje relevantnih parametara od utjecaja na efikasnost ovakvoga načina učenja. Mjerenje relevantnih parametara i analiza dobivenih rezultata je obavljeno pomoću standardnih statističkih metoda.

Najznačajniji znanstveni doprinos provedenoga istraživanja je formalna specifikacija skupa obrazovnih i metoda informatičkih i komunikacijskih tehnologija za realizaciju podrške tradicionalnog obrazovanja, kao i dokaz boljih kvantitativnih i kvalitativnih rezultata u svladavanju nastavnog gradiva u informatički podržanoj nastavi ekologije i drugih predmeta u kojima se izučavaju ekološki sadržaji.

Pojavom Interneta kao globalne mreže računala, on postaje medij u globalnoj komunikaciji i gotovo bezgranični resurs nastavnih sadržaja iz svih predmeta, pa tako i ekologije. Znanstvenom spoznajom o kvaliteti i kvantiteti razine primjene tradicionalnih i IT u svrhe ekološkog obrazovanja učenika strukovnih škola postignuti su osnovni i sekundarni ciljevi istraživanja.

U eksperimentalnom je dijelu provedenoga istraživanja korišten model multimedijalnog interaktivnog udžbenika ekologije, kojega je razvio autor rada, kombinirajući dio udžbenika Biologije za učenike prvih i drugih razreda [10] u kojem se obrađuju ekološki sadržaji s osobnim zabilješkama i zabilješkama nastavnika biologije, te osmišljenim eksperimentima izvedenim u školskom laboratoriju i prirodi i video i zvučnim datotekama.

V. LITERATURA

[1] S. Alessi and S. Trollip, "Multimedia for Learning", Massachusetts: Allyn and Bacon, Pearson Publishing Company, 2001.

[2] R. C. Clark and R. Mayer, "E-Learning and science of Instruction", San Francisco: Pfeiffer, 2001.

[3] R. Marinković, "Inteligentni sustavi za poučavanje", Zagreb: Hrvatska zajednica tehničke kulture, 2004.

[4] V. Matić, "Model informatičke podrške ekološkom obrazovanju učenika strukovnih škola", doktorska disertacija, Zrenjanin: Tehnički fakultet "Mihajlo Pupin", 2011.

[5] V. Matić, P. Hotomski i K. Voskresenski, "Personal computer in use of ecological education pupils vocational schools", Technics Technologies Education Management, vol. 6 (4), pp. 1188–1198, Sarajevo: DRUNPP, 2011.

[6] V. Matić, "Environmental Education of Students in Vocational Schools in a Multimedia Surrounding", IEEE 35th international convention, ISSN: 1847-3946, pp. 1527-1532, Opatija: May 21-25, 2012.

[7] M. Matijević i L. Bognar, "Didaktika", Zagreb: Školska knjiga, 1974.

[8] S. Rodek, "Kompjutor i suvremena nastavna tehnologija", Zagreb: Školska knjiga, 1986.

[9] E. Tobolka, "Model računarski podržane nastave engleskog jezika i metode njene realizacije kao faktor unapređenja i informatizacije obrazovanja", Zrenjanin: Tehnički fakultet "Mihajlo Pupin", 2002.

[10] Lj. Vrček, "Biologija, udžbenik za strukovne škole, svezak C", Zagreb: Profil, 2007.

Primjena e-učenja u NATO okružju (e-NATO)

Darko Možnik
Ministarstvo obrane Republike Hrvatske
Hrvatsko vojno učilište „Petar Zrinski“
Ilica 256B, Zagreb, Republika Hrvatska
Telefon: 01-3784 613 Fax: 01-3784 747 E-mail: dmoznik@xnet.hr

Sažetak - Rad opisuje i objašnjava e-NATO odnosno koncept e-učenja u NATO okružju. Navode se motivi, prednosti, razlozi i mogućnosti obrazovanja i obuke primjenom e-učenja. Elementi i modeli e-učenja (učenje usredotočeno na studenta, SCORM, sustavi za upravljanje učenjem tj. LMS-ovi MOODLE i ILIAS, razine interaktivnosti, „dubine“ znanja, razvojni timovi, resursi i modeli: SAT/ADDIE i SECTIONS), standardi i osiguranje kvalitete. Strategije implementacije. Tehnologije primjene e-učenja (JADL - združeno napredno distribuirano učenje i ADL - napredno distribuirano učenje, CBT - Computer Based Training, immersive learning tj. virtualno učenje, m-learning tj. mobilno učenje, blending learning tj. hibridno učenje i kolaborativno učenje i alati (chat soba, forumi, wiki, webinar - web based seminar i webcast).

Primjena e-obrazovanja na Hrvatskom vojnom učilištu tj. HVU „Petar Zrinski“ (vizija, misija, ciljevi), podizanje kvalitete vojnog obrazovanja, uloga nastavnika i studenata u obrazovnom procesu, povećanje konkurentnosti i dostupnosti učilišta i učilišnih studijskih programa, osposobljavanje studenata za uporabu tehnologija cjeloživotnog obrazovanja i učenja.

Pogled u budućnost - transformacija HVU-a u visoko obrazovnu i znanstveno-istraživačku vojnu ustanovu odnosno instituciju tj. sastavnicu sveučilišne zajednice.

Sadašnje i buduće sposobnosti e-učenja. Zajednica obrazovanja i obuke primjenom e-učenja. Zaključna razmatranja.

I. UVOD

E-učenje i *učenje na daljinu* primjenjuju se u obrambenim i vojnim sustavima zemalja članica NATO saveza kako u pojedinim NATO institucijama tako i kao komplementarni oblici vojnog obrazovanja. U obrambenim i vojnim sustavima zemalja članica NATO saveza *e-učenje* se koristi i primjenjuje pod nazivima: JADL (engl. *Joint Advanced Distributed Learning*) i ADL (engl. *Advanced Distributed Learning*).

O*pen source* sustavi za upravljanje učenjem tj. LMS-ovi (engl. *Learning Management Systems*) **MOODLE** (engl. *Modular Object-Oriented Dynamic Learning Environment*), najviše koriste akademske zajednice i **ILIAS** (njem. *Integriertes Lern-, Informations- und Arbeitsko-operations System*) koristi se u obrambenim i vojnim sustavima zemalja članica NATO saveza i Partnerstva za mir (engl. *Partnership for Peace*).

Od 2009. godine Republika Hrvatska je članica NATO saveza.

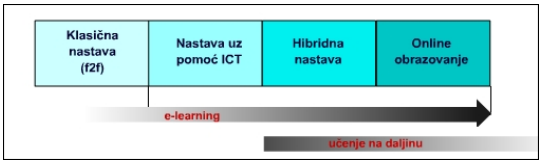
U MORH-u i OS RH u sustavu vojnog obrazovanja za *e-učenje* se koriste LMS sustavi MOODLE i ILIAS kao platforme i skupovi alata. Također, provodi se projekt “*Uvođenje novih načina usvajanja i održavanja znanja*” odnosno *e-obrazovanje u OS RH i MORH-u.* Određeni pripadnici OS RH i djelatnici MORH-a, a poglavito nastavnici Hrvatskog vojnog učilišta “Petar Zrinski” osposobljeni su vlastitim resursima, učinkovitim korištenjem ILIAS-a, razvijati i primjenjivati *online* ADL tečajeve i module za obrambene i vojne potrebe.

II. E-UČENJE I UČENJE NA DALJINU

A. E-učenje

E-učenje je *vrsta komunikacijskog kanala*, kanala kroz koji se odvija učenje.

Postoje dvije skupine definicija *e-učenja* i to: 1. skupina "tehničkih" definicija stavlja naglasak na tehnologiju (na "*e*" dio u imenu). Primjer jedne od "tehničkih" definicija glasi: *"E-obrazovanje* je bilo koji oblik učenja, poučavanja ili obrazovanja koji je potpomognut uporabom računalnih tehnologija, a posebno računalnih mreža temeljenih na Internet tehnologijama.*"* i 2. skupina "pedagoških" definicija ima naglasak na obrazovanju, tj. na učenju i poučavanju ("*learning*" dio u imenu). Jedna od takvih definicija je: *"E-learning* je interaktivan ili dvosmjeran proces između nastavnika i učenika uz pomoć elektroničkih medija, pri čemu je naglasak na procesu učenja, dok su mediji samo pomoćno sredstvo koje upotpunjuje taj proces [1].*“ E-učenje* kao oblik obrazovanja postoji na više razina: kao potpuno samostalan oblik, ali i kao sastavni dio ili nadopuna klasičnog obrazovanja. Proširena klasifikacija ili "vremenska crta" *e-obrazovanja* prikazuje obrazovanje kao kontinuum na čijem je lijevom kraju klasična ili tradicionalna nastava (*f2f – face-to-face* predavanja). Pomak prema *e-learningu* započinje uvođenjem ICT pomagala u f2f nastavu. U središnjem dijelu "vremenske crte" koji se odnosi na *e-learning* je *hibridna* nastava ili *kombinirani* pristup učenju. *Online obrazovanje* kao samostalan oblik nastave smješteno je na desnom kraju ovog kontinuuma. Idući slijeva udesno po "vremenskoj crti" *e-obrazovanja,* u obrazovnim kategorijama se kori-ste sve više i više tehnologije.



Slika 1. Prikaz “kontinuum”a *e-učenja*

Na slici 2. prikazan je odnos *e-učenja*, *učenja na daljinu i online obrazovanja*. Iako se *e-learning* i *učenje na daljinu* često izjednačuju, nije riječ o istim oblicima obrazovanja: postoje vrste *e-obrazovanja* koje se ne odvijaju online, a isto tako postoje i oblici *učenja na daljinu* koji ne koriste

ICT. Među glavnim prednostima *e-učenja* su: a) vremenska i prostorna fleksibilnost; b) interakcija (komunikacija) između polaznika i nastavnika, koja se odvija putem računala (na primjer: e-mail, forumi) te je često neposrednija i intenzivnija nego komunikacija u razredu; c) komunikacija i grupni rad na zajedničkim projektima između polaznika međusobno; d) programi *e-obrazovanja* su dostupni putem Interneta nevezano za fizičku lokaciju na kojoj se polaznici nalaze; e) programe *e-obrazovanja* mogu putem Interneta pohađati polaznici koji zbog svog posla ili drugih obveza ne mogu nazočiti tradicionalnoj nastavi u vrijeme u koje se ona odvija; f) materijali pripremljeni za *e-obrazovanje* relativno se lako obnavljaju, prilagođavaju i ponovo koriste za različite grupe polaznika; g) korištenje *interaktivnih sadržaja za učenje* i različitih *medija* za prezentiranje sadržaja te *dostupnost* sadržaja 24 sata *online*.

Slika 2. Odnos *e-učenja* i *učenja na daljinu*

B. Učenje na daljinu

Učenje na daljinu (engl. *distance learning, distance education*) je *obrazovanje ili obuka koja se nudi učenicima na različitim mjestima odnosno fizički udaljenima od učitelja ili izvora informacija*. U praksi je učenje na daljinu puno složenije od ove definicije jer uključuje koriš-tenje novih tehnologija i novih interaktivnih nastavnih metoda. Termin *učenje na daljinu* u novije vrijeme se zamjenjuje terminom *online učenje*, čime se naglašava da je riječ o posebnom obliku *e-learninga* ili *e-obrazovanja*.

Učenje na daljinu tj. ADL koristi većina zemalja članica NATO-a i PfP-a. Iskustva pokazuju da je i to jedan od oblika učenja, koji ima višestruke prednosti.

Department of Defense (DoD) je uvidio da sa konstantnim brzim i neplaniranim promjenama lokacija vojnih snaga ne može omogućiti konstantnu potrebu za edukacijom u novim tehnologijama. Pošto je vremenski interval razmještaja postrojbi kao i vrijeme implementiranja novih tehnologija kraće sve više pojavljuje se potreba za edukacijom *"right time, right place"*, gdje je vrijeme često ***sada***, a mjesto ***bilo gdje*** na planeti. Dolazi se do zaključka da u informacijskom dobu vojska mora biti spremna odgovoriti na nove prijetnje ***odmah i sada*** sa najnovijim tehnologijama. Kao odgovor na takve izazove razvijalo se i dalje se razvija *e-učenje* odnosno JADL i ADL čija je svrha postavljanje okvira koji će omogućiti *DoD* službenicima (od ratnika - do tajnika) pristup: a) kvalitetnoj edukaciji i treningu prema individualnoj potrebi, b) relativno mali troškovi i c) kada god i gdje god to bude potrebno.

JADL i ADL su definirani kao lideri koji će:

a) iskoristiti maksimalne potencijale (računala, informacijske tehnologije, komunikacijske tehnologije);
b) ne odstupati od standardne programske potpore dostupne širokoj masi računala;
c) omogućiti edukaciju skrojenu prema individualnim potrebama svakog pojedinca dostupnu po principu *anytime, anywhere*;
d) omogućiti *computer-managed instruction* okružje i potporu za: razvojne inženjere, administratore, voditelje (engl. *managers)* i profesore.

Glavne *razlike* između programa za *učenje na daljinu* jesu u tehnologiji koja se koristi, u strukturi programa ili tečaja, te u stupnju nadgledanja polaznika. Kao jedna od najvećih prednosti *učenja na daljinu* smatra se to što omogućuje **stalno učenje** (engl. *lifelong learning*) i **profesionalno usavršavanje**. Kao prednosti *učenja na daljinu* mogu izdvojiti: a) *vlastiti tempo*, b) *odabiranje mjesta učenja, c) dostupnost tema koje ne nude tečaje-vi/programi u tom području*, *d) sudjelovanje u najkvalitet-nijim ili najprestižnijim programima,* e) *odabiranje svog načina učenja* i f) *praktičan rad s različitim tehnologija-ma* omogućuje *samostalno učenje*.

III. KONCEPT E-UČENJA

Danas je sveprisutna potreba ali i želja da se unaprijedi kvaliteta i dostupnost različitih obrazovnih programa. Pri tome je obrazovanje postalo cjeloživotnim procesom i potrebom, pa se briga o dostupnosti i kvaliteti odnosi na sveukupni ciklus cjeloživotnog obrazovanja. Postalo je jasno da se navedeno ne može ostvariti bez odgovarajuće i učinkovite uporabe informacijskih i komunikacijskih tehnologija u obrazovnom procesu. Upravo kvalitetna implementacija tehnologija e-učenja donosi niz prednosti u obrazovni proces i omogućava željeno novo, moderno i kvalitetno obrazovanje.

E-učenje je proces obrazovanja (proces učenja i poučavanja) uz uporabu informacijske i komunikacijske tehnologije, koja pri tome doprinosi unaprjeđenju kvalitete toga procesa i kvalitete ishoda obrazovanja.

E-učenje omogućava: učenje prilagođeno individualnim potrebama studenta (bilo kada, bilo gdje), aktivno sudjelovanje i aktivnu ulogu studenata, aktualnije i ažurnije obrazovne materijale, globalni pristup različitim obrazovnim sadržajima, smanjenje „pritiska“ na fizičke resurse (omogućavajući virtualne učionice, virtualne laboratorije), više komunikacije nastavnik-student i student-student, timski rad, projektni rad nastavnika i studenta, bolju iskorištenost vremena (za kvalitetnije sadržaje) pri kontaktima nastavnik-student, samostalnu provjeru znanja i kontinuirano mjerenje i praćenje napretka studenta [5].

Za zemlje članice NATO saveza, a RH je od 25.10. 2012., potpisivanjem Deklaracije o završenoj integraciji, postala punopravnom članicom NATO saveza, donesen je Koncept e-učenja (8. veljače 2012. godine).

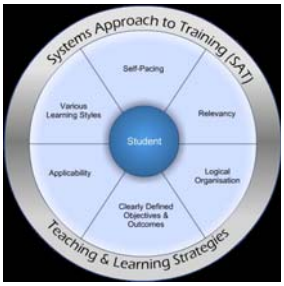
Slika 3. Oznaka za e-NATO

NATO e-učenje nudi povećan i raznovrstan pristup

obrazovanju, nudi mogućnost obuke kroz *on-demand* dostupnost, uštedu u troškovima, korištenje *Self Paced* tečajeve, konzistentnu nastavu i mogućnost za kolaboracijski pristup (suradnju kroz *forum*-e i *chat*-ove). Korištenje e-učenja omogućava NATO-u, svim članicama i PfP partnerskom osoblju pristup visoko kvalitetnom obrazovanju i obuci koja se može prilagoditi individualnim potrebama te omogućiti osoblju da učinkovito doprinosi NATO misijama.

Danas se e-učenje u NATO-u primjenjuje na različite načine i nastoji biti integrirano u sve tijekove obrazovanja i obuke. NATO trenutno nudi preko 300 sati kvalitetnog, relevantnog obrazovanja i obuke u raznim temama i područjima. Do sada je kroz ovakav način obrazovanja obučeno više od 33.000 pripadnika-studenata.

To je učenje i poučavanje koje je usmjereno na pripadnike odnosno studente [5].

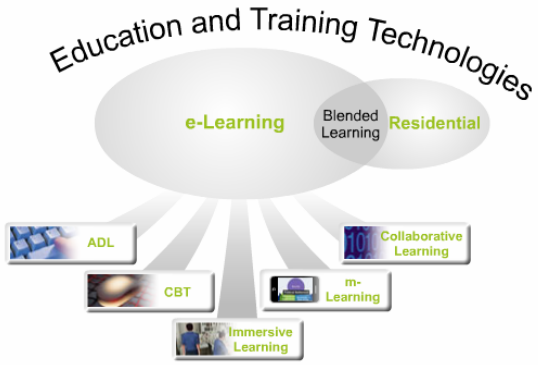


Slika 4. Model učenja usmjeren na studenta

IV. TEHNOLOGIJE E-UČENJA

Postoji više tehnologija obrazovanja i obuke primjenom e-učenja prema NATO-voj podjeli i to:

1. združeno napredno distribuirano učenje - JADL (engl. *Joint Advanced Distributed Learning*) i napredno distribuirano učenje - ADL (engl. *Advanced Distributed Learning*),
2. učenje pomoću ICT-a odnosno CBT (engl. *Computer-Based Training*),
3. „uronjeno“ učenje (engl. *Immersive learning*),
4. mobilno učenje (engl. *M-learning*),
5. suradničko ili kolaborativno učenje (engl. *Collaborative learning*) i
6. kombinirano ili hibridno učenje (engl. *Blended learning*) [5].



Slika 5. Tehnologije e-učenja prema NATO-u

4.1. JADL i ADL

a) JADL

Slika 6. Oznaka za JADL

JADL (engl. *Joint Advanced Distributed Learning*) je združeno napredno distribuirano učenje na daljinu koje razvija NATO-ovo Savezničko zapovjedništvo za transformacije tj. ACT (engl. *Allied Command Transformation*), konkretno dio za JET (engl. *Joint Education and Training*) za potrebe zemalja članica NATO saveza i Partnerstva za mir tj. PfP-a (engl. *Partnerships for Peace*).

Slika 7. JADL i ADL nalaze se na adresi: https://jadl.act.nato.int

b) ADL

Slika 8. Oznaka za ADL

ADL (engl. *Advanced Distributed Learning*) proizvod je *ADL inicijative* uspostavljene 1997. godine u Ministarstvu obrane SAD-a za standardizaciju i modernizaciju upravljanja „treninzima“ i „edukacijama“ te načinima poučavanja. Vizija navedene inicijative je osigurati pristup učenju najviše kvalitete koja može biti „skrojena“ tj. prilagođena pojedinačnim potrebama i isplativa u pravo vrijeme i na pravom mjestu. ADL koristi strukturne i suradničke (engl. *collaboration*) metode iz industrije, akademskih zajednica i vlada koji razvijaju standarde, alate i sadržaje učenja.

Zemlje članice NATO saveza mogu koristiti izrađene obrazovne on-line tečajeve za različita područja u obrambenim i vojnim sustavima.

4.2. CBT (engl. *Computer Based Training*)

CBT je obrazovni ili obučni tečaj dostavljen direktno na studentsko računalo putem CD-ROM-a ili nekog drugog uređaja za pohranu. On se razlikuje od ADL-a po tome što se ne isporučuje preko weba ili putem sustava za upravljanje učenjem i ne zahtijeva usklađenost sa SCORM (engl. *Sharable Content Object Reference Model*) standardom. Studenti se ne prijavljuju na sustav (engl. *login*) CBT, a praćenje napretka polaznika može biti praćeno lokalno ili po završetku tečaja.

Slika 9. Oznaka ili logo za CBT

4.3. „URONJENO“ UČENJE (engl. Immersive learning)

„Uronjeno“ (engl. *Immersive*) učenje koristi računalno simulirano okruženje koje se događa u stvarnom životu ili hipotetsku situaciju u grafički bogatom i dinamičnom okruženju. Studenti su stavljeni u situaciju i uključeni su u proces učenja kroz interaktivne simulacije i igre temeljene na *game-based* aplikacijama. *Immersive* učenje podupire jednog učenika ili više malih timova da rade zajedno kako bi riješili problem, uvježbali tehnike ili poboljšali svoje vještine. Kroz korištenje odobrenih ciljeva i skriptiranje, studentske aktivnosti i odgovori na određene situacije mogu se pratiti i testirati kako bi bili sigurni da su ciljevi dosegnuti i ispunjeni. Ova vrsta učenja može biti web orijentirana, preuzeta na samostalna računala ili mobilne uređaje.

Ozbiljne igre i virtualni svjetovi, oba primjera *immersive* učenja, su dizajnirana da uključe studente u širok spektar obrazovnih aktivnosti.

Ovi tečajevi imaju za cilj poboljšati sposobnost NATO-a koristeći provjerene metode i prakse u učenju kao i tehnologije koje podržavaju simulaciju okoline u kojima se polaznici mogu naći.

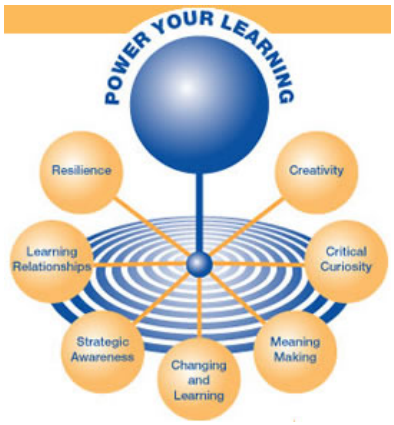
4.4. MOBILNO UČENJE (engl. *m-learning*)

Slika 10. Mobilno učenje (engl. *m-learning*)

Mobilno učenje (engl. *m-learning*) je uporaba prijenosnih računalnih uređaja, kao što su različiti iPad-i i drugi tableti, prijenosna računala, osobni digitalni pomoć-nici (PDA) i pametni telefoni povezani s bežičnim mreža-ma. To omogućuje da se podučavanje i učenje proširi i na prostore izvan tradicionalne učionice te daje nastavnicima i polaznicima povećanu fleksibilnost i nove mogućnosti za interakcije koje su prikladne i dostupne virtualno s bilo kojeg mjesta.

Jednostavnost m-učenja, uz tipičan studentski pristup podacima na zahtjev, na odabranom uređaju otvara nove mogućnosti. Studenti imaju pogodnost i fleksibilnost pri korištenju mobilnog uređaja kako bi pomogli u procesu učenja. S GPS mogućnostima telefona i drugih mobilnih uređaja, tečajevi mogu biti dinamični dok se provode u terenskim uvjetima u prirodi i mogu pomoći studentima prilikom uvježbavanja. Odmah, na zahtjev, mogu osigurati pristup bilo na radnom mjestu ili na terenu. Tradicionalni e-learning tečajevi mogu se prilagoditi manjim veličinama zaslona. Dizajniraju se namjenski za mobilne uređaje.

4.5. SURADNIČKO ILI KOLABORATIVNO UČENJE



Slika 11. Suradničko ili kolaborativno učenje

Suradničko ili kolaborativno učenje (engl. *Collaborative learning*) podupire nastavu kroz interaktivne metode ili online komunikacije koristeći društvene medijske alate kao što su chat sobe, tribine i video alate kao što su webi-nari i web-emitiranja. Ovi alati se mogu implementirani u učionici i online okruženju kako bi podržali tijek nastave. Oni daju mogućnost velikom broju polaznika da rade zajedno s nastavnikom-instruktorom. Također dopuštaju učenje *student-to-student*, na primjer, student može podijeliti svoj uvid o tome kako riješiti složeni problem. Ova aktivna razmjena ideja i rasprava promiče kritičko mišljenje i rješavanje problema. Kolaborativno učenje može biti sinkrono, kada se učenje događa u stvarnom vremenu ili asinkrono učenje, kada se ne odvija u stvarnom vremenu.
Kolaborativni alati za učenje uključuju: chat sobu, forume, wiki, webinar-e/webcast-ove.

4.6. KOMBINIRANO ILI HIBRIDNO UČENJE (engl. *blended learning*)

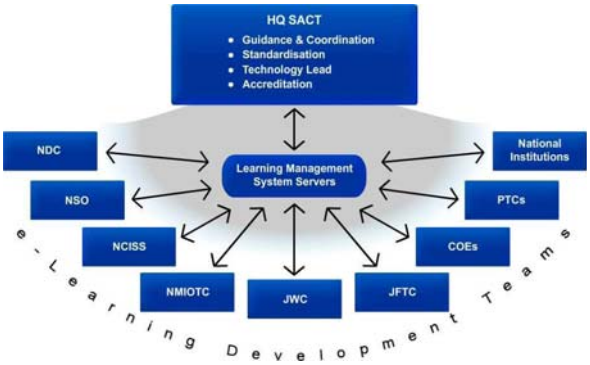
Slika 12. Kombinirano ili hibridno učenje

Ponekad je e-learning prisutan u klasičnom obrazovanju u školama i na fakultetima, tako da se računalo i Internet koriste za vrijeme nastave ili u izradi domaćih zadaća. U drugim slu-ča-

jevima se obrazovni programi djelomično izvode online, tj. uz pomoć Interneta ili uz pomoć materijala distribuiranih na CD ROM-u, a djelomično na tradicionalan način. Taj se oblik zove «blended learning», «hybrid» ili «mixed mode», a često se primjenjuje u izvođenju stručnih seminara i ostalih obrazovnih programa za koje bi bilo preskupo ili nemoguće okupiti polaznike na dulje vrijeme. Na kraju, postoje obrazovni programi koji se u potpunosti izvode online, što znači da se od upisa do diplome, sva komunikacija između polaznika i organizatora obrazovanja odvija putem Interneta.

Kada se neka ustanova odluči svoje obrazovne programe ponuditi putem e-učenja, treba pretpostaviti da pritom ima kao cilj poboljšati kakvoću obrazovanja ili omogućiti pohađanje onim polaznicima kojima bi to inače bilo nemoguće ili omogućiti pristup izvorima i stručnjacima koje je nemoguće imati u neposrednoj blizini.

Ovakve načine e-učenja koriste neke od institucija odnosno čine zajednicu obrazovanja i obuke primjenom e-učenja u NATO-u.



Slika 13. Zajednica obrazovanja i obuke primjenom e-učenja u NATO-u

U zemljama članicama NATO saveza koriste se LMS-ovi (engl. *Learning Management Systems*) odnosno sustavi za upravljanje učenjem MOODLE i ILIAS.

Slika 14. LMS sustav *ILIAS*

Slika 15. LMS sustav *MOODLE*

IV. MODELI E-UČENJA

Postoje dva osnovna modela učenja i to: SAT i SECTIONS. SAT (engl. *Systems Approach to Training*) koji uključuje ADDIE (engl. *Analysis, Design, Development, Implementation and Evaluation*) faze. SAT je proces holističkog i sveobuhvatnog obrazovanja i obuke koji se fokusira na zahtjeve i potrebe organizacije.

Svi NATO-vi programi obrazovanja i obuke, bez obzira da li su vođeni instruktorom ili u obliku tečajeva e-učenja, definirani su, oblikovani, razvijeni, vrednovani i upravljani sukladno SAT modelu učenja. SECTIONS model bi trebao zadovoljiti:

S *Students*: Studente koji dolaze na školovanje

E *Ease of use and reliability*: Potrebno je postići odgovarajuću pouzdanost sustava, a što se najbolje može očitati u kvalitetnom testiranju opreme i sustava

C *Costs*: Nabaviti potrebnu opremu i alate, educirati administratore, autore sadržaja, nastavnike i korisnike

T *Teaching and learning*: Koje vrste učenja zaista trebamo? Potrebno je razviti vlastite e-learning sadržaje; razviti mješoviti način poučavanja (engl. *blended learning*)

I *Interactivity*: Koju vrstu interakcije ova vrsta tehnologije omogućava? Omogućena je f2f interakcija gdje studenti mogu aktivno sudjelovati, postavljati pitanja, davati željenu dinamiku i smjer

O *Organizational issues*: Otkloniti prepreke da bi tehnologija koja bi se koristila polučila uspjeh, izvršiti promjene u organizaciji prije uporabe

N *Novelty*: Učiniti sve da nova tehnologija bude zaista i željena, jedino onda postaje zanimljivija, a samim time i učinkovitija

S *Speed*: Omogućiti da se tečajevi, predmeti mogu implementirati sa tom tehnologijom te da se nastavni materijali mogu brzo mijenjati.

V. POGLED U BUDUĆNOST

Slika 16. Oznaka ili logo NATO saveza

Slika 17. RH je punopravna članica NATO saveza

Slika 18. Hrvatsko vojno učilište „Petar Zrinski“

Slika 19. Primjena elemenata e-obrazovanja na terenu

1. Vizija:

Hrvatsko vojno učilište "Petar Zrinski" uvođenjem e-učenja želi u procesu obrazovanja ostvariti aktivan partnerski odnos nastavnika i polaznika, prilagođen osobinama i potrebama svakog polaznika, s jasno postavljenim obrazovnim ciljevima i jasno dodijeljenim odgovornostima za postizanje tih ciljeva. Želi se postići optimalna kombinacija klasične nastave u učionici i nastave uz pomoć informacijske i komunikacijske tehnologije. Vještine stečene tijekom studija trebaju predstavljati temelj za cjeloživotno obrazovanje i daljnje samostalno učenje.

2. Misija:

Sustavno uvođenje e-učenja doprinosi kvaliteti sadašnjeg i budućeg sveučilišnog obrazovanja kroz stvaranje okruženja u kojem se potiče i omogućava aktivna suradnja nastavnika i polaznika. Uvođenjem e-učenja unaprjeđuje se i znanstveno-istraživački rad i otvara prostor za programe cjeloživotnog učenja, ostvaruje značajnija društvena uloga i povećava konkurentnost učilišta u domaćem i međunarodnom okruženju.

3. Strateški ciljevi:

Uvođenjem i aktivnom primjenom e-učenja Hrvatsko vojno učilište namjerava ostvariti četiri strateška cilja:

- unaprijediti kvalitetu obrazovanja,
- omogućiti nastavnicima i studentima ostvarivanje novih uloga u obrazovnom procesu,
- povećati konkurentnost i dostupnost učilišta i učilišnih programa i
- osposobiti studente za uporabu tehnologija cjeloživotnog učenja.

HVU je u procesu transformacije u visoko obrazovnu i znanstveno-istraživačku vojnu ustanovu odnosno instituciju tj. sastavnicu sveučilišne zajednice.

VI. SADAŠNJE I BUDUĆE SPOSOBNOSTI E-UČENJA

NATO nudi više od 300 sati kvalitetnog i relevantnog obrazovanja i obuke za različita područja u obrambenim i vojnim sustavima i ima više od 33.000 obučenih pripadnika tj. studenata.

Primjeri programa i tečajeva e-učenja koji su na raspolaganju:

- programi i tečajevi preduputne obuke za mirovnu misiju u Afganistanu (engl. *International Security Assistance Force /ISAF/*) koju tijekom priprema uspješno završi svaki pripadnik koji se upućuje u navedenu misiju;

- programi, tečajevi i alati obrazovanja i obuke u području planiranja sveobuhvatnih operacija (engl. *Comprehensive Operations Planning Education and Training programmes and courses and planning tools*);

- programi i tečajevi za područje protupobunjeničke borbe (engl. *COIN – Counter-Insurgency Battle*) i

- programi i tečajevi za obuku protiv improviziranih eksplozivnih naparava (engl. *C-IED training*) i za vizualizaciju računalno poduprtih vježbi (engl. *Visualisation of Computer Assisted Exercises /CAX/*).

Navedeni primjeri programa, tečajeva i alata e-učenja, kao i mnogi drugi, poboljšavaju sposobnosti zemalja članica NATO saveza kroz obrazovanje i obuku koristeći najbolje dokazane metode i prakse primjenom tehnologija za *e-učenje* i šire za *cjeloživnotno učenje*.

Zemlje članice NATO saveza će svojim sposobnostima preoblikovati i proširiti ponudu raspoloživih programa, tečajeva i alata i na ostala područja obrazovanja i obuke u obrambenim i vojnim sustavima [5].

Raspoloživi programi, tečajevi i alati aktivno će primjenjivati i najnovija tehnološka postignuća za e-učenje kroz različite mehanizme i nove tehnološke komponente kao što su npr. virtualne učionice, webcast-inzi i druge interaktivne inicijative.

VII. ZAKLJUČAK

U informacijskom društvu i vremenu *e-učenje* kao i *distribuirano* odnosno *učenje na daljinu* dobivaju sve više na značenju. U vojnom obrazovanju primjenjuje se *e-učenje* putem *JADL*-a i *ADL*-a, odnosno naziv za *e-učenje* u zemljama članicama NATO saveza je *ADL*. U većini institucija u zemljama članicama NATO saveza i Partnerstva za mir za potrebe distribuiranog i e-učenja koriste se sustavi za upravljanje učenjem tj. LMS-ovi MOODLE i ILIAS. Razvijen je i ponuđen veliki broj „gotovih" *online* tečajeva za potrebe obrambenog i vojnog sustava. Stvoreni su potrebni preduvjeti za implementaciju *e-učenja* i u obrambenom i vojnom sustavu RH. U MORH-u i OS RH odnosno u HV-u osobita se pozornost posvećuje i *cjeloživotnom učenju*.

„Znanjem u budućnost" moto je sadašnjih naraštaja i ostat će moto i naraštajima koji će doći na vojno obrazovanje u škole Hrvatskog vojnog učilišta „Petar Zrinski" u Zagrebu.

LITERATURA

[1] A. W. (Tony) Bates, „Upravljanje tehnološkim promjenama - Strategije za voditelje visokih učilišta", *CARNet* - Zagreb i *Benja* - Lokve, 2004. (prijevod), Izvornik: „Managing Techological Change: Strategies for College and University Leaders", *Jossey-Bass Publishers*, San Francisco, USA, 2000.

[2] B. Shackelford, „Project Managing E-Learning", *American Society for Training & Development Press*, Alexandria, USA, 2002.

[3] M. M. Lynch & J. Roecker, „Project Managing E-Learning", A handbook for successful design, delivery and management, *MPG Books Ltd*, Bodmin, Great Britain, 2007.

[4] J. Greene & A. Stellman,"Head First PMP" A Learner's Companion to Passing the Project Management Professional Exam, A Brain-Friendly Guide, *Oreilly Media, Inc*., CA, USA, 2007.

[5] Grupa autora, „E-Lerning Concept", NATO unclassified, *NATO*, *Bruxelles, Belgium*, 8.2.2012.

Sigurnost i mobilnost za sve

Ivana Ružić
I. osnovna škola Čakovec,
Ulica kralja Tomislava 43, Čakovec, Hrvatska
ivana.ruzic@skole.hr

Sigurnost i mobilnost za sve je međunarodni obrazovni program kojem je cilj osvijestiti i potaknuti djecu i mlade na različite aktivnosti kako bi djelovali u svojoj okolini razvijajući i njegujući odgovornije i sigurnije ponašanje pojedinaca i grupa u prometu. Učenici I. osnovne škole Čakovec osmislili su i proveli niz aktivnosti koristeći IKT: istraživanje, izrada obrazovnih sadržaja, izrada obrazovnih računalnih igara, web stranica, izrada letaka, plakata i slikovnice itd. Za bolje planiranje različitih aktivnosti i praćenje njihove realizacije, učenici su se upoznali sa pojmom gantograma, njegovom ulogom u projektu te alatom za izradu gantograma. Projekt se provodio u vremenskom periodu od četiri mjeseca u sklopu izborne nastavne informatike i izvannastavne aktivnosti Informatička skupina, a uz učenike svih dobnih skupina u projekt su bili aktivno uključeni roditelji, suradnici iz PU Međimurske i HAK-a, 17 škola širom Hrvatske i cijela lokalna zajednica.

I. UVOD

U današnje doba svaka organizacija nastoji ostvariti što bolju stratešku prednost kako bi uspješnije funkcionirala. U obrazovnim ustanovama u nastavu se sve više uvodi projektni način rada njegujući timski duh među učenicima i nastavnicima te priprema učenika za budućnost tj. kako bi ostvarili što veću konkurentnost na tržištu rada te se adekvatno pripremili na cjeloživotno učenje.

Kako bi se bilo koji projekt mogao u potpunosti uspješno realizirati potrebno je što točnije isplanirati sve projektne aktivnosti. Postoje različiti alati koji nam pomažu u izradi plana projekta, a u ovom radu ćemo podrobnije objasniti Ganttov dijagram (gantogram) i njegovu primjenu prilikom planiranja i provođenja projektnih aktivnosti u okviru izborne nastave informatike u osnovnoj školi.

Uz Ganttov dijagram u ovom projektu učenici su upoznali, stekla nova znanja i vještine u korištenju različitih programskih alata ovisno o vlastitim interesima i sklonostima: programske alate za 3D modeliranje, stolno izdavaštvo, izradu obrazovnih računalnih igara, izradu multimedijalnih obrazovnih sadržaja, web 2.0 alate itd.

II. SIGURNOST I MOBILNOST ZA SVE

Sigurnost i mobilnost za sve je jedinstven obrazovni program koji su zajednički razvili Renault i obrazovni stručnjaci, provodi se u 45 zemalja širom svijeta posljednjih 10 godina, a ciljano je osmišljen za učenike osnovnih i srednjih škola. Program ima cilj poticati, razvijati i jačati podučavanje i učenje o sigurnosti u prometu i održive mobilnosti. Program usmjerava učenike prema promatranju različitih situacija u prometu, razumijevanju svijeta oko sebe i razmišljanju o značenju i utjecaju njihovog kretanja.

Slika 1. Logo programa Sigurnost i mobilnost za sve

Sigurnost i mobilnost za sve je obrazovni program koji podizanjem svijesti o važnosti vlastitog ponašanja omogućuje mladim ljudima da osmisle, predlažu i provode kampanju podizanja svijesti u svojoj školi, susjedstvu i široj zajednici, koristeći crteže, plakate, video i druge sadržaje i aktivnosti kroz koje će prenijeti snažne poruke na temu sigurnosti na cestama i održivu mobilnost za sve. Ovaj program prate i obrazovni materijali prilagođeni za djecu i mlade prema dobnim skupinama koje su izradili obrazovni stručnjaci.

Program Sigurnost i mobilnost za sve provodio se kao četveromjesečni projekt u osnovnoj školi sa ciljem osmišljavanja i provođenja različitih aktivnosti učenika svih dobnih skupina, izrade edukativnih elektroničkih i tiskanih sadržaja kako bi učenici prepoznali vlastita rizična i neodgovorna ponašanja, opasnosti na koje nailaze u svakodnevnom kretanju do škole, prijatelja i sl., ohrabrili odgovorna i poželjna ponašanja te podigli svijesti o važnosti vlastitog ponašanja u prometu u svom razredu, školi, u obitelji, među prijateljima i široj lokalnoj zajednici.

Aktivnosti u sklopu projekta podržane su IKT te su učenici upoznali i u radu primjenjivali različite programske alate (koji nisu obuhvaćeni propisanim nastavnim planom i programom) za praćenje projektnih aktivnosti, izradu digitalnih i interaktivnih obrazovnih sadržaja, izradu 2D i 3D obrazovnih računalnih igara, izradu 3D modela, izradu i objavu web stranica te stolno izdavaštvo.

III. PROJEKTNE AKTIVNOSTI UČENIKA

Učenici su samostalno vodeći se vlastitim interesima i sklonostima osmislili niz aktivnosti za ostvarenje postavljenih ciljeva projekta Sigurnost i mobilnost za sve.

Aktivnosti učenika:

izrada vlastitog loga programa,

objava natječaja za najbolju učeničku fotografiju na zadanu temu, odabir najboljih fotografija, otvaranje izložbe i objava najboljih fotografija,

snimanje filma o rizičnim ponašanjima u prometu u neposrednoj blizini škole,

izrada animiranog filma o sigurnom ponašanju u prometu,

predavanja i razgovor učenika s djelatnicima PU Međimurske i Auto kluba Čakovec,

identifikacija problema, provođenje oluje mozgova, provođenje istraživanja, obrada i analiza prikupljenih podataka, izrada mentalnih mapa, plakata, digitalnih slikovnica, kvizova i sl. u 6 različitih područja:

a. sigurnost u prometu,

b. sigurnost u cestovnom prometu,

c. rizična i neodgovorna ponašanja u prometu,

d. odgovorna i poželjna ponašanja u prometu,

e. prevencija i

f. automobili budućnosti.

izrada računalnih igara o sigurnom ponašanju u prometu:

a. Nebeski promet - 3D računalna igra o kojoj igrač ispunjava misije gradom vozeći bicikl po prometnim pravilima,

b. Dorianova avantura – 2D računalna igra u kojoj igrač sakuplja zadane vrste prometnih znakova

predstavljanje programa u javnosti i medijima,

izrada web stranica programa i objava svih učeničkih aktivnosti i sadržaja.

Slika 2. Prikaz 3D igre Nebeski promet

Slika 4. Web stranica projekta

Sve aktivnosti su podržane IKT. IKT je osigurala veći interes i motiviranost učenika za sudjelovanje u aktivnostima, potaknula je učenike na samostalno istraživanje i kritički odnos prema IKT.

Aktivnosti učenika I. Osnovne škole Čakovec prepoznate su kao uspješnije među svim sudionicima te su učenici pozvani u Pariz na konferenciju Road safety 2012. gdje su za svoj rad osvojili drugo mjesto u svijetu.

Slika 5. Konferencija Road safety 2012.

IV. GANTOGRAM I NJEGOVA PRIMJENA U PLANIRANJU, PROVOĐENJU I PRAĆENJU PROJEKTNIH AKTIVNOSTI

Gantogram (još i Ganttov dijagram, Ganttov grafikon i metoda grafičkog prikazivanja informacija) tip je stupčanog grafikona koji se koristi za grafički prikaz rasporeda projekta. Gantogrami ilustriraju početni i krajnji datum nekih nepromjenjivih i sažetih elemenata projekta. Nepromjenjivi ili terminalni elementi kao i sažeti elementi obuhvaćaju podjelu radne strukture.

Gantogram se može promatrati kroz koordinatni sustav. Kao takav, na ordinati bi se nalazili objekti terminiranja a na apscisi vremenske (terminske) jedinice. Princip se primjenjuje za slaganje grafikona, dakle na vertikalnu os se postavljaju objekti tj. aktivnosti koje se trebaju obaviti, a na horizontalnu os vrijeme.

Za vrijeme je najvažnije projektno vrijeme. To je vrijeme koje je potrebno ili predviđeno za odvijanje svih aktivnosti u jednom projektu. Uz projektno vrijeme važno je uzeti u obzir i vremensku rezervu. To je veće raspoloživo vrijeme od onog što je potrebno da bi se obavila neka aktivnost. Najčešće se pojavljuje kod onih

aktivnosti koje nisu na kritičnom putu. Kritični put je put koji u projektu traje najdulje, odnosno zbroj trajanja aktivnosti u njima je najveći. Dakle, vremenska rezerva je zapravo vrijeme kojemu ne-kritična aktivnost može biti odgođena bez da se odgodi cijeli projekt.

Bitno je napomenuti da ne mora biti samo jedan kritični put. Ukoliko ih je više, posebna pažnja se posvećuje čvorištima. Čvorišta tada predstavljaju aktivnosti koje su zajedničke više kritičnih puteva. Ukupno vrijeme potrebno ili predviđeno za odvijanje svih procesa u sustavu je sustavno vrijeme, dakle onda bi najkraće sustavno vrijeme bilo vrijeme kritičnog puta.

Važno je uočiti razliku između realnog tj. stvarnog vremena i planiranog vremena. Oni se također mogu paralelno prikazivati na gantogramu jer predstavljaju logičke suprotnosti. Naravno, bitni elementi za prikaz su i milestones/označivači koji su važni događaji koji predstavljaju odnose aktivnosti, te označuju početak tj. kraj.

Gantt bars ili položeni stupci su glavni pokazatelj protoka vremena unutar grafikona za neku aktivnost..

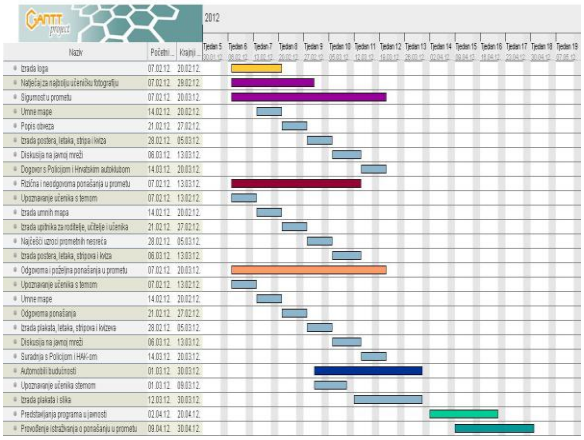
Za planiranje i praćenje realizacije projektnih aktivnosti korišten je alat GanttProject. To je jedan od mnogih besplatnih alata za upravljanje projektima ali mu je jedna od prednosti što radi pod Microsoft Windows 7, Microsoft Vista i Windows XP, Mac OS X i Linux operacijskim sustavima.

Uz velike mogućnosti organiziranja grafikona, dodjeljivanje sredstava za zadatke, te prikazom PERT dijagrama odgovarajućih poslova, čine ga jednostavnim za korištenje. GanttProject je kartično sučelje koje olakšava prijelaz između zadataka, i upravljanje zadacima i resursima. Pomoću dijagrama može se uspostaviti zavisnost između resursa i zadataka. Mogu se generirati PNG slike, PDF, HTML i izvješća za izvoz podataka u tablice u CSV formatu. Također može se izvršiti uvoz i izvoz podataka u alat Microsoft Project.

Gantogram tj. grafički prikaz rasporeda provođenja aktivnosti unutar projekta idealno je sredstvo za bolje razumijevanje, organizaciju i provođenje projektnih aktivnosti. Učenici osmih razreda osnovne škole su se po prvi puta susreli sa ovakvom vrstom grafikona i alatom za izradu tih grafikona na računalu. Radoznalost učenika, otvorenost prema različitim računalnim alatima te razumijevanje prednosti korištenja alata u organiziranju i provođenju projektnih aktivnosti, Ganttproject su učinili zanimljivim i korisnim alatom. Uz spomenuti alat, učenici su bolje upoznali projektni način rada posebice važnost planiranja aktivnosti u vremenu.

Prednosti korištenja gantograma su osigurana preglednost i razumljivost, jednostavna i laka izrada (za mali broj aktivnosti), jednostavno prikazivanje završenosti aktivnosti.

Uočeni su i nedostaci korištenja gantograma, a to su teškoća u izradi i nepreglednost kod velikog broja aktivnosti, ne vodi računa o resursima kod raspoređivanja aktivnosti, prikazivanje završenosti je neprecizno i teško je znati ukupnu završenost u traženom trenutku, podaci koji se dobiju iz gantograma su nedovoljni za kontrolu odvijanja realizacije, a pomoću gantograma je teško uočiti uvjetovanost pojedinih aktivnosti.



Slika 6. Prikaz projektnih aktivnosti pomoću gantograma

V. ZAKLJUČAK

Međunarodni projekti mogu obogatiti cjelokupan nastavni proces, omogućujući učenicima da upoznaju nove zemlje i krajeve, njihove običaje, svakodnevnicu i svoje vršnjake. Ukoliko učenici temu projekta prepoznaju kao korisnu i u svakodnevnom životu primijenjivu, interes i motiviranost učenika u projektnim aktivnostima je veći. IKT je nezaobilazna u svakoj sferi života, a upoznavanje i savladavanje novih programskih alata ovisno o interesima učenika osigurava zadovoljstvo učenika i veću angažiranost. Timski rad, podjela aktivnosti po timovima, praćenje realizacije aktivnosti i njihova evaluacija unutar tima, potiče jačanje suradnje unutar članova tima, bolju komunikaciju i organizaciju u svakodnevnim obavezama unutar grupe tj. razreda.

LITERATURA

[1] Bates, A. W.: Upravljanje tehnološkim promjenama: Strategije za voditelje visokih učilišta.. Zagreb, Hrvatska: CARNet/Benja, 2004.

[2] Jenkins, N.: A Project Management Primer or a guide on how to make projects work, 2005.

[3] http://pdfcast.org/pdf/kori-tenje-ganttovog-dijagrama-za-planiranje-projektnih-aktivnosti

Motivacija učenika za e-učenje

Antonela Czwyk Marić, prof.
Zdravstvena škola, Split, Republika Hrvatska
antonela.czwyk-maric@skole.hr

Sažetak - ***Rad je rezultat istraživanja o motiviranosti i primjerenosti e-učenja za učenike strukovne škole. Učenici su imali priliku slobodnim izborom učiti uz pomoć online tečaja kako bi postigli bolji uspjeh iz predmeta Informatika. Kvalitativnim istraživanjem dobiveni su stavovi učenika te je nastavniku omogućeno izraditi nove smjernice o korištenju e-učenja u nastavi s ciljem postizanja veće motivacije, lakšeg usvajanja znanja kao i osposobljavanja učenika za cjeloživotno učenje.***

I. UVOD

U strukovnim školama nema sustavnog uvođenja e-učenja. To ovisi o uvjetima rada škole, njenoj tehničkoj opremljenosti, informatičkoj platformi, a prvenstveno o entuzijazmu njezinih nastavnika. Mnogi učenici ostaju uskraćeni za takvo iskustvo, iako rado prihvaćaju napredne tehnologije i učenje pomoću digitalnih nastavnih materijala, vješto koriste mnoge usluge interneta, brzo usvajaju vještine rada na računalu, te rado komuniciraju putem računala. Postavlja se pitanje jesu li pri tome zainteresirani samo za zabavu ili su motivirani i za e-učenje ako im je ono dostupno.

A. Cilj istraživanja

Cilj istraživanja je procjena motiviranosti učenika strukovne škole za e-učenje, te primjerenost e-učenja kao oblika učenja u redovnoj nastavi. Da bi se cilj ostvario učenicima je ponuđeno pohađanje e-tečaja, a na temelju svojih iskustava učenici su kroz razgovor izrazili mišljenje o svojoj motiviranosti za ovakav način učenja ali i dali svoju procjenu moguće motivacije za ostale učenike. Učenici su mogli samostalno učiti kroz online tečaj „Excel 2007“ kako bi postigli bolji uspjeh iz predmeta Informatika na kraju školske godine.

Tečaj je osmišljen za samostalno učenje i uključuje samotestiranje, provjeru znanja i predaju zadataka na ocjenjivanje svom nastavniku. Ponuđen je svim učenicima u tri druga razreda koji uče predmet Informatika i to ne kao obavezan zadatak već kao dodatna aktivnost. Tečaj je preporučen grupi od 72 učenice i 18 učenika (ukupno 90) u dobi od 15-16 godina, koji uče informatiku drugu godinu po redovnom programu, a školuju se po novom programu u trajanju od 5 godina za zanimanje medicinska sestra / medicinski tehničar u Zdravstvenoj školi u Splitu.

B. Opis situacije

Tečaj [1] je dizajniran u sustavu za upravljanje učenjem Moodle [2], zahtjeva internet vezu, osigurava online pomoć i povratnu informaciju mentora (nastavnika). Obuhvaća nastavnu jedinicu: „Osnove rada u programu Excel 2007“. Autorica je organizirala tečaj kroz šest poglavlja kroz koja učenici postupno upoznaju osnovne mogućnosti programa. Poseban je naglasak stavljen na praktičan rad i uvježbavanje naučenog. Na kraju svakog poglavlja, kroz završni zadatak i test moguće je provjeriti usvojenost znanja. Ovaj je tečaj prvenstveno namijenjen učenicima srednjih škola kojima je Računalstvo/Informatika redoviti predmet i učenicima osnovnih škola kojima je Informatika izborni predmet. Tečaj mogu koristiti i svi oni koji žele naučiti rukovati osnovnim mogućnostima programa MS Excel 2007 [1]. Učenici su imali sva potrebna predznanja: osnovna uporaba računala (MS Windows), rad s web preglednikom i poznavanje komunikacije putem foruma i elektroničke pošte.

Učenici su pristupali tečaju od kuće uz pretpostavku da na svojim kućnim računalima imaju instaliran odgovarajući program. Za sve učenike koji nisu imali uvjete potrebne za rad kod kuće, osiguran je rad u školi, na računalima koji imaju instaliran traženi program i pristup internetu. Vrijeme korištenja su mogli uskladiti s njihovim potrebama ovisno o njihovom rasporedu nastave i obaveza, te raspoloživim vremenom za učenje. Jedini nedostatak u odnosu na rad od kuće je nemogućnost korištenja školskog računala iza 20 sati, odnosno prije 8 sati ujutro. Učenici su mogli rješavati zadatke u vremenskom periodu od 15 dana (30. svibnja – 13. lipnja 2012.), a samom sadržaju tečaja (lekcijama, vježbama i rezultatima) pristupati do kraja školske godine (31.08.2012.)

Online tečaj pripremljen je u sustavu Moodle na CARNetovom poslužitelju i omogućava samostalno učenje uz pomoć mentora. Za uspješno rješavanje zadataka postavljenih u tečaju učenici su trebali:

- prijaviti se u sustav Moodle svojim AAIEduHr identitetom (korisničko ime i lozinka u sustavu HUSO), redovito pratiti informacije na forumu (diskusija) od strane nastavnika, te postavljati pitanja i tražiti eventualnu pomoć kod rješavanja
- proučiti nastavne materijale, izraditi zadatke za vježbu (vježbanice), izraditi ispitne zadatke na kraju svake cjeline i predati ih nastavniku u zakazanom roku na ocjenjivanje putem online obrasca (ukupno 7 zadataka).
- zadatke rješavati redom, jedan po jedan (nakon što je prethodni bio ocijenjen uz dobivanje povratnih informacija) kako bi se izbjeglo ponavljanje istih grešaka i osigurao napredak u svladavanju gradiva

Učenici su dobili povratnu informaciju o izrađenim zadacima: bodove, ocjenu, komentare o dobrim i lošim

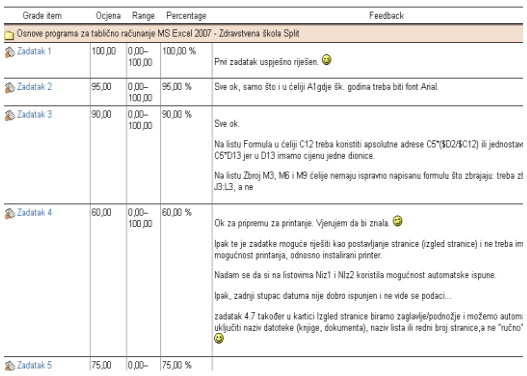
izabranim postupcima, te upute za bolji način rješavanja zadatka (Slika 1.).

Kvantitativni podaci koji se mogu dobiti iz statistike sustava Moodle omogućili su uvid u način rada učenika, njihove aktivnosti, redoslijed i učestalost posjećivanja pojedinih lekcija, izvođenja zadataka za vježbu, vrijeme provedeno u pojedinoj aktivnost, posjećenost foruma i chata. Za bolji pregled aktivnosti učenika bilo bi korisno imati uvid u povijest razgovora učenica u realnom vremenu (chat) koji su prema svojim izjavama često koristili kao sredstvo međusobnog pomaganja i razmjenu ideja o načinima rješavanja zadataka. Time bi se dobio bolji uvid u angažiranost pojedinih učenica u radu, ali i pregled problema na koja su nailazili u radu i pokušavali zajedno riješiti. No, autorica nije omogućila nastavnicima pristup takvim informacijama.

Dobivenim statističkim podacima uočene su neke činjenice:

- učenici su svakodnevno pristupali tečaju, a najčešći termini rada na računalu bili su u večernjim satima, odmah iza nastave (zbog specifičnih uvjeta rada škole učenici imaju organiziranu nastavu uvijek popodne uglavnom od 13 do 19 sati)
- često su pristupali testovima za samoocjenjivanje prije pristupanja samim lekcijama
- ponavljali bi samotestiranje više puta kako bi ostvarili što bolji rezultat
- zadatke su redovito predali nakon što su proučili lekcije i zadatke za vježbu

Iz navedenih podataka ne može sa sigurnošću utvrditi postoji li (i u kojoj mjeri) motiviranost učenika za e-učenje. Jako je važno dobiti informaciju o zadovoljstvu učenika ovakvim načinom učenja, kao i njihova razmišljanja o uočenim prednostima i nedostacima ovakvog rada. U tu svrhu mogla se pripremiti anketa za učenike s pitanjima koji bi dali potrebne informacije. Ipak, relevantnije i pouzdanije odgovore ovakvog subjektivnog tipa svakako je bolje prikupiti služeći se kvalitativnim metodama istraživanja.



Grade item	Ocjena	Range	Percentage	Feedback
Osnove programa za tablično računanje MS Excel 2007 - Zdravstvena škola Split				
Zadatak 1	100,00	0,00–100,00	100,00 %	Prvi zadatak uspješno riješen.
Zadatak 2	95,00	0,00–100,00	95,00 %	Sve ok, samo što i u ćeliji A1gdje šk. godina treba biti font Arial.
Zadatak 3	90,00	0,00–100,00	90,00 %	Sve ok. Na listu Formula u ćeliji C12 treba koristiti apsolutne adrese C5*($D2/$C12) ili jednostav C5*D13 jer u D13 imamo cijenu jedne dionice. Na listu Zbroj M3, M6 i M9 ćelije nemaju ispravno napisanu formulu što zbrajaju: treba zb J3:L3, a ne
Zadatak 4	60,00	0,00–100,00	60,00 %	Ok za pripremu za printanje. Vjerujem da bi znala. Ipak te je zadatke moguće riješiti kao postavljanje stranice (izgled stranice) i ne treba im mogućnost printanja, odnosno instaliran printer. Nadam se da si na listovima Niz1 i Niz2 koristila mogućnost automatske ispune. Ipak, zadnji stupac datuma nije dobro ispunjen i ne vide se podaci... zadatak 4.7 također u kartici Izgled stranice biramo zaglavlje/podnožje i možemo autom uključiti naziv datoteke (knjige, dokumenta), naziv lista ili redni broj stranice,a ne "ručno"
Zadatak 5	75,00	0,00–	75,00 %	

Slika 1. Ocjene i povratne informacije nastavnika

II. METODA ISTRAŽIVANJA

A. *Izbor metode istraživanja*

Denzin i Lincoln (1994) odredili su kvalitativno istraživanje kao višemetodski pristup koji počiva na tumačenju subjektivnih pojava [3]. Kvalitativna analiza može produbiti naše spoznaje o izučavanoj pojavi, te baciti na nju posve novo svjetlo, odnosno uputiti na ono što kriju subjektivna iskustva nedohvatljiva drugim sredstvima. Dakle, istraživač nastoji putem duboke pozornosti i empatičnog razumijevanja „iznutra“ prikupiti građu o viđenju samih aktera, te provesti analizu riječima (opisom, transkriptom) što su neke od temeljnih obilježja kvalitativnog pristupa. [4]

Kvalitativna metoda u ovom istraživanju ostvaruje slijedeće ciljeve:

- opisuje i tumači učenička iskustva o e-učenju
- produbljuje nastavnikovo poimanje o načinima korištenja e-učenja u svojoj redovnoj nastavi
- približava učeničke osjećaje i doživljaje tokom e-učenja radi dubljeg razumijevanja što može pomoći u daljnjem planiranju nastavnih aktivnosti
- uočava neočekivane i nepredvidljive pojave (poteškoće u pristupu sadržajima i samom izvođenju zadataka, nerazumijevanje nastavnog sadržaja, neprimjerena težina…)
- saznaje što ih motivira na ovaj oblik učenja (razumijevanje uzročnih objašnjenja)
- omogućava nastavniku nesputano i kreativno razmišljanje o primjeni e-učenja u nastavi, osvježava pristup i eventualno generira nove, slikovitije i sadržajnije pretpostavke
- prikuplja polazne podatke za konstrukciju upitnika koji bi se mogao implementirati u sam tečaj kao način evaluacije e-učenja i mjerenje učenikove razine zadovoljstva

Uloga ispitanika (učenika) je da zajedno sa istraživačem (nastavnikom) budu subjekti u ovom istraživanju kako bi uzajamnom suradnjom ostvarili razumijevanje dobivenih podataka [4]. Učenici imaju važnu ulogu jer aktivno sudjeluju u procjeni valjanosti. Sam nastavnik kao istraživač vlastitom perspektivom utječe na tumačenje pojma e-učenja i njegove primjene u redovnoj nastavi, ali i svojom uključenošću i empatijom potiče učenike da slobodno i iskreno, bez ustezanja govore o svojim iskustvima.

Prikupljeni su podaci u obliku:

- statističkih podataka
- transkripata razgovora (podaci sadržani u riječima, tekstu temeljnom na intervjuu)

U kvalitativnoj metodi koriste se razne istraživačke metode: dubinski intervju, fokus grupe i opažanje. Namjera nastavnika je doći do opisa različitih unutarnjih doživljaja (misli i osjećaja) i ponašanja učenika u ovoj određenoj situaciji.

B. *Fokus grupa*

S obzirom broj sudionika i ciljeve ovog istraživanja smatralo se najprikladnijim koristiti metodu fokus grupe (6-9 učenica). Fokus grupa primjerena je promatranom istraživanju jer ima slijedeća svojstva po Kruegeru [5]: provode sa na manjim grupama osoba koje dijele neka zajednička obilježja, razgovor je usmjeren prema određenom problemu, provode se u nizu grupnih razgovora, pružaju podatke kvalitativne prirode. Pokazalo se zanimljivim ispitati dvije fokus grupe:

1. grupa učenika koja je uspješno riješila većinu zadataka i ostvarila svoj cilj (usvojila znanja i postigla bolje zaključne ocjene iz predmeta Informatika)
2. grupa učenika koji su se prijavili u sustav, pokazali interes za rad, započeli proučavati lekcije, ali nisu predali ni jedan zadatak na ocjenjivanje te nisu ostvarili zadani cilj.

III. POSTUPAK ISTRAŽIVANJA

A. *Planiranje fokus grupe*

Istraživanju su bili pozvani svi učenici koji su pristupili tečaju ili su u bilo kojem trenutku to namjeravali. Svi su obaviješteni o načinu i cilju istraživanja. Planirano je provesti ponavljanje i provođenje fokus grupe u nizu (nastavak istraživanja). To ovisi o organizaciji nastavne godine i raspoloživosti nastavnika i učenika za daljnji rad. Kroz pripremu je trebalo informirati učenike o temi istraživanja, dogovoriti prikladan termin i osigurati prostor. Pitanja su bila otvorenog tipa (bez ponuđenih odgovora). Prilikom planiranja i traženja prikladnog termina, samo se jedna učenica iz 2. fokusne grupe odazvala pozivu i iskazala interes za sudjelovanje, ali ipak nije pristupila istraživanju. Moguće je da su ostali osjećali nelagodu zbog neuspjeha, iako je nastavnik istaknuo kako ima potpuno razumijevanje za njihovo odustajanje i da ih nikako ne smatra neuspješnim. Iz toga razloga ispitana je samo 1. fokusna grupa usmjeravana od strane nastavnika/ moderatora.

B. *Provedba (pro/vođenje razgovora)*

Scenarij je bio osmišljen pomoću pripremljenih pitanja po uzoru na metode istraživanja u obrazovanju [6]. U transkriptu su navedena postavljena pitanja i dogovori ispitanika. Postavljeno je ukupno 16 pitanja. U tablici (TABLICA I.) su navedena neka zanimljivija pitanja i pripadni odgovori učenica. Izostavljeni su samo odgovori koji su se ponavljali.

Istraživanju je u zakazano vrijeme pristupilo 7 učenica. Grupa je homogena: sve su učenice (spol), završile su 2. razred (dob), smjer medicinska sestra (usmjerenje i predznanje). Učenice su bile motivirane i zainteresirane za razgovor o temi, iako nisu za to bile ni na koji način nagrađene (ne uče više taj predmet, već su usmjerene na sasvim nova nastavna područja - struka i novu grupu nastavnika). Motivaciju za razgovor su crpile iz osjećaja zadovoljstva nakon uspješnog e-učenja, poštovanja prema nastavniku, ali i ponosa jer se uvažava njihovo mišljenje. Razgovor je sniman na računalu programskim alatom Adobe Conect Pro (video i audio) uz pristanak učenica.

C. *Analiza rezultata*

U ovom istraživanju analiza podataka je izvedena interpretacijom jer ljudsko ponašanje zahtjeva duboko razumijevanje, empatiju; pomno se iščitavaju prikupljeni podaci (ne sažimajući ih) tražeći suštinu [7].

Iz transkripta (TABLICA I.) može se uočiti slijedeće:

- pitanja su navedena spontano i prirodno, slijedom unaprijed zacrtanog plana, razumljivo, logično, otvoreno i nesugestivno (da se izbjegne utjecaj već oblikovanih stavova o e-učenju)
- pitanja su bila usmjerena na problem (fokusirana na iskustvo učenika u e-učenju)
- učenice su se međusobno poznavale, iako nisu bile iz istog razreda, što ipak nije ometalo samu raspravu, jer su na početku postavljena pravila: redoslijed iznošenja mišljenja nakon postavljenih pitanja uvijek je započeo od druge učenice (ciklička izmjena)
- moderator (nastavnik) nastojao je poticati grupnu interakciju i dinamičnost (u nekoliko navrata želeći što prije izreći svoje mišljenje učenice su bile spontane i zaboravljale na dogovoreni redoslijed na što se nije reagiralo dajući im vremena da nadopunjuju svoje iskaze)
- ostvaren je neformalan ugođaj (učenice su ponuđene sokom i čokoladom) kako bi se opustili i slobodno iznosili vlastita mišljenja i stavove, nesputan razgovor je ostvaren jer je sastav grupe bio homogen
- posebno opažanje istraživača: na licima učenica vidljivo je bilo iskreno oduševljenje što su koristile e-učenje, zadovoljstvo postignutim i ponos da mogu preporučiti ostalim učenicima ovakav način učenja

Analiza podataka se temeljila se na tri međusobno povezana postupka

- redukcija - sažimanje postojeće građe sukladno teorijskom okviru
- prikaz podataka - tablično bez brojeva / verbalni opisi (transkript)
- izvođenje zaključaka - davanje smisla podacima i njihovom tumačenju

D. *Valjanost nalaza istraživanja*

Izvori problema u ovom istraživanju moguća su kao reaktivnost [4]: istraživačev/nastavnikov utjecaj na pojavu koju opaža, a puno manje kroz istraživačku pristranost koja se očituje kao holistička greška (peglanje podataka), bez elitističke pristranosti (precjenjivanje podataka dobivenih od informatora višeg statusa naobrazbe i intelektualnih sposobnosti) ili pristajanje uz „urođenička“ objašnjenja (usvajaju mišljenja skupine koju istražuju) [7]. Treba imati na umu da nastavnik ima već unaprijed razvijen pozitivan stav o e-učenju i time može

TABLICA I. TRANSKRIPT KVALITATIVNOG ISTRAŽIVANJA NA TEMU: E-UČENJE U STRUKOVNIM ŠKOLAMA

PITANJE	ODGOVORI
1. Kako vam se sviđa ideja samostalnog učenja?	- radimo u slobodno vrijeme - zanimljivije je - iščekivanje rezultata za svaki pojedini zadatak, ocjene od profesorice (mentor) - lakše zapamtim gradivo jer imam više izvora učenja: udžbenik, lekcije, vježbenice - zanimljiviji je rad na računalu nego klasičan rad u školi (zapisivanje u bilježnice, slušanje) - lakši način učenja, zadaci se mogu rješavati po želji, zadaci su zanimljivi - način da se zabavimo, ali i kroz zabavu naučimo
2. Kako ocjenjujete radne materijale za učenje-lekcije objavljene u digitalnom obliku?	- lekcije su bile jako korisne kao da koristimo udžbenik - pregledno, jasno, sve na jednom mjestu - kao da nam nastavnik predaje lekciju u razredu - korisne vježbenice koje su nas pripremale na zadatak
3. Jesu li vam testovi za samoocjenjivanje bili korisni?	- dobri jer smo odmah vidjeli rezultat testa i svoju ocjenu, uvidjeli što nismo naučili, te smo ponovili lekcije koje nam još nisu bile jasne - korisno ih je više puta ponavljati jer pitanja nisu uvijek bila ista pa smo mogli bolje provjeriti znanje - korisna su čak i kada su rezultati odlični, jer se može pokušati riješiti ih brže i spretnije, pa služe i za bolje pamćenje činjenica - više puta ponoviti testove da bismo bolje zapamtili
4. Procjenjujete li da vam je online učenje ometalo vaše uobičajene obaveze, učenje drugih predmeta?	- nije nas ometalo - to smo i tako radili uvečer umjesto gledanja televizije - ponekad je možda ipak trebalo učiti nešto drugo, ali nam to nije ugrozilo ocjene iz drugih predmeta
5. Smatrate li biste postigli iste rezultate i klasičnim načinom učenja (u razredu, pomoću udžbenika)?	- više smo ovako naučili, jer smo radili samostalno: čitali, listali, istraživali - radili smo zbog nekog cilja, da sebe zadovoljimo a ne nekog drugog - više smo naučili jer u razredu budu i oni koje to ne zanima , a ovdje su svi bili koji su zainteresirani da uče na ovaj način i o ovoj temi - zanimljivije nego klasična nastava, jer je svaka promjena dobra - dobro je da možemo raditi od kuće gdje smo opušteniji
6. Procjenjujete li da ste tečajem osim gradiva usvojili i neka druga korisna znanja? Koja?	- najviše smo naučili o Excelu, a sve ostalo je bilo usput - o načinu slanju zadataka - što sam više radila bila sam vještija i brža, pa mi je tijekom vremena za iste radnje trebalo sve manje vremena - naučila sam nešto i o Moodlu, kako ga koristiti, jer do sada nisam imala priliku raditi u online tečaju - timski rad (često smo išli na chat)
7. Jeste li bili zadovoljni sa svojim rezultatima?	- jako sam zadovoljna - ugodno sam se iznenadila kada sam vidjela veću zaključnu ocjenu - nisam dobila veću ocjenu, jer sam već imala odličan, ali sam bila sretna jer mi je Excel postao drag - na kraju mi je bilo žao što nisam stigla sve naučiti jer nije bilo više vremena - pomoglo mi je za bolji opći uspjeh - zaboravila sam na zaključnu ocjenu, jer mi je zadovoljstvo bilo učiti na ovaj način - osjećam da sam sposobna učiti i drugo na ovaj način, a to je uistinu dobar osjećaj - nije mi žao ni trenutka izdvojenog za rad
8. Smatrate li da je e-učenje primjereno učenicima vaše škole?	- zanimljivo je - treba ga koristiti već i u prvom razredu kako bi se prije upoznali sa ovakvim radom i navikli se - samo za one koji žele, jer tko ne želi to mu je komplikacija - svakako ih motivirati većom ocjenom - veća zaključna ocjena nije nužna, ali neka dobra ocjena u toku godine sigurno može biti nagrada - danas u doba računala i razvoja tehnologije obavezno bi trebalo biti e-učenje
9. Imate li kakvu poruku za nove generacije koje će imati priliku za ovakav način učenja?	- jako je zabavno i dobro, izgleda kao igra - isplati se - osim učenja, uživaš u istraživanju, dobro se osjećaš - kada probate vidjet ćete da je zanimljivo - u početku bude teško, ali brzo se zabaviš i krene ti u rutinu - treba probati, jer uz knjigu i vježbenice na tečaju nije teško, već zabavno

nehotice usmjeriti razgovor u smjeru isticanja prednosti e-učenja, a manje njegovoj kritici.

Provjera valjanosti kvalitete dobivenih podataka izvedena je prema Miles i Hubermanu [7] provjerom reprezentativnosti: učenici su tipični za strukovnu školu, iste dobi, istog spola i životnih interesa. Planira se nastaviti niz istraživanja sa novim fokus grupama kako bi se potvrdila valjanost. Preporuka je valjanost osigurati i triangulacijom koja bi se u nastavku istraživanja ostvarila: dodatnim intervjuom sa učenicima koji nisu sudjelovali u fokus grupi, provođenjem još jedne fokus grupe učenika koji su namjeravali koristiti e-učenje ali nisu uspjeli iz raznih razloga (loša internet veza, nedostatak vremena, neispravno računalo, nesnalaženje u tečaju). Postoji moguća reaktivnost od strane nastavnika koja će se u nastavku istraživanja izbjeći uključivanjem drugih nastavnika koji nisu toliko aktivni u e-učenju.

IV. ZAKLJUČAK

E-učenje je potrebno što više koristiti u redovnoj nastavi strukovnih škola, ali kao dodatni sadržaj, za učenike koji su voljni samostalno učiti. Potrebno ga je provoditi češće, za manje nastavne jedinice i u zadanom vremenskom razdoblju. Važan je za usvajanje informatičke pismenosti i osposobljavanja učenika za cjeloživotno učenje.

Učenici rado, sa zadovoljstvom prihvaćaju nove tehnologije i e-učenje. Zato im je potrebno to omogućiti.

Važno je osvijestiti prednosti takvog učenja kod što većeg broja učenika.

Istraživanje bi trebalo nastaviti u smjeru ispitivanja stavova učenika koji nisu htjeli prihvatiti e-učenje i uočiti koje su poteškoće utjecale na takvu odluku. Time bi nastavnik mogao bolje planirati implementaciju e-učenja u redovnu nastavu primjerenu potrebama i mogućnostima svih učenika.

LITERATURA

[1] Šutalo, Sanda (2011). Osnove programa za tablično računanje MS Excel 2007, http://www.carnet.hr/e-tecajevi_za_primjenu_u_nastavi/Osnove_Excel

[2] CARNet, sustav za online učenje Moodle, MS Excel 2007 – Zdravstvena-škola,Split,(10.04.2013.) https://moodle.carnet.hr/course/view.php?id=1986 ,

[3] Milas, Goran (2005). Istraživačke metode u psihologiji i drugim društvenim znanostima, Naklada Slap, 17, str.571.-613.)

[4] Ajduković, D. (2008) Odgovornost istraživača i valjanost kvalitativne metodologije. U:N.Koller-Trbović i A.Žižak (ur.) Kvalitativni pristup u društvenim znanostima. Zagreb:Edukacijsko rahabilitacijski centar

[5] Richard A. Krueger 1994). Focus groups**:** a practical guide for applied research, SAGE

[6] L.Cohen,L.Manion, K.Morissonn (2007). Metode istraživanja u obrazovanju, N. Slap, IV,15, 267-292str.

[7] Matthew B. Miles, A. Michael Huberman (1994). Qualitative Data Analysis, SAGE

Obrazovanje za poduzetništvo pomoću računalne tehnologije u nastavi vježbeničke tvrtke

Ivana Ivančić
Ekonomska i trgovačka škola Ivana Domca Vinkovci, Vinkovci, Hrvatska
ivanai753@gmail.com

Sažetak - Globalno okruženje ukazuje na sve značajniju potrebu obrazovanja za poduzetništvo kako bi se omogućilo razvijanje poduzetničkih kompetencija potrebnih suvremenom tržištu rada i povećanju konkurentnosti nacionalnog gospodarstva. Obrazovanje za poduzetništvo, koje je jedna od temeljnih europskih kompetencija, postalo je dio nacionalnog kurikuluma Republike Hrvatske i provodi se u ekonomskim školama kroz modularnu nastavu predmeta vježbenička tvrtka. Vježbenička tvrtka je simulacija poslovanja poslovne organizacije u kojoj učenici obavljaju radne zadatke u simuliranom uredu. Poslovanje vježbeničkih tvrtki omogućava učenicima i nastavnicima rad bez vremenskih i prostornih ograničenja te pruža veću fleksibilnost i primjenu različitih strategija i metoda učenja. Vježbenička tvrtka povezuje teorijsku nastavu sa stvarnom poslovnom praksom, a korištenje informatičkih aplikacija za stjecanje poduzetničkih kompetencija sastavni je dio njenog kurikuluma. Cilj rada je analizirati i prezentirati korištenje računalne tehnologije u nastavi vježbeničke tvrtke pomoću informacijskog modula SUVT kao on-line potpore u radu učenika strukovnih škola.

I. UVOD

Suvremeno tržište rada, koje zahtijeva sposobnost brze prilagodbe i fleksibilnost, ukazuje na sve značajniju potrebu obrazovanja za poduzetništvo kako bi se omogućilo razvijanje potrebnih kompetencija radnika i povećala konkurentnost nacionalnog gospodarstva. Radnici na suvremenom tržištu rada moraju biti sposobni stvarati, analizirati i transformirati informacije, djelotvorno komunicirati te organizirati i koordinirati poslovne aktivnosti. Od njih se traže razvijene komunikacijske vještine, informatička znanja te spremnost na daljnje učenje i usavršavanje [1].

Da bi odgovorila izazovima razvoja društva znanja i svjetskoga tržišta, Europska unija je odredila osam temeljnih kompetencija za cjeloživotno obrazovanje [2]. To su: komunikacija na materinjem jeziku, komunikacija na stranim jezicima, matematička kompetencija i osnovne kompetencije u prirodoslovlju i tehnologiji, digitalna kompetencija, učiti kako učiti, socijalna i građanska kompetencija, inicijativnost i poduzetnost te kulturna svijest i izražavanje.

Digitalna kompetencija odnosi se na osposobljenost za sigurnu i kritičku upotrebu informacijsko-komunikacijske tehnologije za rad u osobnom i društvenom životu te u komunikaciji. Njezini su ključni elementi osnovne informacijsko-komunikacijske vještine i sposobnosti: upotreba računala za pronalaženje, procjenu, pohranjivanje, stvaranje, prikazivanje i razmjenu informacija te razvijanje suradničkih mreža putem interneta. Inicijativnost i poduzetnost su pretpostavke uspješnog života pojedinca i razvoja društva.

Navedene kompetencije pružaju jednake šanse za učenje u društvu znanja, za zapošljavanje i očuvanje zaposlenosti, međugraničnu mobilnost, za razvoj kvalitetne osobnosti i aktivno uključivanje u društvo. One nisu samo sastavnica nacionalnih obrazovnih standarda, već omogućuju prijelaz između obrazovanja i tržišta rada što je moguće jedino ukoliko se uspostavi suradnja između obrazovnih ustanova i poslovne prakse [3].

Ovaj rad analizira koncept obrazovanja za poduzetništvo koji se, kao dio Nacionalnog kurikuluma Republike Hrvatske, provodi kroz program stručnog predmeta vježbenička tvrtka u zanimanju ekonomist.

Cilj rada je prezentirati i analizirati korištenje računalne tehnologije u nastavi predmeta vježbenička tvrtka. Učenici postaju zaposlenici koji s drugim tvrtkama posluju putem interneta. Pri tome koriste informacijski modul SUVT kao on-line potporu u svom radu.

II. OBRAZOVANJE ZA PODUZETNIŠTVO

A. Koncept obrazovanja za poduzetništvo

Danas postoje mnoge definicije poduzetništva koje navode različite liste osobina poduzetne osobe. Iako među njima postoje razlike, pokazuje se da su neke od osobina zajedničke kao što su: motivacija za postizanjem uspjeha, autonomija u izvršavanju zadataka, kreativnost, inicijativnost, spremnost na preuzimanje rizika, vjera u mogućnost kontrole vlastite sudbine i ustrajnost [4]. Poduzetništvo podrazumijeva kreativnost, inovativnost, sposobnost planiranja, organiziranja te vođenja projekata kako bi se postigli poslovni ciljevi. Kod zaposlenika potiče svjesnost o cjelovitosti rada i razvija sposobnost iskorištavanja prilika.

Nužno je razlikovati širi i uži koncept obrazovanja za poduzetništvo. Širi koncept u funkciji je razvijanja poduzetničkih sklonosti, vještina i sposobnosti. Uži,

specifični koncept obrazovanja i osposobljavanja za poduzetništvo, je u funkciji pokretanja poslovnih subjekata i upravljanja njima.

Razvoj poduzetničkih kompetencija ključan je za pretvaranje ideja u konkretne poslovne pothvate i zbog toga se reforme u europskim obrazovnim sustavima usmjeravaju na razvoj znanja o poslovanju i poduzetništvu. Iskustva europskih zemalja u obrazovanju za poduzetništvo pokazuju da se i u obrazovnoj praksi razvijaju i primijenjuju modeli koji se zasnivaju na širem razumijevanju pojma poduzetništva. Takvi modeli, koji ovisno o dobi djeteta, pored stjecanja znanja i vještina potrebnih za pokretanje poslova i samozapošljavanje, impliciraju i razvoj osobina ličnosti učenika koji se nalaze u osnovi poduzetničkih stavova i ponašanja [5].

Obrazovanje za poduzetništvo kroz poslovanje vježbeničkih tvrtki omogućava učenicima i nastavnicima rad bez vremenskih i prostornih ograničenja te pruža veću fleksibilnost i primjenu različitih strategija i metoda učenja. Vježbenička tvrtka povezuje teorijsku nastavu sa stvarnom poslovnom praksom, a korištenje informatičkih aplikacija za stjecanje poduzetničkih kompetencija sastavni je dio njenog kurikuluma.

B. Vježbenička tvrtka

Vježbenička tvrtka je virtualna učenička tvrtka koja posluje na virtualnom tržištu s ostalim vježbeničkim tvrtkama. Sve poslove obavlja u suradnji s tvrtkom partnerom koja je stvarni gospodarski subjekt [6].

Učenici na osnovi usvojenih znanja i vještina otvaraju vježbeničke tvrtke kroz čije poslovanje stječu i praktična iskustva zaposlenih u svim organizacijskim jedinicama, ali i iskustva vođenja i upravljanja poslovnim procesima. Uspostavljaju poslovne odnose sa stvarnim tvrtkama iz gospodarskoga sektora, čime stječu i vještine komuniciranja, poslovnog dopisivanja, pregovaranja i ugovaranja.

Suvremena ideja vježbeničke tvrtke u Hrvatskoj potječe iz sredine devedesetih godina prošloga stoljeća, kada su pojedini nastavnici i ravnatelji uspostavili kontakt s vježbeničkim tvrtkama u Austriji, priređivali prve skupove za stručno usavršavanje nastavnika te osnivali prve vježbeničke tvrtke. Od 2003. godine Ministarstvo znanosti, obrazovanja i sporta ovu ideju sustavno podržava. Danas je obvezni predmet strukovnog programa za zanimanje ekonomist u 3. i 4. razredu srednje škole.

Prije uvođenja koncepta vježbeničke tvrtke, obrazovanje za poduzetništvo u hrvatskim ekonomskim školama temeljilo se na poučavanju učenika u programima nastavnih sadržaja pojedinih strukovnih predmeta, npr. poduzetništva ili marketinga. U okviru njih učenici i danas primijenjuju stečena teorijska ekonomska znanja o stvaranju i realizaciji poduzetničke ideje, o izboru djelatnosti, imena i zaštitnoga znaka tvrtke, o postupku osnivanja tvrtke, raspoređivanju radnika na odgovarajuća radna mjesta, određivanju ciljnoga tržišta i slično. Vježbanje je usmjereno na poticanje kreativnosti učenika, logičko povezivanje i razmišljanje te promatranje stvarnosti.

Međutim, vježbenička tvrtka je viša razina simulacije poslovanja u kojoj učenici igranjem uloga postaju zaposlenici i obavljaju stvarne radne zadatke u simuliranom uredu. Ovakav model omogućuje oponašanje svih aktivnosti stvarnoga poduzeća na različitim stupnjevima zahtjevnosti jer je svaka vježbenička tvrtka podijeljena na odjele: odjel za ljudske potencijale, administracija, prodaja i marketing, nabava, financije i računovodstvo te logistika. Učenici prolaze kroz poslovanje svih odjela i u svakome od njih obavljaju specifične poslove.

Simulirani ured vježbeničke tvrtke zahtijeva veliku učionicu. Udaljenost između dva radna mjesta trebala bi biti oko 85 cm. Svaki odjel ima na raspolaganju računalo s odgovarajućim softverom koji uključuje programski paket MS Office te knjigovodstveni program. Računala su umrežena s računalom nastavnika, a od ostale opreme koristi se pisač, fotokopirni stroj, skener, telefon s faksom, internetski priključak, e-adresa, pano-ploča, ormar za dokumentaciju i uredski pribor.

Ključni poslovi koji se obavljaju su poslovi administracije, nabave, prodaje i marketinga, financija i računovodstva, unos podataka i obrada informacija,uporaba uredskih strojeva i računala, telefoniranje, fotokopiranje, kopiranje, slanje faksova i elektroničke pošte te pretraživanje internetskih stranica.

Važna prednost vježbeničke tvrtke je mogućnost ostvarivanja poslovne suradnje s vanjskim partnerima. Kao što poduzeća surađuju u stvarnome gospodarstvu, tako i vježbeničke tvrtke održavaju međusobne poslovne kontakte pa učenici mogu upoznati poslovnu kulturu svojih partnera u zemlji i inozemstvu.

III. RAČUNALNA TEHNOLOGIJA U NASTAVI VJEŽBENIČKE TVRTKE

A. Vježbeničke tvrtke i SUVT

Razvoj informacijskih i komunikacijskih tehnologija zahtijeva vještine kao što su sposobnost rješavanja problema, kreativnost, kritičko razmišljanje, preuzimanje rizika, suradnja i poduzetništvo [7]. Koncept vježbeničke tvrtke ostvaruje sveobuhvatnu metodu učenja i podučavanja. On je odgovor na stalne zahtjeve gospodarstva za primjenom praktične nastave, za boljim poznavanjem stranih jezika i uvođenjem najnovije tehnologije u nastavu.

Učenici radom u vježbeničkoj tvrtki stječu stručne kompetencije kao što su: samostalno korištenje informacijske i komunikacijske tehnologije te rad na računalu, odgovarajuća znanja iz računalstva koja se primjenjuju za pronalaženje podataka važnih za organizaciju i rad te njihovo korištenje u obavljanju postavljenih radnih zadataka, korištenje interneta i slanje elektroničke pošte.

Agencija za strukovno obrazovanje i obrazovanje odraslih osnovala je Središnji ured vježbeničkih tvrtki (SUVT) u okviru informacijskog sustava za strukovno obrazovanje VETIS [1].

[1] Više o sustavu VETIS na: http://www.aoo.hr/

SUVT je informacijski modul koji pruža potporu u radu svim učenicima kao mogućim vlasnicima vježbeničke tvrtke pri njezinoj registraciji, u vođenju njezina poslovanja kao i u slučaju potrebe njezina zatvaranja. Učenici mogu, korištenjem usluga SUVT-a, usvojena znanja iz teorije primijeniti u praksi i na taj način jačati timski rad. Radom u skupini uče uvažavati rad drugih uz mogućnost iznošenja vlastitog mišljenja. Usvajaju različite tehnike kako doći do poslovne ideje, samostalno izraditi poslovni plan, razviti prezentacijske vještine i potrebu za cjeloživotnim učenjem. Ovaj modul analizira poslovanje vježbeničkih tvrtki uz objavljivanje njihovih rezultata te pruža sve informacije o pravnim propisima važnim za poslovanje pravnih subjekata u Republici Hrvatskoj.

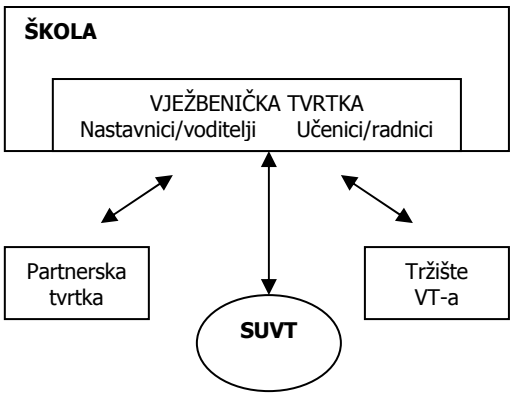
Za ulaz u modul SUVT potrebno je putem internetske veze ući u VETIS. Učenicima i nastavnicima dodjeljeno je korisničko ime i zaporka pomoću kojih mogu pristupiti SUVT-u, a svaki od njih može prijaviti i osnovati novu tvtku.

SUVT je istodobno i sustav za e-učenje. Pomoću njega je moguće iz učionice obavljati sve poslove registracije tvrtke, njezine prijave i prijave zaposlenika na mirovinsko i zdravstveno osiguranje, prijave platnog prometa i vođenja ukupnog poslovanja tvrtke. Slika 1. prikazuje kako izgleda početno korisničko sučelje SUVT-a.

Slika 1. Početna stranica Središnjeg ureda vježbeničkih tvrtki

SUVT potiče i koordinira poslovanje vježbeničkih tvrtki u zemlji i inozemstvu stvaranjem njihovog registra u Republici Hrvatskoj. Trenutno se u evidenciji SUVT-a[2] nalazi 196 vježbeničkih tvrtki čiji je popis prema županijama, školama, nazivima i registriranim djelatnostima lako dostupan kako bi se poslovanje između njih uspješno odvijalo.

Vježbeničke tvrtke surađuju s drugim vježbeničkim tvrtkama i SUVT-om te održavaju poslovne kontakte sa stvarnim tvrtkama na tržištu, a koje su njihove tvrtke partneri a na tržištu imaju svoje kupce i dobavljače. Na slici 2. prikazana je mreža vježbeničkih tvrtki kao otvoren sustav vježbeničkih tvrtki i SUVT-a koji međusobno komuniciraju.



Slika 2. Mreža vježbeničkih tvrtki

B. *Usluge SUVT-a*

Osnovna zadaća SUVT-a je pružanje usluga Poslovne banke, Trgovačkog suda, Zavoda za statistiku, Porezne uprave, Zavoda za mirovinsko i zdravstveno osiguranje i Registra osiguranika Rad institucija SUVT-a uključuje i prihvaćanje ili odbijanje pristiglih zahtjeva u obliku obrazaca ili dokumenata i nadzor nad radom vježbeničkih tvrtki.

Korištenjem usluga Poslovne banke učenici provode postupke otvaranja kunskoga i deviznog žiro-računa, uplate temeljnog kapitala potrebnog pri registraciji trgovačkog društva, razumiju i zaključuju kako pravne i fizičke osobe u stvarnosti posluju posredstvom banke u zemlji i inozemstvu. Sva teorijska znanja o bankarskim poslovima koriste izvođenjem pojedinih poslova i komuniciranjem s bankom koja obavlja poslove: vođenja kunskih i deviznih računa, provođenja naloga za plaćanje, naplate potraživanja i izdavanja izvoda o promjenama na žiro-računu putem internet bankarstva.

Korištenjem usluga Trgovačkog suda učenici provode postupak upisa trgovačkog društva odnosno vježbeničke tvrtke u Registar Trgovačkog suda. Svoja teorijska znanja o postupku registracije primijenjuju popunjavanjem prijave za upis. Nakon što obave uspješno upisivanje trgovačkog društva, dodjeljuje im se matični broj subjekta (MBS) te nakon toga mogu pristupiti sljedećem koraku cijelog postupka registracije trgovačkog društva.

Zavod za statistiku vježbeničkih tvrtki pruža uslugu dobivanja matičnog broja za trgovačko društvo prema Nacionalnoj klasifikaciji djelatnosti koja vrijedi u Republici Hrvatskoj. Tako učenici imaju priliku provjeriti jesu li odabrali dobru klasifikaciju djelatnosti svoga trgovačkog društva. Dobiveni će matični broj služiti i za Poreznu upravu, jer se prema njemu registrirana vježbenička tvrtka evidentira kao pravna osoba u svim registrima SUVT-a.

Učenici obavljaju i prijavu trgovačkog društva te zaposlenika vježbeničke tvrtke Zavodu za mirovinsko osiguranje i Zavodu za zdravstveno osiguranje. Moraju se pridržavati zakonskih rokova prijave subjekata kako bi svi

[2] Dostupno na: http://www.aoo.hr/suvt/default.aspx?id=69

zaposlenici mogli koristiti svoja prava iz radnog odnosa, a trgovačko društvo pravodobno izvršavati svoje obveze plaćanja propisanih doprinosa. Simulacijom tog postupka uče koliko je važno da u budućnosti, kao stvarni zaposlenici trgovačkog društva, zaista budu prijavljeni zavodima za zdravstveno i mirovinsko osiguranje.

Postupkom registracije trgovačkog društva i obavljanjem njegova redovitog poslovanja, učenici primijenjuju porezne propise i propise o plaćanju obveznih doprinosa kroz Porezni kalendar vježbeničkih tvrtki u kojem je poslovna godina prilagođena trajanju školske godine i za obračun poslovanja traje od 1. rujna do 31. svibnja.

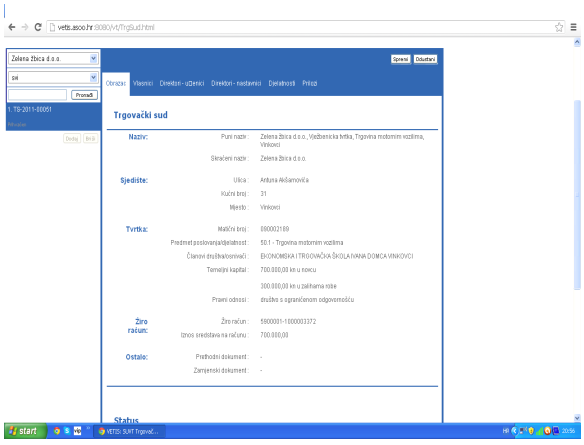
Korištenjem usluga Središnjeg registra osiguranika REGOS učenici mogu izabrati, prijaviti i promijeniti obvezni mirovinski fond, pratiti registar podataka o doprinosima za obvezna mirovinska osiguranja po osiguranicima te obraditi ponuđene obrasce. Uspješna prijava na REGOS omogućava praćenje mirovinskog osiguranja osiguranika, zaposlenika tvrtke i kapitaliziranje štednje buduće mirovine.

C. Korištenje SUVT-a za registraciju vježbeničke tvrtke

Usluge što ih pruža SUVT određene su postojećim propisima i dokumentacijom i zbog toga otvaranje nove tvrtke prolazi kroz korake točno određenim redoslijedom od kojih će se detaljno objasniti postupak registracije tvrtke na Trgovačkom sudu.

Za predaju zahtjeva za otvaranje tvrtke Trgovačkom sudu potrebno je odabrati aplikaciju SUVT Trgovački sud u padajućem izborniku na uvodnoj stranici.

U predloženom obrascu potrebno je popuniti polja: puni naziv tvrtke, skraćeni naziv tvrtke, ulica, kućni broj, mjesto, predmet poslovanja/djelatnost, članovi Društva/osnivači, temeljni kapital, pravni odnosi (d.o.o.), iznos sredstava na računu. Primjer tvrtke upisane u registar prikazan je na slici 3.



Slika 3. Registrirana tvrtka Zelena žabica d.o.o.

Nakon unosa ovih podataka, popunjavaju se ostale kartice u zaglavlju dokumenta, kao što su kartice Vlasnici, Direktori-učenici i Direktori-nastavnici. Ove kartice moguće je popuniti ako je škola prethodno pripremila bazu podataka u VETIS-u pa tada sustav automatski preuzima podatke i predlaže ih u padajući izbornik.

Kartica koja se odnosi na djelatnost koju vježbenička tvrtka obavlja nosi naziv Djelatnost. Za svaki novi unos koristi se aplikacija Dodaj te se otvara mogućnost izbora različitih djelatnosti u padajućem izborniku.

U kartici Prilozi potrebno je popisati priloge koji se predaju Trgovačkom sudu: izjava o osnivanju, potvrda o uplati osnivačkog pologa, izvješće revizora, popis članova osnivača, odluka o imenovanju člana Uprave - direktora Društva, popis osoba ovlaštenih za zastupanje, ovjera potpisa člana Uprave Društva i odluka o poslovnoj adresi.

Nakon što je tvrtka upisana u Registar poslovnih subjekata, ovlaštena osoba tvrtke popunjava osnovne podatke o tvrtki koji se ne popunjavaju u obrascima za registraciju na Trgovačkom sudu. Zatim jedna od ovlaštenih osoba treba prijaviti tvrtku u Registar Zavoda za statistiku, Registar obveznika PDV-a i Registar HZMO-a.

IV. ZAKLJUČAK

Obrazovni sustav ima glavnu ulogu u razvoju i oblikovanju stavova i vještina mladih ljudi koji su spremni pokrenuti i razvijati vlastite poduzetničke pothvate i sudjelovati u razvoju poduzetničke kulture.

Suvremena nastava predmeta vježbenička tvrtka naglašava važnost samostalnog učenja te zahtijeva aktivnu ulogu učenika u procesu obrazovanja.

Poslovanje vježbeničke tvrtke pomoću informatičkog modula SUVT ima za svrhu stjecanje poduzetničkih kompetencija i informatičke pismenosti koje su u skladu s trenutnim zahtjevima suvremenog tržišta rada.

LITERATURA

[1] Miyamoto K., "Human Capital Formation And Foreign Direct Investment In Developing Countries", OECD Technical Papers, No. 211. Paris, 2003.

[2] Key competences for lifelong learning, Recommendation of the European Parliament and of the Council of 18 December for lifelong learning (2006/962/EC), preuzeto sa: http://europa.eu/legislation_summaries/education_training_youth/lifelong_learning/c11090_en.htm (16.02.2013.)

[3] http://eacea.ec.europa.eu/education/eurydice (16.02.2013.)

[4] B. Baranović, M. Štibrić i V. Domović, "Obrazovanje za poduzetnost – perspektiva osnovnoškolskih učitelja i nastavnika" Sociologija i prostor, Vol.45 No.3/4 , 2008, str.339-360.

[5] http://www.insme.org/files/805 (14.02.2013.)

[6] http://www.asoo.hr/UserDocsImages/Dokumenti/Vjezbenicka%20tvrtka-web.kon.pdf (10.02.2013.)

[7] http://www.gencat.cat/salut/ccfcps/html/ca/dir3612/docs/thefuturelearning.pdf (10.02.2013.)

CTS

International Conference on

COMPUTERS IN TECHNICAL SYSTEMS

Steering Committee

Chairs: Vlado Sruk, University of Zagreb, Croatia
Leonardo Jelenković, University of Zagreb, Croatia
Dejan Škvorc, University of Zagreb, Croatia
Jadranko Novak, Nomen, Rijeka, Croatia

Members: Matjaž Colnarič, University of Maribor, Slovenia
Željko Hocenski, University of Osijek, Croatia
Niko Guid, University of Maribor, Slovenia

INVITED PAPER

New Wave of FPGA Prototyping for System-on-Chip Designs

Prof.dr.sc. Hrvoje Mlinarić
Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
hrvoje.mlinaric@fer.hr

System-on-Chip Designs using a processor inside an FPGA has many advantages. Specific peripherals can be chosen based on the application, with unique user-designed peripherals being easily attached. A variety of memory controllers enhance the FPGA embedded processor systems interface capabilities. FPGA embedded processors use general-purpose FPGA logic to construct internal memory, processor busses, internal peripherals, and external peripheral controllers (including external memory controllers). As more pieces (busses, memory, memory controllers, peripherals, and peripheral controllers) are added to the embedded processor system, the system becomes increasingly more powerful and useful. However, these additions reduce performance and increase the embedded system cost, consuming FPGA resources. Likewise, large banks of external memory can be connected to the FPGA and accessed by the embedded processor system using included memory controllers. Unfortunately, the latency to access this external memory can have a significant, negative impact on performance.

The Field Programmable Gate Array (FPGA) is a general-purpose device filled with digital logic building blocks. The two market leaders in the FPGA industry, Altera and Xilinx, are the focus of this presentation. The most primitive FPGA building block is called either a Logic Cell (LC) by Xilinx or a Logic Element (LE) by Altera. In either case, this building block consists of a look-up table (LUT) for logical functions and a flip-flop for storage. In addition to the LC/LE block, FPGAs also contain memory, clock management, input/output (I/O), and multiplication blocks. For the purposes of this study, LC/LE consumption is used in determining system cost.

Both Xilinx and Altera produce FPGA families that embed a physical processor core into the FPGA silicon. A processor built from dedicated silicon is referred to as a hard core processor. Such is the case for the ARM922T inside the Altera Excalibur family, the PowerPC 405 inside the Xilinx Virtex-II Pro and Virtex-4 families and the ARM dual cortex A9MP inside Zynx family.

A soft processor is built using the FPGA general-purpose logic. The soft processor is typically described in a Hardware Description Language (HDL) or netlist. Unlike the hard core processor, a soft core processor must be synthesized and fit into the FPGA fabric. In both soft and hard core processor systems, the local memory, processor busses, internal peripherals, peripheral controllers, and memory controllers must be built from the FPGA s general-purpose logic.

An FPGA embedded processor system offers many exceptional advantages compared to typical microprocessors including customization, obsolescence mitigation, component and cost reduction and hardware acceleration.

The FPGA embedded processor system is not without disadvantages. Unlike an off-the-shelf processor, the hardware platform for the FPGA embedded processor must be designed. The embedded designer becomes the hardware processor system designer when an FPGA solution is selected. Because of the integration of the hardware and software platform design, the design tools are more complex. The increased tool complexity and design methodology requires more attention from the embedded designer. Since FPGA embedded processor software design is relatively new compared to software design for standard processors, the software design tools are likewise relatively immature, although workable. Significant progress in this area has been made by both Altera and Xilinx. Within the next year, this disadvantage should be further diminished, if not eliminated. Device cost is another aspect to consider. If a standard, off-the-shelf processor can do the job, that processor will be less expensive in a head-to-head comparison with the FPGA capable of an equivalent processor design. However, if a large FPGA is already in the system, consuming unused gates or a hard processor in the FPGA essentially makes the embedded processor system cost inconsequential.

In this presentation we will show the trend in prototyping for System-on-Chip Designs using FPGA integrated with processor. The flexible nature of programmable logic and its tight integration to the processor based system offers designers the possibility to add virtually any peripheral they want and create accelerators to extend the performance of standards processor systems.

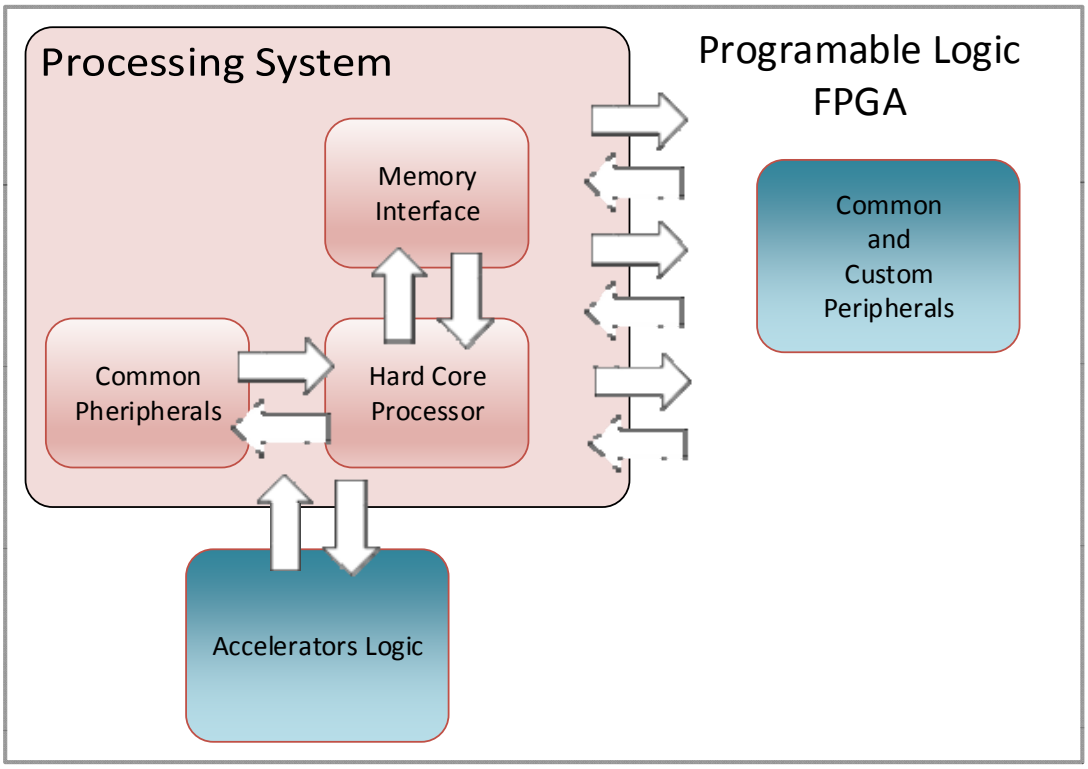


Figure 1. Embedding a processor inside an FPGA

PAPERS

Model-based Verification of the SIP Invite Scenario

Siniša Matetić and Marina Bagić Babac
Faculty of Electrical Engineering and Computing, Zagreb, Croatia
sinisa.matetic@fer.hr, marina.bagic@fer.hr

Abstract – The Session Initiation Protocol (SIP) is a signaling protocol for multimedia control over the Internet. It is used for initiating, maintaining and terminating multimedia sessions. In this paper we have carried out the verification of the SIP Invite scenario with Coloured Petri nets (CP-nets or CPNs). First, we have specified and verified the protocol model using a simulation tool. Then, we have verified the scenario using a state space analysis tool. We have concluded that the SIP Invite transaction is free of livelocks, and has only desirable deadlocks. Therefore, no additional remodeling for the protocol Invite transaction is needed.

I. INTRODUCTION

The Session Initiation Protocol (SIP) has been developed as a signaling protocol for creating, modifying, and terminating sessions with one or more participants [8]. These sessions include Internet telephone calls, multimedia distribution, conferences calls, and multi-player online games. SIP is a pure signaling protocol, that is, it is not concerned with the transport of the session's content. Instead, it carries meta-information that describes the media content of the session, which is then used to set up the communication channels. During the establishment phase of a session, communication between the entities is carried out through proxy servers that perform important tasks such as locating the communication partners, forwarding of calls, creating billing information, and so on [10].

Since a well-defined and verified protocol specification can reduce the cost for its implementation and maintenance, modeling and analysis are important steps of the protocol development life-cycle from the point view of protocol engineering [6]. Therefore, the need for formal specification and verification of the SIP is identified to increase the correctness, unambiguity, and clarity.

In this paper we have verified an example of SIP Invite scenario using Coloured Petri nets (CP-nets or CPNs). CP-nets provide a framework for the design, specification, validation, and verification of systems [13]. CP-nets have a wide range of application areas and many CPN projects have been carried out in industry, e.g., in the areas of communication protocols, operating systems, hardware designs, embedded systems, software system designs, and business process re-engineering [11, 14]. We have chosen CP-nets as a graphical language for specification and verification not only due to many qualities of CP-nets, but also because of their strong tool support by CPN Tools [5]. The tool is maintained and extended with new releases regularly. CPN Tools provides us with the editor, simulator and state space analysis tools, and many more features [1].

There has been some work by others on formal verification of SIP protocol in recent years. CP-nets have been used for verification of the INVITE transaction when the medium is unreliable in [6, 7] and reliable in [9]. Our model uses reliable communication medium. While papers [6, 7, 9, 12] have no detailed explanation of the particular Coloured Petri net model, we have made a test project to investigate properties and complexities of using CP-nets in formal verification. We put emphasis on the details of CPN modeling, so our paper can be used as an example of the complexities investigation in formal verification of systems. Also, each of the related papers [6, 7, 9, 12] provides its own idea of the places and transitions of the CP-net, so the models differ but the verification results principally match.

The paper is organized as follows. In Section II basic principles of SIP protocol and an example of message exchange are given. Section III presents CPN modeling of the high abstraction overview of our SIP Invite scenario, while Section IV describes details of the CPN model at the lower abstraction level. Section V. gives the CPN model analysis. Section VI concludes the paper.

II. SIP INVITE SCENARIO SETUP

To model the message exchange scenario of SIP via CPN we first have to decide about the complexity of the scenario itself.

Figure 1. Example SIP message exchange

This work was carried out within research project 036-0362027-1640 "Knowledge-based network and service management", supported by the Ministry of Science, Education and Sports of the Republic of Croatia.

The main stakeholders in the initiation of the SIP session are the caller, SIP proxy and callee. Caller is the only one that can send the INVITE message to initiate the session, but both the caller and the callee have the power of terminating the session by sending the BYE message.

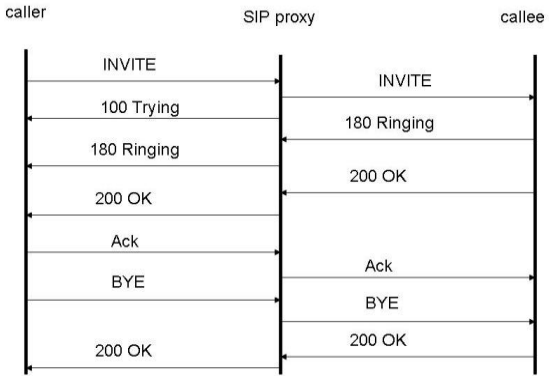


Figure 2. Message sequence chart for the scenario setup of SIP Invite transaction

For example, to invite Bob to a call, Alice could send the message from Figure 1 to her SIP proxy server. The structure of this message is typical for a SIP request message [10].

The first line identifies the request kind, the SIP destination of the request, and the SIP version used. It is followed by a sequence of headers that, for example, identify the SIP proxy at which the addressee is known to be reachable (Via). Then, there is the specification of a limit on the number of times the messages may be forwarded between proxies (Max-Forwards), the originator of the request (From), and additional information about the payload of the message (Content-Type) with its length (Content-Length) [8].

In the given example, the payload contains the initial parameters for the voice call. This is one possible payload that may be carried by a SIP message: media session parameters specified in the Session Description Protocol.

After this request has finally reached Bob and Bob has agreed to establish the session with Alice, Bob might send the response back to Alice via his proxy. This response consists of the SIP version followed by the response code (200 OK), followed by a sequence of headers. The Via headers specify the sequence of proxy servers that the message shall take on its way back to Alice. The From and To headers are repeated from the request, and again a description of the payload type and its size in octets are specified. In the example, the payload would contain a description of the kind of session that Bob is willing to participate in [10].

In this paper we have verified the basic message exchange for the SIP Invite transaction as shown in Figure 2. For the sake of simplicity, we have omitted some of the message headers. We have also assumed that the communication medium is reliable.

In this particular case from the Figure 2, after the establishment of the session, the caller terminates the session and the callee confirms the termination by sending a status message.

III. MODELING THE MESSAGE EXCHANGE SCENARIO

CPN allows modeling the network in multiple hierarchies for more complex nets. Even the simplest protocol that can be modeled by CPN can result in numerous states and transitions which make the message flow hard to track. Since the proposed scenario of the SIP protocol can be easily divided into three parts: Sender, Proxies and Receiver, it is best to make a higher abstraction of the CPN by showing only these three modules and the communication channels between them.

In Figure 3 there are three main modules (SIP-SENDER, TRANSMISSION, SIP-RECEIVER) connected through four states (A, B, C, D). State A represents the output from the SIP-SENDER module where the message requests from the caller arrive. Upon arrival of any messages state A represents the input to the TRANSMISSION module.

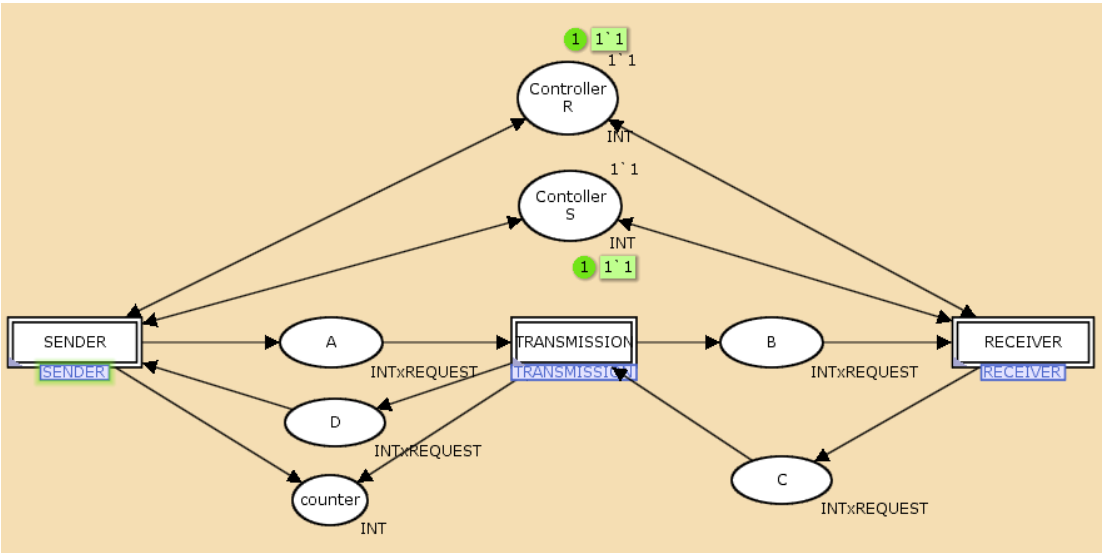


Figure 3. CPN model of the high level abstraction overview for the SIP Invite scenario

TRANSMISSION module is responsible for sending that request further to the SIP-RECEIVER via state B which represents the output state of the TRANSMISSION module and in the same time the input state for receiving requests intended for the SIP-RECEIVER. We can say that states A and B represent a one way channel for transmitting messages from SIP-SENDER. In the same way, states C and D represent a one way channel from the SIP-RECEIVER, either for sending responses to incoming requests or for initiating its own requests (e.g. BYE message in our SIP CPN example scenario).

In Figure 3 there are also three additional states: counter, Controller R, Controller S. Their meaning is not intuitive as they are not directly related to the SIP message flow which our CPN models. Additional states were required to fulfill the exact running sequence of messages in the SIP protocol.

Since the states A, B, C, D, counter, Controller R and Controller S are not parts of the internal structure of modules Sender, Receiver and Transmission, they are tagged with In, Out or I/O notes in all of the following figures.

Namely, the SIP CPN model is designed to withstand only one message per request, without the possibility that the message will be lost (as messages are carried over the reliable communication medium). After the request or response is sent from one party (e.g. caller), the party awaits the confirmation message or the next expected message defined by the SIP protocol. Until then, the party is blocked and cannot send any additional messages, nor repeat the previously sent ones.

Counter state regulates the message order from the SIP proxies (TRANSMISSION module) to the initiator of the SIP session. This is particularly needed when the SIP proxy receives the INVITE request. Consequently, the proxy has to send the "100 Trying" message back to the initiator before forwarding the request further to the receiver. After the 100 Trying is sent, the counter state enables the SIP proxy to continue with the message transmission.

Controller S and Controller R are also states for regulating the message order in the SIP-SENDER and SIP-RECEIVER. They represent I/O states which block the caller and the callee after one message is sent.

For example, after the INVITE message has been sent Controller S regulates that the caller can now expect the "180 Ringing" message and cannot send nor receive anything before the arrival of that message. Opposite to that, Controller R regulates that the callee has to send "180 Ringing" message followed by "200 OK" message upon receiving the INVITE request.

IV. DETAILED SIP CPN DESCRIPTION

The in-depth picture of each module from our CPN presented in high level abstraction overview (Figure 3) will be described in detail in further subsections. After the insight into the CPN message model and variables in the next subsection, we explain the idea under the TRANSMISSION module followed by the SIP-SENDER and SIP-RECEIVER.

A. CPN message model and variables

The main colset (shortcut for the "colour set", i.e. a data type in CPN modeling) used in our CPN model is a simplified SIP protocol message consisted of a sequence number representing the order number for requests and status code for responses (e.g. 200 for an OK message). Record type from CPN tools is used to form this complex message consisted from an integer and a string. Other fields that are available in regular SIP messages are not used in the model since the model recognizes only one Sender and one Receiver. It is obvious that the SIP messages initiated from the Sender have to go through the available proxy and are intended for the only available Receiver. For message transmission a product between the previously mentioned record and an integer is used. Basically, all the message regulations go through that additional integer. Semantically we could have used the sequence number from the record type, but by presenting an additional number we have divided the meaning of the "OK" messages available in the protocol flow. For example, "200 OK" message sent as a response for the INVITE request is characterized with integer 200, but the "200 OK" message sent as a response to the BYE request sent from the Sender is characterized by integer 210. The division made it simpler to understand and analyze the message flow when running the simulation. Additionally, the CPN contains numerous control variables with predefined values. These are used for protocol message order regulation.

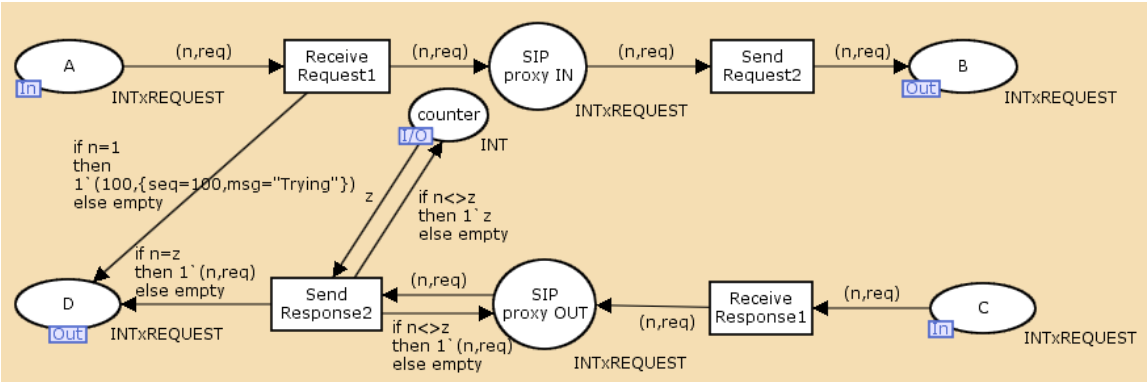


Figure 4. CPN model of the TRANSMISSION module

All of the variables have assigned numbers that are chosen randomly to suit the idea of the authors. Note that none of them follow any status code of messages or control variables found in the literature.

B. Transmission module

In Figure 4 the TRANSMISSION module is shown with its inner states and transitions. As described earlier it is consisted of two input states (A for inbound requests/responses from SIP-SENDER, C for inbound requests/responses from SIP-RECEIVER) and two output states (B for outbound requests/responses to SIP-RECEIVER, D for outbound requests/responses to SIP-SENDER). Additionally we see the counter state for regulating the order of incoming responses after the first INVITE request and SIP proxy IN, SIP proxy OUT states that simulate the queue for messages held in the SIP proxy regarding the input or output channel.

After receiving the first message in state A, the "Receive Request1" transition becomes available for execution. It automatically checks if the message in A is an INVITE request and if it is "100 Trying" message is sent to state D which represents the response channel. Also, the INVITE message is forwarded to the SIP proxy IN state for further transitions. Consequently, transition "Send Request2" becomes available for execution. It only forwards the message from the queue to the output state B making it available for the SIP-RECEIVER. We can see that there is no regulation of the message order defined by the server for the SIP initiation; proxy awaits the messages and just forwards them not worrying about their contents (expect the INVITE – 100 Trying case).

Upon arrival of the message in the state C, transition "Receive Response1" can be executed which transfers the message to the SIP proxy OUT queue. Since the response messages from SIP-RECEIVER are "180 Ringing" and "200 OK" to confirm the INVITE request we have to regulate their delivery order to the SIP-SENDER. Transition "Send Response2" checks from the counter its current state which is defined by the SIP-SENDER module. If the message sequence number is the same as the one in the queue, counter state emptied is and message is forwarded to the state D. If not, counter state remains the same and no messages are forwarded. Since the counter is now empty, new forwarding of the messages in queue can be done after the SIP-SENDER receives the previous messages and sets the counter state for the next message. State D becomes empty after the SIP-SENDER receives "180 Ringing" message and the path for "200 OK" is now open. The counter is afterward set to expect either a BYE request from SIP-RECEIVER or a "200 OK" as a response to the SIP-SENDERS BYE request.

C. Sender module

The SIP-SENDER module is shown in Figure 5. The main state in the module is the Caller. It contains all the possible messages that can be sent from the SIP session initiator. Also the module contains one input state D for receiving requests/responses from the SIP-RECEIVER, two output states A and a counter for transmitting its own requests/responses for SIP-RECEIVER and regulation of the receiving order.

Controller R and Controller S are two I/O states for message regulations where Controller R is only used for announcing the initiation of BYE request to the SIP-RECEIVER while Controller S regulates the order of sending and receiving messages from both inbound and outbound channel.

At the beginning, transition "Send Request1" is available for execution. Controller S states the message that can be used from the Caller's available messages. The message is characterized by 1'1 initial state in the Controller S which corresponds to the INVITE request from the Caller.

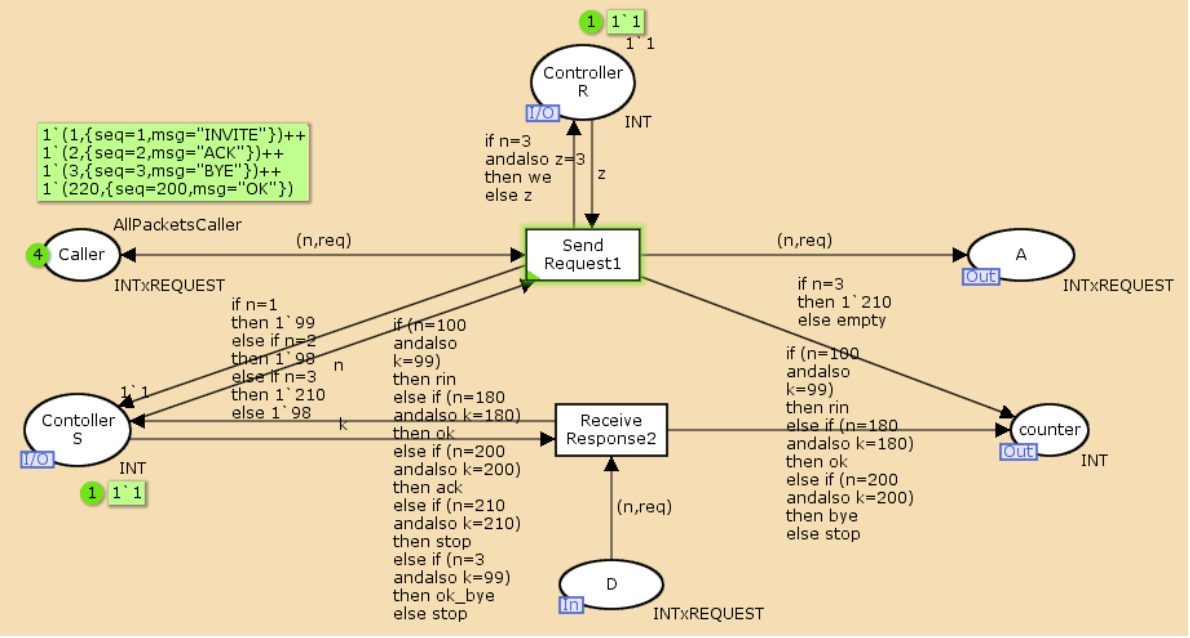


Figure 5. CPN model of the SENDER module

After the execution of the transmission, Controller S's state is changed to 1'99 which blocks any additional execution of that transmission. SIP-SENDER module now awaits a message in the input D state.

Upon receiving "100 Trying" from the SIP proxy and after executing "Receive Response2" transition, Controller S changes its state to enable the "180 Ringing" message to be received. Consequently, after receiving "180 Ringing", Controller S is regulated to enable the final "200 OK" message.

After receiving final response from the SIP-RECEIVER, Controller S enables the Caller and transition "Send Request1" to send an ACK message to SIP-RECEIVER. As the message goes to the outbound channel and state A, Controller S awaits a successful arrival of the ACK to the SIP-RECEIVER. SIP-RECEIVER will inform Controller S after the arrival and then the Controller enables the Caller to initiate BYE messages for terminating the session.

Crucial moment in the CPN model is when one of the parties either the SIP-SENDER or the SIP-RECEIVER decides to terminate the session. Immediately after the message has been sent, other party is blocked by its controller and cannot send the BYE message by itself, yet only wait for the message. Upon receiving it, it sends a "200 OK" stating that everything went right. Note that this is a simplification to the CPN design since it is very common that both parties can initiate termination at the same time or with small time differences when none of them received the other party's request for termination. This feature is not important for protocol analysis and was then easily controlled by regulator states.

D. *Receiver module*

The SIP-RECEIVER module's states and transitions are shown in Figure 6. It contains one input state B for receiving requests/responses from SIP-SENDER, one output state C for sending requests/responses to the SIP-SENDER and two I/O states: Controller R and Controller S used for regulating message orders in the same way as in SIP-RECEIVER. Also, the module has one main state which contains all the available messages that can be sent from the Callee. The Callee state contains different set of available messages since it cannot initiate the session, yet just respond to the initiation performed by SIP-SENDER.

At the beginning of the simulation SIP-RECEIVER is blocked by Controller R and cannot send and messages. Transition "Receive Request2" awaits the message in state B to be received so it transition process can be executed. It is important for the first message to be the INVITE request characterized by sequence integer "1". After executing the "Receive Request2" transition, Controller R's state is changed from 1'1 to 1'180 which enables the execution of "Send Response1" transition. Callee state provides the "180 Trying" message which is transferred to the output state C.

Consequently, Controller R's state is changed to 1'200 enabling the "Send Response1" transition to execute the sending of the next "200 OK" message from the Callee.

Upon that response, the SIP-RECEIVER is blocked for further sending until the ACK message has been received from the SIP-SENDER. ACK message enables both the SIP-SENDER and SIP-RECEIVER for transmitting the BYE message and the situation is the same as described in the case of SIP-SENDER.

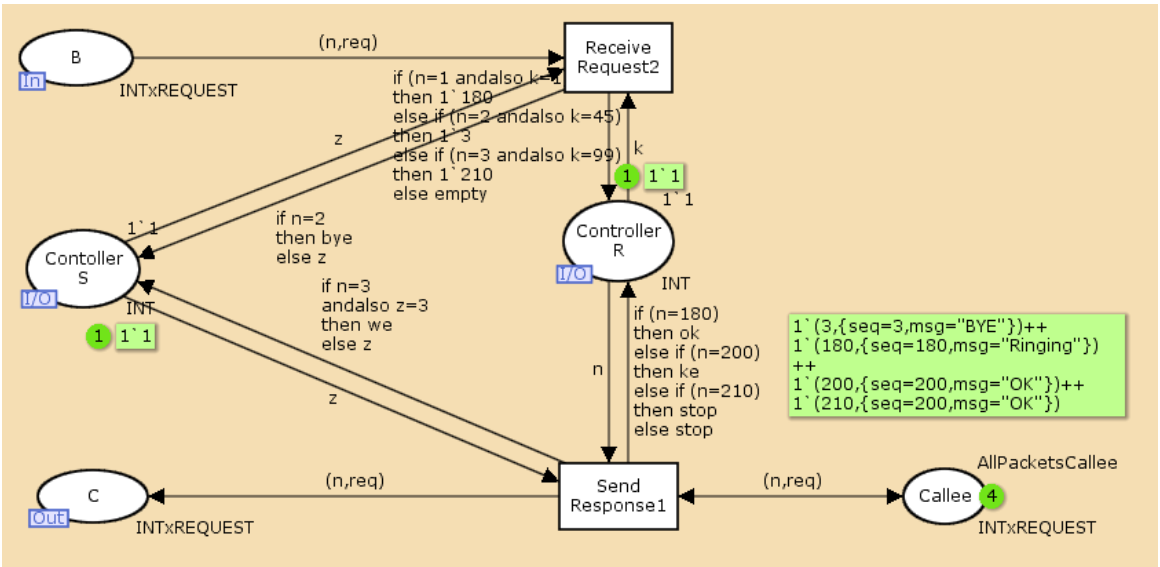


Figure 6. CPN model of the RECEIVER module

V. CPN STATE SPACE ANALYSIS

CPN Tools [5] provides us with the state space report when running state space analysis tool [4]. The first part of the state space analysis report contains statistical information about the size of the state space. We see that the state space has 48 nodes and 66 arcs. From the information it is clear that this is the full state space of our CPN. Also, the statistical information available contains information about the SCC-graph of the state space. The information represents the number of strongly connected components and the number of arcs that start in one component and end in another. A strongly connected component is a maximal subgraph in which it is possible

to find a path from any node to any other node [1]. Strongly connected components are very useful to determine certain kinds of behavioral properties, and they can be calculated by standard algorithms (which are linear in time and space) [3]. From the report we see that the number of nodes is also 48 and number of arcs is slightly less than state space statistics, 63. Reason for this similarity between the number of nodes and arcs in both cases is in the fact that our CPN SIP model is very restricted. Number of unwanted transitions is regulated by Controller states and there are no repeated messages.

The second part of the state space report contains information about integer and multi-set bounds. These values represent the maximal and minimal number of tokens which the individual places may have. We see that each of the places A, B, D which represent the channel always has between zero and one token. The exception is the state C which can have two tokens (part of the simulation where the messages "180 Ringing" and "200 OK" are sent quickly one after another from the SIP-RECEIVER). Places Callee and Caller always have 4 tokens; the tokens do not go away after a message has been sent to the other party. Counter place, as place C, has between zero and two tokens for the same reason regarding two response messages from SIP-RECEIVER that go after each other. Finally, we see the Controller places have maximum 1 token and minimum zero tokens, with the exception of Controller S where minimum is set to 1 allowing the simulation to start (initial sequence number for INVITE message).

The third important part is the liveness properties of the state space. We see that the state space has two dead markings where the simulation can end. Nodes 47 and 48 present those two final markings. In node 32 the decision is made which leads to two different endings and it depends on the first sender of the BYE message. If SIP-SENDER initiates the termination of the session we end up in the marking number 47 while termination by SIP-RECEIVER leads us to marking 48. Also, from the report we see that our CPN has no dead transition instances, i.e. the simulation always ends in the two final markings without stopping because of any errors during simulations.

VI. CONCLUSION

In this paper we have modeled and verified an example of the Session Initiation Protocol Invite Scenario. We have used graphical modeling language Coloured Petri nets for the protocol specification, simulation and state space analysis. The very detailed description of the protocol model is given at both higher and lower levels of abstraction. At the higher level, there are three modules (Sender, Transmission and Receiver) simulating our example of the SIP Invite scenario, while at the lower level their internal places and transitions were given and explained. Therefore, our example is also useful as a tutorial example for the in-depth understanding of CPN modeling and simulation. Both simulation and state space analysis have shown the correctness of the protocol model, where only desirable deadlocks were present being our final markings of the protocol.

REFERENCES

[1] K. Jensen and L. M. Kristensen, "Coloured Petri Nets", Springer-Verlag Berlin Heidelberg, 2009.

[2] L. M. Kristensen, S. Christensen, and K. Jensen, "The practitioner's guide to coloured Petri nets", International Journal on Software Tools for Technology Transfer, pp. 98-132, vol. 2, 1998.

[3] K. Jensen, "An Introduction to the Practical Use of Coloured Petri Nets", Lectures on Petri Nets II: Applications, Lecture Notes in Computer Science vol. 1492, Springer-Verlag, pp. 237-292, 1998.

[4] K. Jensen, "Coloured Petri Nets. Basic Concepts, Analysis Methods and Practical Use", Volume 3, Practical Use. Monographs in Theoretical Computer Science, Springer-Verlag, ISBN: 3-540-62867-3, 1997.

[5] CPN Tools, http://cpntools.org/, CPN Tools homepage, 2013.

[6] S. Baraković, D. Jevtić, and J. Baraković-Husić, "Modeling of Session Initiation Protocol INVITE Transaction Using Colored Petri Nets", International Journal of Computer and Communication Engineering, vol. 6, pp. 22-29, 2012.

[7] L. Liu, "Verification of the SIP Transaction Using Coloured Petri Nets," in *Proc. of the 32nd Australasian Computer Science Conference*, Wellington, New Zealand, January 2009.

[8] J. Rosenberg, et al., "SIP: Session Initiation Protocol," Technical Report RFC 3261, Internet Engineering Task Force (IETF), June 2002.

[9] L. G. Ding and L. Liu, "Modelling and Analysis of the INVITE Transaction of the Session Initiation Protocol Using Coloured Petri Nets", In K.M. van Hee, Rüdiger Valk (eds.): Applications and Theory of Petri Nets. Proceedings of the 29th International Conference, PETRI NETS 2008, Xi'an, China, June 23-27, 2008. Lecture Notes in Computer Science. Vol. 5062, Springer-Verlag, pp. 132-151, 2008.

[10] C. Willcock, S. Tobies, T. Deiss, S. Keil, F. Engler, and S.Schulz, "An Introduction to TTCN-3", John Wiley \& Sons Ltd, 2005.

[11] V. Gehlot and A. Hayrapetyan, "A CPN Model of a SIP-Based Dynamic Discovery Protocol for Webservices in a Mobile Environment", In K. Jensen (ed.): Proceedings of the Seventh Worskhop and Tutorial on Practical Use of Coloured Petri Nets and the CPN Tools, Department of Computer Science, University of Aarhus, PB-579, pp. 197-216, October 2006.

[12] G.Vijay, Colored Petri Net model of the Session Initiation Protocol (SIP), IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society, pp. 2150-2155, 2010.

[13] L.M. Kristensen, J.B. Jřrgensen, and K. Jensen, "Application of Coloured Petri Nets in System Development", Lectures on Concurrency and Petri Nets - Advanced in Petri Nets. Proc. of 4th Advanced Course on Petri Nets. Vol. 3098 of Lecture Notes in Computer Science, pp. 626-685, Springer-Verlag, 2004.

[14] L.M. Kristensen and K. Jensen, "Specification and Validation of an Edge Router Discovery Protocols for Mobile Ad Hoc Networks", In Integration of Software Specification Techniques for Applications in Engineering. Vol 3147 of Springer Lecture Notes in Computer Science, pp. 248-269, Springer-Verlag, 2004.

Improved Automatic DTV/STB Menu Navigation for Functional Verification

V. Marinković*, B. Kordić*, M. Popović* and V. Peković**
* Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
** RT-RK Computer Based Systems LLC, Novi Sad, Serbia
vladimir.marinkovic@rt-rk.com

Abstract – Menu navigation is the basic functionality of DTV/STB devices. This paper presents improved automatic menu navigation and integration of tools such as MaTeLo in the Black Box Testing (BBT) system for automated functional verification of DTV/STB devices. Specifically, the method for accurate determination of the source of an error in the case of failing test is given. In addition, the mechanism of selecting either recovering from error in the case of test failure or restoring the device to its initial state, as well as the mechanisms for recovering or returning to its initial state are proposed.

I. INTRODUCTION

Devices for home multimedia entertainment, such as high definition television (HDTV), enable the high quality transmission of video content using traditional transmission methods (cable, terrestrial or satellite) and additionally enable transmission over IP networks [1]. Initially, for reproducing a digital audio and video content to the end user Set-Top Box (STB) devices were used rather than digital TV (DTV) [2]. At the time, the basic functionality of STB devices was demodulation and decoding of digital video broadcast (DVB) in the audio-visual content that is shown on an analog TV which is practically used as a monitor, or could be stored to some external storage device.

Further development of STB devices have introduced new functionalities related to new interactive and noninteractive services and applications [3], and consequently the complexity of the STB devices architecture was increased. With increased complexity of STB devices, functional testing became more difficult. Since the tests are performed several times during the development of the device, and the time to market has to be reduced to a minimum, fast and efficient functional verification has become imperative. Therefore, the manual testing as a slow, time consuming, and prone to errors process is unacceptable, and automated functional verification comes as a natural choice, and is crucial for producers of STB devices [4,5]. In order to achieve greater efficiency and reliability, automatic approach can significantly reduce the time and the cost of testing. However, the implementation of automated testing brings with it the difficulty of automatic generation of test cases, automated test execution on a STB device under test, and automatic data collection and analysis.

Previously proposed efficient Black Box Testing (BBT) test system for automated functional verification of STB and digital TV devices, which considered the device under test as a black box with known outputs expected for the corresponding inputs [5]. The proposed system has the ability to control both devices under test and other devices that make up the testing environment, placing the device in the desired state.

As the menus, one of the major features of STB devices, become more complicated and used in various applications, they also become one of the most important functional items to be tested. It is important to clarify that testing menu navigation entails testing other parts and functionalities of STB as well (applications, peripherals, routines, etc.). For example, when we test menu navigation it is often the case that we test the search TV channels menu level, which also tests functionality of searching channels. That means if STB's tuner is failing, theses test cases would also fail.

In the automatic functional verification of devices, for the functionality of menu navigation and other functionalities, it is necessary to automatically generate and execute tests. For automatic test case generation, various tools, like a MaTeLo tool for statistical testing of both software and hardware, can be used integrated in the BBT system for automatic tests execution. If usage model represents menu navigation system of a DUT, by using MaTeLo tool it is possible to automatically generate test cases which enable automated menu navigation once they are joined to BBT system. Then MaTeLo tool applies Markov chains logic to define an abstract model of the system which is to be tested. Based on an abstract model, in compliance with the selected test strategy and transition probabilities defined in the model, the tool automatically generates abstract tests cases. Abstract test case contains abstract test steps that are mapped to a particular action of the system under test. In accordance to the target test bed platform, abstract test cases are translated into a specific format (Python language in the case of BBT system) [6].

After test suite execution, using the BBT system all test results are collected and forwarded to MaTeLo tool for system reliability and mean time to failure (MTTF) measures calculation. In order to make these results accurate, it is necessary to determine the exact cause of the error in the test case execution (is the device under test really causing the error) and how many steps in the test case are actually performed unsuccessfully. In other words

it is necessary to provide the possibility of continuing the execution of the test case even after an unsuccessful execution of a step in the test case.

Thus, besides the implementation of the system for verification, it is necessary to reliably determine the cause of the error, and if it is really the device under test, to register the error and to return the device under test to the desired, known state. Taking into account that for the verified functionality menu navigation is selected, restoring the device to a known state is a challenging task due to the complexity of menus and many special cases, which means that the solution cannot be easily generalized.

An approach for reliable determination of the cause of failing test case is proposed in this paper. Besides, it offers a way for selecting the mechanism for recovering the device from errors, i.e. for returning it to the initial state. These mechanisms are also described.

Section II describes automatic testing framework within which the study has been made. In Section III the mechanism for reliably determining the cause of the test case error is introduced, while the Section IV describes the proposal of efficient solutions for automated error-free menu navigation. The Section V depicts the system testing procedures, while the Section VI gives a brief conclusion by examining the advantages of this solution and disadvantages as the subjects of further development.

II. AUTOMATIC TESTING FRAMEWORK

Automatic functional verification of the DTV and STB devices [4,5] is usually done by BBT approach [7]. BBT approach considers the DTV/STB DUT as a black box what means its inner software/hardware architecture is not known. Only input and output interfaces and the conditional DTV/STB states in which it can be set into are known, according to the DTV/STB design specification. In order to enable automated testing, it is necessary to generate/stream the input signal, monitor its output, and to automatically control the DTV/STB under test. Based on the matching between the output of the DTV/STB under test and the expected output the decision is made concerning the functionality of the DTV/STB under test.

The BBT testing framework in which the study has been made (described in [5]), consists of: (1) PC workstation with RT-Executor application for controlling automated testing process and collecting test results, (2) DekTec card with RF and DVBS2 modulators as a signal generator, (3) RT-AV100 device used for image capturing [8], and (4) Remote controller emulator [9]. The testing system is shown in Fig. 1 with its typical application in production line with STBs under test.

III. DETERMINATION OF THE CAUSE OF AN ERROR

In order to calculate the reliability and mean time to failure of a device, it is essential to isolate the errors in test execution that are caused by that device only. As it is impossible to conclude whether the device causing the error is the device under test, the natural solution is to exclude the possibility that any other device included in testing environment is a potential cause of failing test.

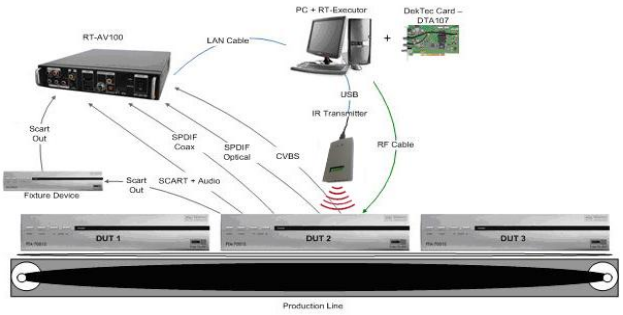


Figure 1. BBT system for automatic testing used as the framework.

Two possible approaches are isolated: (1) constant validation of a single device from testing environment, and (2) to trust a single device that it is working correctly based on its high reliability. The first approach is better, more confident, and safer, but it is not possible to carry out in every situation, because the simple way to validate some device during its operation is not always available. The second approach is less efficient because we cannot state that some device has not caused the error no matter how high reliability it has. In that case, the reliability of each device from testing environment has to be included in final calculation and of reliability and mean time to failure of the device under test. The reliability (if not already known) of each single device has to be determined by statistical functional testing of the same.

The first approach can be easily applied to the controllers included in test bed. As the first item, the controller of testing process already has the mechanism for checking if every command (test step from test case) is sent to the device driver and additionally for checking whether the device driver received the command and executed it successfully or not. In the case that some of these conditions are not met, it is considered that the error is caused by the controller of the testing process. The second most often used controller in testing process of DTV/STB devices is the remote controller emulator because remote controller is usually the only interface between user and the digital receiver. As it is most often infrared remote controller emulator, its reliability is not satisfactory due to used parts with small range, not very precise, and with small immunity to the noise introduced by similar infrared devices [9]. In order to reduce or even eliminate the possibility that the cause of failing test case is the remote controller emulator, it is possible to use two of these emulators – emulator A that sends commands to the DTV/STB under test, and emulator B which is positioned near the infrared receiver of the device under test and constantly receives commands sent from emulator A. Immediately after the test step referring to sending some command over the emulator A, it is checked whether the emulator B have received the same command. Emulator B must not receive neither more (means that some noise is present), nor less commands than actually sent from emulator A (means that command is probably not sent). Unrecognized commands are not acceptable as well. If any of listed criteria is not met, the remote controller emulator is considered as the cause of the error and it is not ascribed to the device under test. However, there is still a case of failing IR commands caused by bad IR receiver or decoder of a DTV/STB device. But this

case is not really a problem, in fact it is a failure of a DUT, thus it is natural to ascribe it to DUT.

Operation of signal generators used in the system is also easy to check. The output from it is connected to both video capturing device RT-AV100 and the DUT, using a signal splitting device. That way, it is possible to check the output of the signal generator in one test step, every time it is necessary in process of testing some DUTs functionality (e.g. TV channel search).

For all other devices that are included in test bed, it is necessary to apply the second approach, i.e. their reliability has to be determined as precisely as possible. This applies to audio and video capturing device RT-AV100 [8], in the case of described testing framework.

IV. Automatic Navigation Through the Menu

In the case of successfully executed test case, the state that the device under test is expected to run into is known. Thus, after the execution of a test case it is checked if the device is in the expected state. It is done by capturing current image given on the output of the tested receiver and comparing it to the referent image afterwards. If these two images match with a certain tolerance, it is concluded that the test case is successfully executed and the device under test is returned to its initial state, or it is left in the current state (depends on the type of tests, i.e. whether the tests are interdependent or not). Thereby, the device is ready for the next test case. If two images do not match, the test case is considered as failed.

When the test case or even one of its steps is executed with failure, it is necessary to start the certain process of recovering from the failure. The recovery procedure comes down to running device into some known state. When setting the device in some known state there are two possible scenarios. The first cause of failing test is running device into incorrect state, but the device is still functioning well, and the second one is the device stuck in some state. In the first case, it is possible to force the device to the initial state or to try once again to make it go into the state it should have been entered once the failing test case is executed (depends of test cases, i.e. whether one test case relies on the previous one or they are completely independent). In the second case, the device has to be completely restarted, set to factory default settings, and set to the initial state if needed. It seems like the second method could be used in both cases, what is correct, but restarting the device, setting it to the factory defaults, and setting it to the initial state takes a long time (up to five minutes) extending the test execution time. For this reason, it is preferred to apply the first method whenever it is possible.

In order to force the device to go into some known state, it is obligatory to check in which state the device currently is (which item in menu is chosen), and to carry out the actions for the navigation to the desired state. V. Zlokolica, N. Milutinovic, V. Marinkovic, V. Peković, and J. Zloh have offered a solution for automatic error-free menu navigation of STB devices in [10]. In that solution they have used the idea of capturing the image from the device under test, reading the text from defined area of the image by Optical Character Recognition (OCR) algorithm, and finally comparing the text to the one that is joined to the each menu item. Based on the defined complete structure of the menu, it is possible to determine which actions have to be carried out in order to reach the desired state, i.e. menu item.

Similar, but significantly extended method is used in this solution. For the navigation through the menu, the remote controller emulator is used, while image capturing from device under test is done by the audio and video capturing device RT-AV100. The text is extracted from defined region of image captured from tested digital receiver, using the plug-in for the BBT system with OCR algorithm suited with captured image and the region of interest like shown in Fig. 2. Based on the extracted text, the unique hash key is calculated. Each menu item also has the hash key joined to it, what makes a menu item hash table. Searching this hash table it is possible to determine the current state of the device. Furthermore, each state is joined to the set of commands which can force the device from that state to the initial one in the case that such a scenario is needed. If one test case relies on a previous one, and consequently the test cases have to be executed sequentially, and that everything that should have been executed in a failing test case is necessary to be executed again, the navigation based on the menu model defined in MaTeLo tool is done. That way, the device is returned to the initial state of next test case. In both cases, after the navigation, it is checked once again if the navigation was successful. If the navigation fails once again, the device is considered to be blocked in some state and it has to be restarted.

For the particular example, the same approach form [10] is not used because the menu model is already defined using MaTeLo tool, thus it would be redundant to define it again. Nevertheless, this solution is more convenient for particular applications because the device is often returned to its initial state, so that the navigation from item A to item B is not always required.

As stated before, in the case of concluding that the device has been stuck after test execution, it is necessary to restart it. Restarting the device under test is done by the switch controlled from the PC [11], i.e. from the test execution controller. Setting the device to factory defaults and to initial state which is performed afterwards, is done by the remote controller emulator.

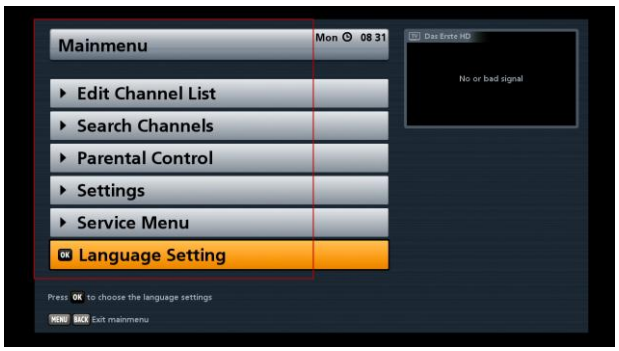


Figure 2. The example of captured image showing menu items and the region of insterest marked with the red rectangle.

V. TESTING AND RESULTS

In order to verify the proposed system, the particular STB from Kathrein is tested by the statistical testing methodology. That way, the reliability and mean time to failure of the tested STB is determined. Test bed is consisted of: Kathrein UFS 903si digital STB receiver as the device under test, audio and video grabber device RT-AV100, remote controller emulator, RT Executor – controlling application of automatic test execution, and DekTec card as a signal generator.

The device is tested using test cases generated by MaTeLo tool for statistical testing, based on the produced usage model. Overview of cumulative testing results is shown in Table I. Table I also gives comparison of testing results with and without improvements described in this paper.

In Table I it can be noticed that the number of test cases executed with failures during automatic functional verification is remarkably lower in the case when described improvements were used. Errors that are excluded after improvements were mainly the consequence of fault IR signals transmission, but some of them were introduced by signal generator.

In other words, the reliability and mean time to failure are significantly improved. There is still a problem of ascribing a certain amount of failures to the device under test even though it may not be the real cause of them. Including these failures in the calculation of final result as it is the only work around this issue, they still affect device under test reliability and mean time to failure measures. In order to reduce or minimize this error, it is necessary not only to determine the reliability of each device included in test bed, but to deal with suppressing these failures as well.

MaTeLo tool is used for calculation of reliability and mean time to failure without going into details of how it does it, what is the mathematics behind it, and with high confidence that it works properly. For obtaining the values of reliability and mean time to failure of device under test, it is necessary to define an internal parameter k, which indicates the number of activations (in this case the number of commands sent by a remote control device) in one hour. Given that this is a very sensitive parameter, it is determined by the experimental method on a larger sample and the value of 15 is taken.

TABLE I. TESTING RESULTS

Results (number of activations k = 15)		
	Without improvements	***With Improvements***
Number of executed test cases	2800	**2800**
Number of test cases executed with failures	24	**13**
Nuber of successfully executed test cases	2776	**2787**
Reliability	0.564613	0.924134
Mean time to failure [days]	1.716309	15.457343

VI. CONLUSION

In this paper, the efficient way for automatic menu navigation for digital TV/STB receiver is proposed. It can be opportunely used in automatic functional verification of digital receivers. Based on testing particular digital receiver from Kathrein, the process of automatic navigation through main menu and its submenus during statistical testing using MaTeLo tool and the system for automatic testing is described. Besides, the analysis of all potential problems and limitations introduced by the used system is given.

Applying the described solution, it is necessary to determine reliability of each device included in testing environment, and to include these values in the final calculation of the reliability of device under test, what will be the next step of the development of this system. It would be useful to connect usage model from MaTeLo tool to the mechanism for recovering from failing test steps as well. Based on that model, it would be possible to determine which activities are needed to be executed in order to force the device to run into the state it should have been entered if the test step was not failing.

ACKNOWLEDGMENT

This work was partially supported by the Ministry of Education and Science of the Republic of Serbia, Grants III 44009 and TR 32031, from year 2011.

REFERENCES

[1] Y. Wu, S. J. Hirakawa, and U. H. Reimers, “Overview of digital television development worldwide”, Proceedings of the IEEE, Vol. 94, No.1, Jan. 2006, pp.8 - 21.

[2] S. Pekowsky and R. Jaeger, “The set-top box as multimedia terminal”, IEEE Transactions on Consumer Electronics, Vol. 44, No. 3, Aug. 1998, pp. 833 - 840.

[3] F. Lonczewski, R. Jaeger, “An extensible set-top-box architecture for interactive and broadcast services offering sophisticated user guidance”, IEEE International Conference on Multimedia, New York, NY, USA, 2000.

[4] A. N. Rau, “Automated test system for digital TV receivers”, IEEE International Conference on Consumer Electronics, Los Angeles, CA, USA, 2000, pp. 228-229.

[5] D. Marijan, V. Zlokolica, N. Teslić, V. Peković, T. Teckan, “Automatic functional TV set failure detection system”, IEEE Transactions on Consumer Electronics, Vol. 56, No. 1, February 2010.

[6] A. Guiotto, B. Acquaroli, A. Martelli, “MaTeLo: automated testing suite for software validation”, Proceedings of DASIA 2003, Jun. 2003, Prague, Czech Republic.

[7] B. Beizer, "Black-box testing: techniques for functional testing of software and systems," *Wiley*, 1995.

[8] Network attached video capture device – NAVIC RT-AV100, http://www.rt-rk.com/audiovideo-capture-devices/navic.

[9] V. Marinkovic, I. Kastelan, N. Vranic, Z. Marceta, V. Pekovic, “Protocol independent remote controller emulator for automated testing of multiple DTV/STB”, TELFOR 2012, pp. 1064 - 1067.

[10] V. Zlokolica, N. Milutinovic, V. Marinkovic, V. Peković, J. Zloh, “Automatic set-top box menu navigation scheme for STB menu functional verification”, IEEE 19th International Conference and Workshops on Engineering of Computer Based Systems (ECBS), Novi Sad, Serbia, April 2012, pp. 60-63.

[11] V. Nuhijevic, S. Vukosavljev, B. Radin, N. Teslic, M. Vucelja, “An intelligent home networking system”, 2011 IEEE International Conference on Consumer Electronics - Berlin (ICCE-Berlin), Berlin, Germany, 2011, pp. 50-53.

Reliability Sensitivity Analysis for Yahoo! Pipes Mashups

G. Delac*, M. Silic* and K. Vladimir*
* University of Zagreb, Faculty of Electrical Engineering and Computing,
Consumer Computing Laboratory, Zagreb, Croatia
{goran.delac, marin.silic, klemo.vladimir}@fer.hr

Abstract - Development of Web technologies has pushed forward a wide array of tools which can be used by the end-users to customize their own environment. Mashup development frameworks stand out as means to create Web applications by combining the existing content or functionality provided by various, disparate, providers on the Web. Like for all composite systems, reliability is one of the primary design concerns in mashup development. This is due to the fact that the content used in the mashup is often not under the supervision of the mashup developer and is rather deployed on the Web by a third-party provider. In this paper a reliability sensitivity analysis is conducted for the collected dataset of Yahoo! Pipes mashups. Reliability of each mashup in the dataset is estimated using a reliability model based on belief networks. A sensitivity analysis is conducted on the reliability models to ascertain which building components have the greatest impact on the overall reliability of the data set.

I. INTRODUCTION

Development and proliferation of web technologies has brought into focus a variety of end-user development frameworks. Such frameworks allow users to customize their environments by combining the existing functionalities into new added-value web applications. Common examples of such frameworks include web service composition frameworks based on the SOA architectural [1][2] style where functionalities are distributed on the Web and are accessible through a unified interface. The advantage of such an approach is in the ability to utilize the existing building blocks as programming components to create new functionalities. However, apart from the functional properties, the quality of service (QoS) for the given composite application is dictated by a variety of non-functional parameters like reliability, availability or latency [3]. Moreover, since the application building components are usually deployed by various third party providers, application developers often do not have the ability to manage non-functional properties. In that sense, reliability stands out as one of the most important non-functional parameters that impact the ability to sustainably develop composite applications [4]. In order to enable development of large-scale composite applications (applications with a large number of building components) it is necessary to analyze how the elementary building components impact the overall application reliability. That way it is possible to selectively strengthen the weakest points in the constructed application. This approach is particularly useful when a large number of building components makes the traditional reliability assurance methods, like dynamic compositions, resource intensive [5] [6] [7].

In this paper we present a reliability sensitivity analysis method for web mashups built using the *Yahoo! Pipes* [8] framework. The goal of the analysis is to evaluate how the individual building components impact the overall reliability of a mashup. To do so, we applied a belief network based reliability model. Such a formal model was chosen as it better suites the hierarhical structure of mashups then other commonly used Markov models [14]. For the purposes of the analysis, a data set containing 8569 mashup definitions has been collected. Since the collected data does not contain information on the reliability of individual building components, the reliability is defined within the experimental setup. We present the results of the sensitivity analysis, outline the most influential components and compare the results with the reliability influence measure based on the total number of component executions.

The rest of the paper is structured as follows: In Section 2 we define the reliability sensitivity analysis problem in a greater detail and give a brief introduction to *Yahoo! Pipes*. In Section 3 we describe the applied reliability model and its generation from the input mashup definitions data set. Reliability sensitivity analysis method along with experimental setup and results is described in Section 4. Finally, concluding remarks are given in Section 5.

II. PROBLEM STATEMENT

The goal of the presented research is to analyze how the reliability of basic building components affects the overall reliability of the constructed web mashup. The overview of the proposed solution is presented in Fig. 1. In order to perform the proposed analysis, a set of mashup definitions needs has been collected. Mashup definitions contain a set of rules which bind the atomic (elementary, indivisible) or complex building components into web application workflows. In the first step, the mashup definitions need to be translated into reliability models. The reliability models define the influence a particular atomic component has on the overall mashup reliability. The reliability sensitivity analysis is then conducted for the constructed models. At its output, the sensitivity analysis yields a list of building components sorted by their influence on the overall mashup reliability. The remainder of this section provides a brief introduction into the *Yahoo! Pipes* framework with emphasis on the set of available atomic components.

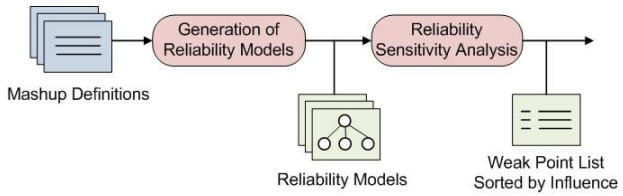


Figure 1. Overview of the proposed solution

A. Yahoo! Pipes Framework

Yahoo! Pipes is a development framework used to construct web mashups by performing GUI-level operations. The framework is exposed as a web application which enables developers to combine various web sources such as: web feeds, web pages and web services. The mashups are built using 54 atomic components (out of which 4 are deprecated) or using other existing mashups as complex building components. Atomic components are divided into 8 categories: *sources*, *user inputs*, *operators*, *string*, *url*, *date*, *location* and *number* [8].

III. Reliability Model Generation

The first step of the presented analysis is to generate reliability models form the mashup definitions. In this section we formally describe the applied reliability model and briefly discuss its generation process. Finally, we present an overview of the reliability models generated from the input data set.

Throughout the section we will refer to a Yahoo Pipes! mashup example defined by the UML activity diagram in Fig. 2. The presented mashup is used to fetch feeds defined by user inputs and his location. It is constructed using *text input*, *string builder*, *url builder* and *fetch feed* elementary building components as well as a complex building component *pipe* C_2. In addition, the mashup contains an obligatory *out* component used to forward data to its output. The query is constructed using user inputs defined through two *text input* components that are joined by a *string builder* component. The location information is acquired using *pipe* C_2 and forwarded along with the combined user input to *url builder*. Finally, the *fetch feed* component retrieves the data for the specified query and passes it to mashup's output.

A. Belief Network Based Reliability Model

In this paper, we apply a model for overall mashup reliability prediction based on belief networks [13]. The model is defined as a directed acyclic graph $M=G(V, E)$, where V is a finite set of random variables and E a finite set of edges which represent dependency between the random variables. For instance, if an edge is pointing from random variable X to Y, this means that Y is conditionally dependant on X. Each reliability model has multiple input variables, representing atomic components and a single output variable representing the total mashup reliability. In between is a hidden layer which defines how the input variables impact a single output variable.

Let X_i be a random variable that describes the behavior of component i within a mashup. For the purposes of the presented model, the space of all possible outcomes for variable X_i has two values: *fault occurrence* and *normal operation*. Thus, the random variable X_i is defined as follows:

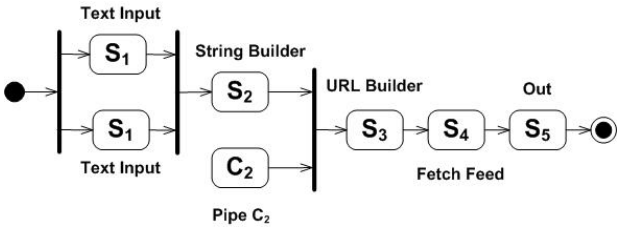


Figure 2. Example of a mashup workflow

$$X_i(\gamma) = \begin{cases} 1, & \gamma = \textit{normal operation} \\ 0, & \gamma = \textit{fault occurence} \end{cases} \quad (1)$$

The probability mass function for the given random variable can be defined as:

$$\rho_{X_i}(x) = \begin{cases} p(X_i = 1), & x = 1 \\ p(X_i = 0), & x = 0 \end{cases} \quad (2)$$

where $p(X_i=1)$ is probability that the outcome $X_i=1$ will occur. In other words, $p(X_i=1)$ stands for reliability of the component i within a mashup. The outcomes $X_i=1$ and $X_i=0$ will be simply denoted as X_i and $\neg X_i$ throughout the rest of the section.

In case a reliability model has n input variables, 2^n outcomes are possible due to (1). A set of all possible outcomes for the input variables is defined as an n-tuple:

$$P_o = \{(e_1, \cdots, e_n)\} \mid X_i = e_i, e_i \in \{0,1\} \quad (3)$$

By applying (3) and the law of the total probability, the probability of an outcome X_i ($\neg X_i$) can be calculated as follows:

$$p(X_i) = \sum_{j=1}^{2^n} \left(p(X_i \mid o_j) \cdot p(o_j) \right) \quad (4)$$

where $p(X_i \mid o_j)$ is the conditional probability of variable X_i on outcome $o_j \in P_o$ and n is the total number of edges leading into variable X_i.

Hidden nodes are used to model the influence a particular input variable has on the overall reliability. Thus, the conditional probability for the hidden nodes is influenced by the mashup's workflow constructs. As redundant execution paths and branching are not explicitly supported in the *Yahoo! Pipes* framework, we define the following model nodes: *mashup*, *sequence*, *repetition* and *loop*.

1) Sequence and Mashup

The node *sequence* is a random variable which models the probability that multiple sequential workflow elements will be executed without a fault occurrence. For instance, if a sequence consists of executing components $S_1, \ldots, S_n$ the only outcome in which a fault will not occur is ($e1$=1,

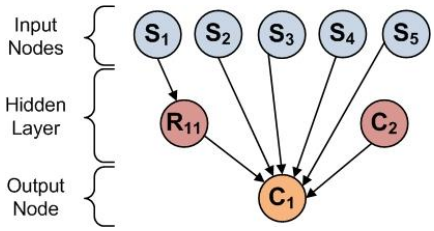


Figure 3. Reliability model for mashup example in Fig. 2

…, e_n=1). Thus the conditional probability for the *sequence* node is defined as follows:

$$p(Seq \mid o_i) = \begin{cases} 1, & o_i \in P_o \mid e_1,..., e_n = 1 \\ 0, & otherwise \end{cases} \quad (5)$$

The *mashup* node is equal to *sequence* in its conditional probability and serves as an output variable of a reliability model. It should be noted that *mashup* nodes can occur in hidden layer or input nodes of the model.

2) Repetition and Loop

The *repetition* node is used to model multiple executions of a component within a framework. Thus the conditional probability for *k* executions of component *i* is defined as follows:

$$p(R \mid X_i) = P(X_i)^{k-1} \quad (5)$$

The framework provides a *loop* component within the *operators* category which is used to explicitly repeat the execution of an assigned component. In such a case, the overall reliability is impacted both by the reliability of the *loop* component and the assigned component. Therefore, this scenario is modeled by placing the *loop* and the assigned component in a *sequence* which has an edge pointing to a *repeat* node.

For the mashup example in Fig. 2, a reliability model is constructed as presented in Fig. 3. The model contains two mashup nodes C_1 and C_2, where C_1 is the mashup's output node and C_2 is the node representing a complex building component. Since the components $S_3 - S_5$ are executed sequentially, their respective nodes in the reliability model are directly connected to node C_1. Although the components S_2 and C_2 can be executed concurrently, both of them need to execute properly to ensure normal mashup operation. Thus, this workflow construct is modeled in the same manner as a sequence. Finally, since the component S_1 is invoked two times per each mashup execution, a repeat node R_{11} (k=2) is added to the hidden layer.

B. Construction of Reliabilty Models From Yahoo! Pipes Mashup Definitions

In order to obtain the input data set for the reliability sensitivity analysis, it is necessary to translate the JSON mashup definitions into the reliability model described in the previous subsection. A simplified JSON structure for *Yahoo! Pipes* mashup definitions is presented in Fig. 4. Generally, the mashup definitions contain a set of metadata describing a particular mashup. The metadata

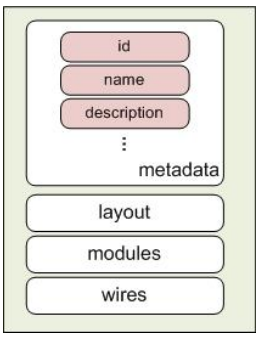


Figure 4. JSON structure of a mashup definition

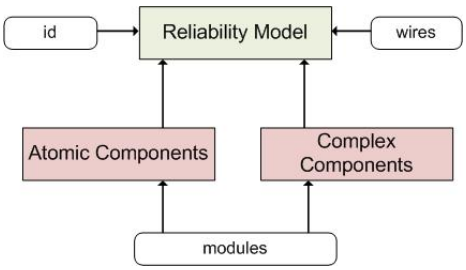


Figure 5. Reliability model generation process

contains values such as: unique mashup identifier, name, description, list of domains, list of users and time of creation. The object *layout* defines how the graphical representations of components are placed on the canvas. The *modules* object defines what atomic building blocks are used to construct a mashup. Since the *Yahoo! Pipes* framework supports the nesting of mashups, i.e. a particular mashup can be a building block in another mashup, *modules* also defines a set of complex building blocks. Finally, the *wires* object describes how particular components in *modules* are binded together into a workflow.

A high level overview of the reliability model generation process is given in Fig 5. For the purposes of this paper, we present a brief description of the model generation process. In the first step, a list of used building components is extracted from *modules* of each mashup definition. The components are stored into *atomic components* and *complex components* tables. It should be noted that duplicates are not stored into tables: the key in *atomic components* is the atomic component name, and the key in *complex components* is mashup id. The tables are maintained throughout the model generation process in order to detect incomplete mashup definitions, i.e. mashups constructed using complex components for which a definition does not exist. Such entries are discarded form the data set.

By knowing both the mashup's building components and their connections, activity diagrams, as for mashup example in Fig. 2, can be constructed. Thus, reliability models are built for the mashup data set by combining data from *modules* and *wires* objects as described in Section IIIA.

C. Overview of the Generated Data Set

The generated data set contains 8569 reliability models for the collected *Yahoo Pipes!* mashups. In addition, the data set contains 1391 (16.2%) mashups which are built

using complex components, i.e. other mashups form the data set. To give an insight into the structure of mashups, histogram in Fig. 6 shows how many mashups were built using a certain number of distinct atomic or complex components. The data shows that most mashups were built by using 4 components (1354 mashups) after which the number of mashups exponentially drops with the increase of the number of building components. From the total number of mashups, 66.97% were built using 2-6 components. Apart from the data presented in Fig. 6, the mashup definitions contained 2 mashups build using 23 components, one using 25 components and one using 46 components.

To further analyze the usage of the atomic building components, we present the total number of executions across the entire data set in Fig. 7. The presented data contains values for the 30 most executed components. The results indicate that the most executed component is *fetch* from the *sources* component category (on average 1.53 times per mashup) followed by the *textinput* component from the *user inputs* category (on average 1.02 times per mashup). Since it is an obligatory component, *out* is executed once per mashup. Other commonly used components include *urlbuilder* from the *url* category (on average 0.704 times per mashup) which is commonly used alongside the *fetch* component. The rest of the most commonly executed components are mostly used to manipulate string entries.

IV. Reliability Sensitivity Analysis

In this section we describe the algorithm used to perform the sensitivity analysis on the generated reliability models. We present and discuss the results of the analysis.

A. Sensitivity Analysis Method

Since the influence of each building component is bounded to a single random, as defined by the reliability model presented in Section 3, the reliability sensitivity analysis can be conducted utilizing the algorithm presented in Fig 8. The algorithm iterates over all the building components of the mashup and ascertains their influence on the overall reliability. This is done by setting the reliability value of a particular component to 1.0 (step 5 of the algorithm) and then recalculating the reliability of the entire mashup. That way it is possible to observe how reliable the mashup would be in case a particular component was perfectly reliable. The reliability value of the analyzed component is then reset and the procedure is repeated for all the other components. Difference between the initial reliability and the reliability achieved when $P(X_i) = 1.0$ is referred to as the reliability growth for the component X_i. The components are inserted into a list sorted decreasingly by the value of the reliability growth. Thus, the components that have a greater influence on the reliability of the mashup are at the beginning of the list as they have achieved greater reliability growth during the sensitivity analysis.

B. Experimental Setup

In order to apply the presented reliability sensitivity analysis method, it is necessary to have the reliability values for each atomic building component. It should be noted that the collected data set does not contain reliability values. In practice, it is difficult to collect reliability data for distributed systems (e.g. web services). Apart from performing direct measurements [9][10], prediction models are often used to estimate reliability values without putting additional strain on the monitored system [11][12]. For the purposes of this paper, we will assume that all the atomic building components have an equal reliability value 0.9. We acknowledge that this is a significant simplification and threats to validity it causes will be discussed in the following subsection.

The goal of the experiment is to order the components by their average impact on the reliability of mashups for the given mashup definitions and reliability values. The experiment consists of multiple reliability sensitivity analysis steps. At each step the most influential component is calculated using the algorithm in Fig. 8. The reliability of the component is then increased by 10% (to 0.99) in order to strengthen the detected weak point. The

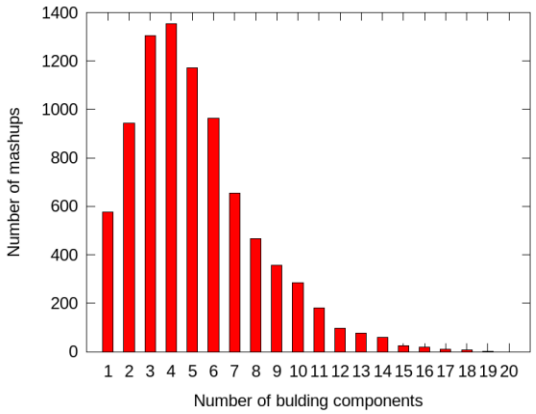


Figure 6. Number of components per mashup

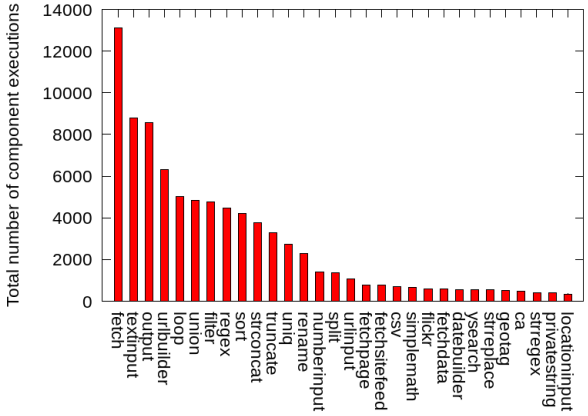


Figure 7. Number of component executions

```
1:  function RELSENSAN(model)
2:    wpList ← {}                          ▷ List of components sorted by influence
3:    for all node in model.inNodes do     ▷ inNodes - list of inputs
4:      relTmp ← node.Rel
5:      node.Rel ← 1
6:      finRel ← model.calcRel()           ▷ Recalculates the mashup reliability
7:      node.Rel ← relTmp
8:      wpList.insert({finRel, node})      ▷ Inserts (finRel,node) pair
        into the list sorted by finRel
9:    end for
10:   model.calcRel()                      ▷ Resets the mashup reliability
11:   return wpList
12: end function
```

Figure 8. Reliability sensitivity analysis algorithm

procedure is then repeated until all the 54 atomic components are strengthened. The results yield a reliability growth model for the input data set of mashup definitions.

C. Results and Discussion

The results for the first step of reliability sensitivity analysis are shown in Fig. 9. The presented results include average reliability growth achieved on the entire data set for the 30 most influential atomic components. When comparing the results of reliability sensitivity analysis to the total number of component executions presented in Fig 7, it is clear that in both cases the most influential component is *fetch*. Thus, the total number of component executions over all mashups is a good reliability influence measure for the *fetch* component. On the other hand, the reliability sensitivity analysis indicates that the *out* component has greater impact on average reliability than the *textinput* component. In this case the total number of component executions is not a good measure of reliability influence as the structure of mashups impacts the influence a particular component has on the average reliability. Specifically, if the reliability of the *out* component is set to 1, it causes an average reliability increase of 0.049192, while the *textinput* component causes the increase of 0.025448 (51.73% less than the *out* component).

To further analyze the discrepancy between the number of component executions as reliability influence measure and the proposed reliability sensitivity analysis we present the results for all the remaining analysis steps in Fig. 10. Specifically, the average reliability of all the mashups in the data set is shown in the relation to the reliability sensitivity analysis steps. The average reliability for the mashup data set is calculated as follows:

$$AvR_i = \frac{\sum_{j=1}^{N} rel_{i,j}}{N} \quad (6)$$

where i is the reliability sensitivity analysis step and $rel_{i,j}$ is the reliability of the j^{th} component in the data set and N total number of mashups.

The results indicate a steep initial increase in average mashup reliability. Specifically, 89.02% of the possible reliability increase for the given experimental setup is achieved after performing 15 sensitivity analysis and component strengthening steps. This is mostly due to the fact that the number of mashups is dropping with the number of building components, as shown in Fig. 6. Detecting such thresholds in the reliability growth model can be used by system operators to decide on how many components to strengthen. Furthermore, the results indicate that the number of component executions is a good measure for the influence a certain component has on the overall mashup reliability. This is mostly due to the fact that all the components are assigned an equal initial reliability. In addition, the *Yahoo! Pipes* workflow does not explicitly support conditional or redundant execution paths. Thus, each component in the workflow is executed at each mashup execution, regardless of the application state. In order to evaluate the difference between the

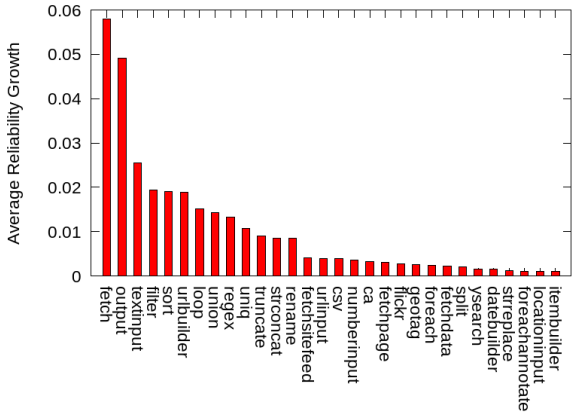


Figure 9. Reliability sensitivity analysis results (1. step)

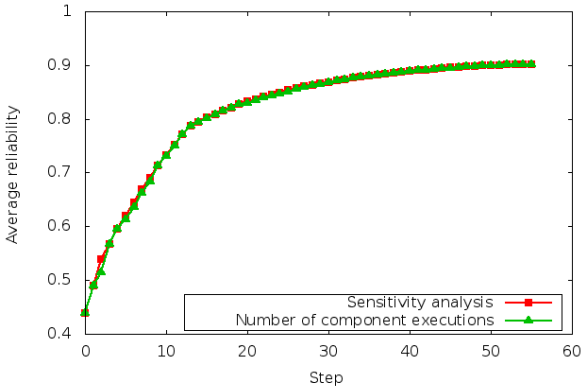


Figure 10. Average mashup reliability

reliability sensitivity analysis and the approach based on the number of component executions, we apply the root mean square error measure (RMSE) as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (re_i - ra_i)^2}{N}} \quad (7)$$

where i is the reliability sensitivity analysis step, re_i and ra_i are average reliabilities in the i^{th} step for the approach based on the number of component executions and reliability sensitivity analysis respectively, and N is the total number of component strengthening steps. For the given data set and the presented experimental setup, the RMSE measure for average reliability of the two observed approaches equals 0.004068682.

To further analyze the properties of the two approaches, we present the results for the average stepwise reliability growth in Fig. 11. The average reliability growth is calculated as follows:

$$RGr_i = \frac{\sum_{j=1}^{N} (rel_{i,j} - rel_{i-1,j})}{N} \quad (8)$$

where i is the reliability sensitivity analysis step, $rel_{i,j}$ reliability of mashup j in step i, and N the total number of

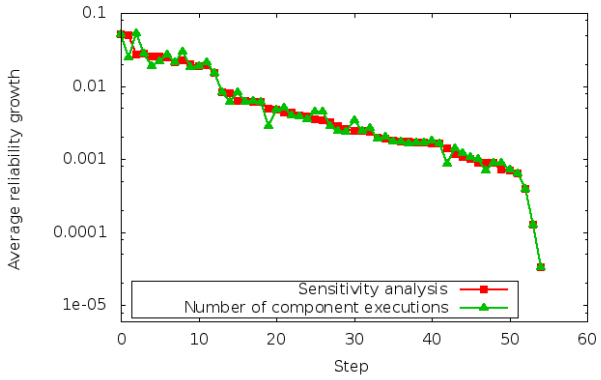


Figure 11. Average stepwise reliability growth

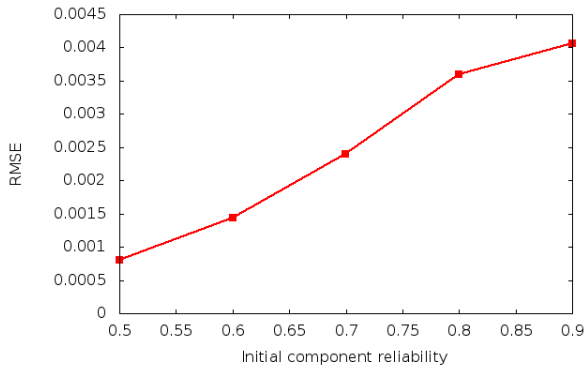


Figure 12. RMSE for varying initial component reliabilities

mashups. The reliability growth given on the y-axis of the graph is shown in logarithmic scale. The results indicate that the method based on the number of the execution counts is less stable than the presented sensitivity analysis. It is visible that errors are present in the method based on number of executions throughout the component strengthening process.

It should be noted that the main threats to validity for the presented results stem from the fact that the results are influenced by the input values defined in the experimental setup, i.e. by the initial reliability values for the building components. For instance, input values influence the RMSE value for the two described methods as shown in Fig. 12. The graph shows the dependency of the RMSE measure for a range of initial component reliabilities with identical experimental setup as described in Section 4B. Addition error can be accumulated in cases when all the components do not have an equal initial reliability as defined by the experimental setup.

Collecting reliability for atomic components is difficult in practice. In fact, the reliability of the components from the *input* category is further influenced by the reliability of attached data sources. However, we argue that the presented results give an insight into how atomic components influence the overall reliability given the collected mashup workflow definitions. In addition, the results can be used to argue how appropriate is to use the number of component executions as reliability influence measure for the given data set and reliability inputs.

V. CONCLUSION

In this paper we presented a reliability sensitivity analysis method for *Yahoo! Pipes* mashups. The analysis was conducted on a data set containing 8569 mashup definitions. Since the collected data set does not contain reliability values for individual atomic components, in the experimental setup the components were assigned an equal reliability. The results indicate that for the given experimental setup the number of component executions is a good measure for the reliability influence of atomic components. Specifically, the RMSE measure for the sensitivity analysis and the measure based on the number of executions equals 0.00407. This result is mostly influenced by the fact that atomic components were assigned equal initial reliability and that the *Yahoo! Pipes* framework does not support redundant or conditional execution paths. Finally, the constructed reliability growth model indicates that 89.02% of the possible reliability growth is achieved after strengthening 15 most influential atomic components.

REFERENCES

[1] D. Krafzig, K. Banke, and D. Slama, “Enterprise SOA: Service-Oriented Architecture,” Best Practices (The Coad Series), Upper Saddle River, NJ, USA: Prentice Hall PTR, 2004.

[2] W3C, “Web Services Architecture,” http://www.w3.org/TR/ws-arch/, February 2004.

[3] Q. Liang, X. Wu, and H. C. Lau, “Optimizing Service Systems Based on Application-level QoS,” IEEE Transactions on Services Computing, Vol. 2, No. 2, pp. 108 –121, April-June 2009.

[4] F. Curbera, R. Khalaf, N. Mukhi, S. Tai, and S. Weerawarana, “The next step in Web services,” Communications of the ACM - Service-oriented computing, Vol. 46, No. 10, October 2003.

[5] S. Y. Hwang, E. P. Lim, C. H. Lee, and C. H. Chen, “Dynamic web service selection for reliable web service composition,” IEEE Transactions on Services Computing, Vol. 1, No. 2, pp. 104 –116, April-June 2008.

[6] S. Kalasapur, M. Kumar, and B. Shirazi, “Dynamic service composition in pervasive computing,” IEEE Transactions on Parallel and Distributed Systems, Vol. 18, No. 7, pp. 907 –918, July 2007.

[7] L. Zeng, B. Benatallah, A. Ngu, M. Dumas, J. Kalagnanam, and H. Chang, “Qos-aware middleware for web services composition,” IEEE Transactions on Software Engineering, Vol. 30, No. 5, pp. 311 – 327, May 2004.

[8] M. Pruett, Yahoo! Pipes, 1st ed. O’Reilly, 2007.

[9] Z. Zheng and M. Lyu, “Ws-dream: A distributed reliability assessment mechanism for web services,” in Dependable Systems and Networks With FTCS and DCC, DSN 2008., June 2008, pp. 392 –397.

[10] M. De Barros, J. Shiau, C. Shang, K. Gidewall, H. Shi, and J. Forsmann, “Web services wind tunnel: On performance testing largescale stateful web services,” in Dependable Systems and Networks, 2007. DSN ’07, June 2007, pp. 612 –617.

[11] M. Silic, G. Delac, I. Krka and S. Srbljic, “Scalable and Accurate Prediction of Availability of Atomic Web Services,” IEEE Transactions on Services Computing, accepted paper.

[12] Z. Zheng, H. Ma, M. Lyu, and I. King, “Qos-aware web service recommendation by collaborative filtering,” IEEE Transactions on Services Computing, Vol. 4, No. 2,pp. 140 –152, April-June 2011.

[13] D. Koller and N. Friedman, “Probabilistic Graphical Models: Principles and Techniques”, MIT Press, 2009.

[14] F. Brosch, H. Koziolek, B. Buhnova, and R. Reussner, “Architecture-Based Reliability Prediction with the Palladio Component Model,” IEEE Transactions on Software Engineering, vol. 38, no. 6, pp. 1319-1339, Nov.-Dec., 2012

Requirement Bottlenecks in a Cloud based SCADA system

D. Sutic* and B. Atlagic *
* Faculty of Technical Sciences/Department for Electronics, Energetics and Telecommunications, Novi Sad, Serbia
davor.sutic@gmail.com
batlagic@gmail.com

Abstract – High expectations of the cloud hold great promises for computer scientists and engineers. However, its applications in realtime systems, especially in SCADA, are still in infancy stage. Besides, while there were a lot of researches working on integrating scientific systems on cloud systems, even the study on the mentioned topic was limited. Therefore, the authors seek to address some major issues in creating a cloud compatible realtime SCADA system. These are: data storage and database management, security of a system hosted on the cloud and system failure and recovery on the cloud.

I. INTRODUCTION

The phrase "in the cloud" has entered into our colloquial language. The cloud, this amorphous entity that is supposed to represent the future of modern computing. But, under all those statements of a new-age and unprecedented approach, the cloud is in reality a technology that had been in use for quite some time now - the Internet. When the Internet, with all its associated protocols and standards, is perceived as a cloud, one of the most fundamental and essential characteristics of the cloud is shown: abstraction. Resources are pooled and partitioned as needed with a standard-based communication.

The Internet was begun as a network of networks, with an architecture that was redundant and could survive massive disruption. What the original system architects of the Internet could not have anticipated is that the size of resources attached to it would become massively scalable, which is the second characteristic of cloud computing.

Google's infrastructure, for example, spans 30 datacenters around the world with over a million computers; infrastructure that Google now leases out to developers upon which applications may be staged. So the third and equally important characteristic of cloud computing is that the cloud is a "utility", which has basically been the promised land for computer scientists and engineers, and that services are provided using a pay-as-you-go model.

Cloud computing represents a real paradigm shift in the way in which systems are deployed. The massive scale of cloud computing systems was enabled by the popularization of the Internet and the growth of some large service companies. Cloud computing makes the long-held dream of utility computing possible with a pay-as-you-go, infinitely scalable, universally available system. That's why cloud computing is revolutionary, even if the technology it is built on is evolutionary.

The paradigm of realtime computing is different. The business data model, which is dominant in cloud applications, and which presumes that data is stored somewhere on the cloud and that a user accesses it through a web page in order to manipulate it (example: an e-mail client), is not applicable to realtime requirements.

The criticality implies that a very large amount of data needs to be processed before a deadline. The consequences of failing can be severe, in so-called hard realtime systems even pose a threat to human lives or escalate to a catastrophe. Under such circumstances the "best effort" approach is not deemed enough.

It is understandable that engineers are pessimistic towards such solutions, yet, on the other hand, they are thrilled to make the best out of a new technology by submitting it to one of the most challenging purposes today. The cloud may be considered the next logical step in the development of computer science and as such it will require a paradigm shift as well in thinking as in practice [1].

This paper will use a SCADA (Supervisory Control And Data Aquisition) system as a relevant real-time system. We will discuss some major features that are expected from a SCADA system, but are not obviously cloud specific, and propose ways to solve them.

In the third chapter, we examine the data flow that is expected from a realtime environment and how a cloud system could deal with the increased quantities of data and the imperatives of low-latency and criticality.

The fourth chapter deals with basic cloud security and ways that make communication over the internet possible without posing a threat to the firewall integrity of the SCADA control system.

In the fifth chapter the importance of redundant data paths is discussed, as the system should be able to react in milliseconds to any critical failure of the infrastructure and have at least one backup system ready to take over the control in such a case.

It is assumed throughout the paper, that the goal is to ensure that the system architecture gets the most out of the cloud benefits while retaining the realtime paradigm. The cloud system is the center of the architecture, it performs most operations and stores virtually all data. The field outstations transmit their data to the cloud and receive commands, by themselves, they store as little data as possible, ideally nothing that is not present in the cloud database, and do only the basic processing. The same applies to client machines, whose primary task should be displaying data to the user and sending instructions back to the cloud. Ideally, the client is supposed to be operated via a "light" application, arguably even through a web browser, with little requirements for processing power and close to none for memory.

II. PUBLISH/SUBSCRIBE APPROACH JUSTIFICATION

Data storage and database management has traditionally been of great interest for many system architectures and solutions since it deals with bottlenecks which are most difficult to overcome. Issues like disproportionate operational times of storage media and processing units and access collision management demand special attention in design.

As Buchmann defines it, a realtime database is a processing system designed to handle workloads whose state is constantly changing [2]. Such database concepts are in common use in most process control systems; a number of field stations communicates with a central system, which holds information about all of them and is in charge of transferring changes, at any time and as close to realtime as possible.

When deploying a SCADA system on the cloud, there are several requirements to be met.

Traditional relational databases are not designed to handle the high amounts of read/write operations or the criticality a realtime environment requires. A reason more is that such data structures are not optimized, they take much time to build and too much bandwidth to transmit. Related to that, it is obvious that an on-disc database is also unsuitable, since it introduces additional time losses in communication. Therefore, we suggest the use of a non-relational, in-memory database, as it is still valid for the most of high-end SCADA solutions currently available on the market [3]. Some may raise the question if an in-memory database is a happy choice. It is clear that the cloud environment offers near limitless physical resources, as this represent one of the basic ideas, so it is safe to assume that the storage capacity won't be exhausted. Additionally, the database should, by design, keep compact structured data, not only for the sake of the size, but also for the sake of transmission. Finally, the main concern may be the case when the system goes offline. The issue of an outage will be approached bellow. Regarding a regular shutdown, a realtime system is almost never expected to be offline, but, of course, there should be the possibility and in that case there should be a mechanism to transfer the data to the disc or some other external storage. Note, that in this paper, we focus solely on the realtime aspects, it is expected and often even

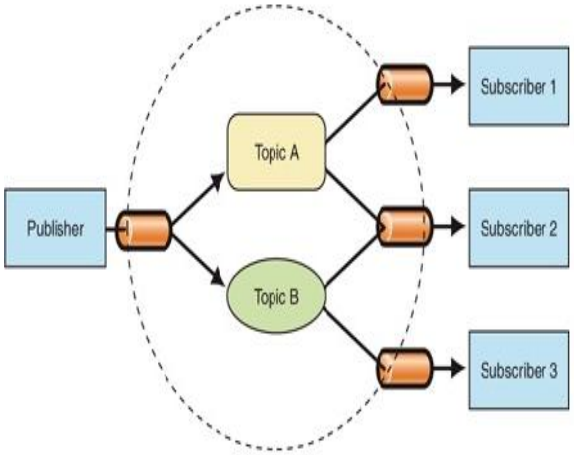


Figure 1 Publish/Subscribe approach

required, to keep, e.g., statistical data on the cloud, but this kind of database is of no issue, since it is not a part of the realtime environment and can be deployed and used as a regular relational database, the kind that is presently utilized on the majority of public cloud servers.

Since relatively large amounts of data are expected to flow frequently through the network, a low-latency approach is mandatory. As mentioned above, some data structures take up too much bandwidth to transmit. The minimal bandwidth is required when the data stays during transmission in its raw form. Sure, the data will eventually be required to be cast into SQL, HTML, XML, etc. but that point should be delayed as much as possible, i.e. ideally at the final destination. The bottom line is that all data traffic between the cloud and its peripheries should be in the raw form, every adjustment and reinterpretation should be done at the peripheries. Additional benefits from this approach are easy scalability and the ability of the cloud to communicate with various clients. A possible drawback would be that the peripheries need to convert received data and prepare data for sending, which would result in a requirement for performance increase of them. But, today, such minor expenses should be considered tolerable and, besides that, a client handles only a small fraction of data in comparison with the cloud, so this additional workload is justified.

The last point implies that a mechanism is needed which will ensure that each user gets only the data intended for him and this acquisition is expected to be automatical, since the user is in no position to manually "ask" the cloud for new values. A common solution is an event-based architecture in form of a "publish/subscribe" approach (). Namely, the client is expected to subscribe itself to a set of data, this is noted by the cloud system and it sends the client the updates of that subscribed data. Of course, this relation is supposed to work bilaterally since the client can make changes to that data and those changes should be propagated to the cloud. This approach greatly reduces the stress on the transmission and bandwidth, because each subscription, i.e. each client, will have a relatively small amount of data to handle, expectedly at most a few hundred points. These "small amount" subscriptions will change often, but it can be assumed that

these changes carry little data and require a bit more processing power, which shouldn't be considered an issue.

III. KEEPING THE CLOUD SAFE

Today, typical SCADA and control systems support high-value production and distribution lines, which are expected to be of strict security, because every irregularity, intrusion or security breach could potentially have serious consequences. Engineers in charge of such institutions are thus very skeptic to potential loopholes in the security system. LAN networks are now accepted, although their introduction was reluctant, but this is mainly done under the condition that the LAN is not connected to the Internet. Opening a port to the cloud, which is located on the Internet, is still considered a threat [4].

On the other hand, more "open" systems, systems that are geographically dispersed obviously rely on an Internet link, but they consider the trust in a cloud server provider a threat.

Today, most cloud security providers offer, so called, end-to-end security solutions. That means that the whole network path from the client and outstation machines to the cloud-hosted server is to be covered by a third party security system, e.g. Cisco ASA 1000V. Of course, such systems are designed to support hybrid architectures, which consist of physical, virtual and cloud infrastructures.

The main issues with these highly sophisticated defense mechanisms lie in their self-management and complexity. Although less clarifying, perhaps, it would be more accurate to summarize those weaknesses under one problem: the cloud firewall being a third-party component.

Every insight into the protection system would have to be negotiated with the provider, who, ultimately, has no or limited obligations to reveal details to the customer. Always under the shroud of "The protection software is designed to do its job and every detail that is disclosed might lead to a security compromise".

Such sophisticated protection implies a layered structure which covers almost every bit of communication and equipment. While this is by no means a bad thing, due to the general purpose of these systems, it is sure to produce some overhead.

Since real-time systems are sensitive to every amount of possible optimization, measures should be taken to ensure best performance. On the other hand, if a sufficiently acceptable security system is found and favorable conditions can be negotiated with the provider, there is nothing wrong in accepting third parties in the system.

Ultimately, it is reasonable to expect that at least a minor security management will be enforced by the cloud provider. However, we recommend that a full-scale security system should be custom built for a sensitive system, such as a real-time environment. This paper proposes guidelines for designing such a security infrastructure. With the explosive growth of network-based services and attacks, the complexity and cost of firewall deployment and management have been increasing rapidly. Yet, each private network, no matter big or small, has to deploy and manage its own firewall, which is the critical first line of defense [5].

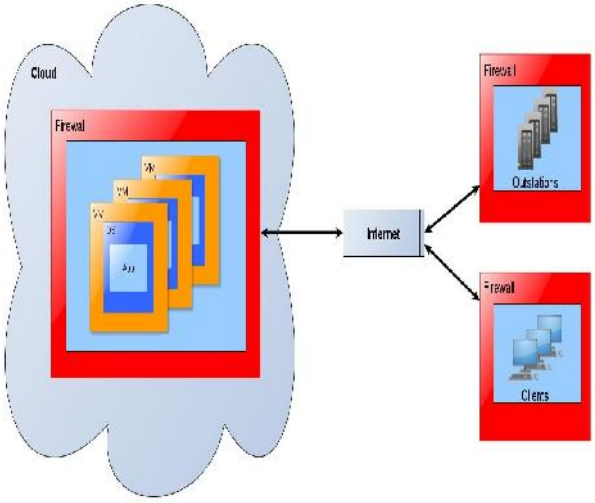


Figure 2 Firewall structure

As shown in Figure 2, a realtime server would be spread over a set of virtual machines (VMs) assigned to the user by the cloud provider. Since this already gives the user the illusion that he is handling separate servers, there is no need to trust the cloud firewall, or, for that matter, to make it a greater part of the security plan than it necessarily already is. For easier maintenance it is even more convenient to set up one's own firewall around the system, as part of a middleware, which would correspond to the security system the clients and field stations utilize. This would focus the security system on only the domains it should be covering, eliminating the need for complex security software and involvement of third parties.

Naturally, this solution does not remove the security risks and mistrust to the cloud provider, but hardly any would, except the unrealistic task of actually owning a cloud provider.

IV. REPLICATION IN CLOUD

For realtime systems, especially SCADA process control systems, a mandatory requirement is a 100% uptime, as even minor disruptions could have serious consequences. This is even more true as the scale of a system goes up and with it the responsibility and geographical coverage.

Fortunately, one of the basic premises of cloud computing is that the cloud will ensure a near to 100% uptime through its abstraction of physical machines to the user. But, as relatively recent outages of the Amazon EC2 cloud service show, the reliability of such systems has still a way to go to improve. Namely, in 2011 parts of the Amazon EC2 server, which hosts an alarmingly high percentage of the internet's top popularity content, e.g.

Netflix, Instagram, etc., went down in Virginia, United States. To make matters worse, the outage was induced by severe thunderstorms in the area. Now, although the mentioned web-sites suffered downtimes, which beyond doubt frustrated their users, no real harm was done, except maybe to the image of cloud reliability. Obviously, the damage would be much more severe if there were any realtime systems hosted at that time.

In SCADA systems, redundancy has been used to ensure correct functionality even if parts of the system shut down. That means that the risk of a complete system failure is diminished by having a replica of the system ready to take over when an outage occurs. This backup system is idle, waiting in standby, as long as the main system is functioning. Depending on the type of the standby mode, the backup system can be "cold", "warm" and "hot". In this paper, we will focus only on "hot-standby" systems, as they provide the most realtime-friendly option, because they represent virtually active systems and a switch over can take a very short time, in the scale of milliseconds [6].

The term replication refers to the process of constantly copying the state of the main system to the backup system, i.e. all transactions, data and commands. The general concept, adapted to the cloud environment, is demonstrated in Figure 4.

In Figure 4 is shown only one direction of information flow – from the data sources to the server and from there to the clients. Of course, there are other paths, but this was omitted since it is not relevant for the matter at hand. The same case applies to the concept of middleware. Remote stations and clients are required to have a software interface to the cloud server deployed. The details regarding this piece of software are beyond the scope of this paper. For the ideas explored here it is important to perceive the middleware as a union of the data source or client and the cloud interface with a communication foundation that connects them.

The red lines indicate active communication paths, while the black show alternative data paths. The active cloud system constantly replicates its state on the backup system, so it can in a case of failure take up the work. Ideally, the middleware should be able to abstract from the clients and remote stations the redundancy mechanisms, i.e. clients and remote stations should be oblivious to system failure and the resulting switch over. Finally, when a switch over occurs, the alternate data paths become active and the system resumes to function normally.

We propose that at any time at least one cloud system is used as backup for the main. This may seem a somewhat radical solution, given that a cloud is deemed stable based on its basic ideas of architecture. But, it is necessary to tend to two major bottlenecks in communication.

The first is the outage of the whole cloud system, like the Amazon EC2 case. Although very rare, this scenario is to be expected and there should no risks to be taken that allow that a critical process depends on the stability of one service.

Name	CPU Cores	RAM [Gb]	Archi. [bit]	Disk [Gb]	Cost [$/h]
Amazon EC2					
m1.small	1	1.7	32	160	0.1
m1.large	2	7.5	64	850	0.4
m1.xlarge	4	15.0	64	1690	0.8
c1.medium	2	1.7	32	350	0.2
c1.xlarge	8	7.0	64	1690	0.8
GoGrid(GG)					
GG.small	1	1.0	32	60	0.19
GG.large	1	1.0	64	60	0.19
GG.xlarge	2	4.0	64	240	0.76
Elastic Hosts (EH)					
EH.small	1	1.0	32	30	£0.042
EH.large	1	4.0	64	30	£0.09
Mosso					
Mosso.small	4	1.0	64	40	0.06
Mosso.large	4	4.0	64	160	0.24

Figure 4 Cost/Performance table

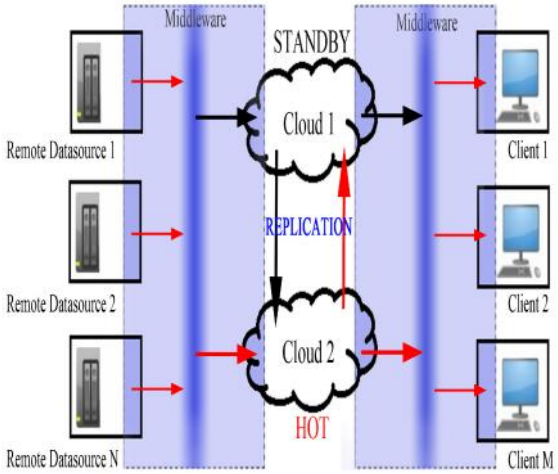


Figure 3 Redundancy principle

The second is a more common one, the communication links to the cloud system themselves. As all network connections, these are prone to malfunctions and can cause the failure of certain parts of the system.

An obvious drawback is the additional cost which effectively doubles for each backup system. For comparison, on the table in Figure 3 are the prices of some influential vendors shown [7]. Regarding this issue, the cloud provider should be able to scale at a satisfying rate. That means, quickly enough to compensate eventual computing bursts and idle periods, which will save money. Real-time systems are expected to be up and running 24 hours 365 days per year, so even minor optimizations could significantly lower funding demand. If the fact is taken into account, that an idle computer uses about 65% of the power of a busy machine, careful implementation of the whole system could reduce the negative impact on the environment, which is currently a major global concern. These considerations point out the importance of choosing both cloud providers, main and backup, wisely and according to that choice, efforts

should be made to optimize the SCADA system in order to obtain optimal resource control.

Another "financial" aspect is software licensing. If the SCADA system uses some third-party proprietary software, issues could arise when the licensing is concerned. Today, it is common that a license is issued for a specific machine, which, of course, proves to be complicated when a cloud provider is concerned. Many providers advocate open-source software, but this cannot always be a solution, sometimes a licensed commercial program is necessary. One possibility is the vendor issuing a cloud specific license, which is a more expensive solution. Typically, those services are charged by a price per hour, e.g. running Microsoft Windows on EC2 costs around $0.15 per hour. In any case, since this is not directly a problem that can be solved in implementation, it should be considered which third-party products the SCADA system is obliged to use. This can significantly raise the maintenance cost, so it is important to discuss these matters with the cloud providers early on.

Also each backup system should be from another provider, which induces additional effort to implement the communication between the various clouds. Basically, this can be solved with a special communication mode, e.g. "BackUpLink", which is in charge of distributing the state of the main system to all the backup systems. This does not only places the requirement for an appropriately wider bandwidth, but opens also a whole plethora of questions dealing with the issue of implementing a secure distributed network of cloud systems.

One of the most pressing issues in this matter is the possible data lock-in which might occur on a particular cloud provider, i.e. the hazard of data staying trapped on a cloud site. This difficulty of extracting data from the cloud concerns certain potential customers so much that it prevents their organizations from adopting the cloud computing paradigm. For example, an online storage service called The Linkup shut down in 2008 after losing access as much as 40% of customer data. The Linkup, in turn, had relied on the online storage service Nirvanix to store customer data, and now there is finger pointing between the two organizations as to why customer data was lost. Meanwhile, The Linkup's 20,000 users were told the service was no longer available and were urged to try out another storage site.

Cases like this one point to a somewhat darker aspect of cloud provider business, the customer lock-down. It is quite attractive since without a standardized cloud API, the customers are much less likely to abandon their providers and switch to another, but, as the example above shows, it can under circumstances drive the provider out of business or, in a minor case, inflict damage to the trust of customers. On the other side, it makes users more vulnerable to price increases, to reliability problems and even to data loss, as shown before.

Of course, the best solution for this issue would be the global one: a common API for all cloud providers. That would effectively allow customers to deploy their data across multiple providers and rendering them immune to a single cloud company's failure. In addition to the above mentioned customer lock-down policy, providers could fear the rise of a "race to the bottom" of pricings and cause profits to dwindle. However, it is our opinion that such fears are not justified for reasons of basic economics: just because something is cheap, it doesn't mean that it is the best. As in every other aspect of trade, if the offered service matches the price, the product will have customers. Furthermore, a standardization could create possibilities for a deeper collaboration between providers and lead to better quality of services.

Be that as it may, at the time of writing of this paper, there are no tendencies of creating a common API between cloud providers. So, our concept of replication requires that such considerations are taken into account and a specific API for the above mentioned concept of a distributed network of cloud systems is implemented. Given the criticality of real-time systems, it would probably be advisable to implement this API into the solution even if, eventually, a standardized API is deployed, even if only a wrapper would be sufficient. However, at the time being these are considerations for future works.

Implementing the redundancy concept in a cloud environment might arguably prove the most challenging task. This seems somewhat of a paradox since it seeks to eliminate the danger of outage in a system that is deemed to have an over 99% stability. However, taking into account the nature of a realtime system and the possible consequences of this less than 1% outage time, redundancy will be a mandatory feature.

V. CONLUSION

The next step in development of computer systems is expected to be the transfer to the cloud. The paradigm of completely abstracting the physical architecture and shifting the developer and user concerns to only service issues was the endeavor for many generations of computer scientists and will surely show its benefits after the initial adaptation period. The impact of the new idea is felt also in the domain of realtime applications, which opens the particularly interesting challenge of hosting a SCADA system on the cloud.

This paper proposes approaches in dealing with the some issues that are expected from SCADA realtime systems, but not necessarily part of the cloud paradigm.

All SCADA systems have a sensibility to database management, thus, the largely static databases, which are commonly utilized by major cloud providers, are not suitable to deal with the highly dynamical data environment of a realtime application. Low-latency policy and large amounts of data, which are expected to be quickly processed and distributed, lead to the need to utilize a non-relational database and to distribute data in its raw form, i.e. to postpone any conversions into, and, of course, back from, SQL, XML, etc. as close as possible to the client. Furthermore, a publish/subscribe event based approach should be implemented, which will reduce the bandwidth requirement and ensure that each client only receives data relevant for it.

Given the criticality of SCADA applications, many trust issues have been raised regarding cloud security. Engineers in a high security institution are understandably reluctant to opening ports to a third party cloud provider over the Internet. Additionally, it would be difficult to maintain a security system that incorporates the cloud provider. So, we propose to implement the security firewall around the cloud service, deployed on a cloud system, itself, which is paired with only the security on the outstations and clients.

Finally, special concern represents the case of a system failure. Although rare and marginalized in the cloud environment, an outage is a valid high risk scenario. Again, “regular” cloud applications are not affected that much, but a realtime system failure could have serious consequences. We propose that a cloud SCADA system should have at least one backup system, which can take the main ones place in case of an outage. Since the switch should not degrade the performance of the system, the backup system has no time to be configured from the scratch. This implies that the backup system should always be prepared, thus a constant replication from the main system of all data and performed transactions is to be carried out. In order to provide adequate protection, the each backup server should from a different cloud provider.

From the previous exposure it is apparent that the mentioned issues could be realized through a specific middleware, which could be subject of further related work. Future research can deal with additional issues that would arise in the process of designing the middleware.

At the time of writing of this paper, there are only experimental Cloud-based SCADA systems, however, intense research is being conducted on this topic and its derivatives. More practical results of the endeavor of migrating a SCADA architecture to the Cloud can be expected when a few competitive commercial solutions arise and their engagement is analyzed .

REFERENCES

[1] H. Jin, S. Ibrahim, T. Bell, L. Qi, H. Cao, S. Wu and X. Shi, “Tools and Technologies for Building Clouds,” Cloud Computing: Principles, Systems and Applications, Computer Communications and Networks, Springer-Verlag, Berlin, 2010

[2] Buchmann, A. "Real Time Database Systems." Encyclopedia of Database Technologies and Applications. Ed. Laura C. Rivero, Jorge H. Doorn, and Viviana E. Ferraggine. Idea Group, 2005.

[3] B.Atlagic, M.Sagi, D.Milinkov, S.Culaja, B.Bogovac, “A way towards efficiency of SCADA infrastructure”, ECBS 2012, Novi Sad 2012.

[4] R. L. Krutz, “Securing SCADA Systems,” Wesley Pub-lishing, Inc., Indianapolis, 2006.

[5] Khakpour, A.R.; Liu, A.X., "First Step toward Cloud-Based Firewalling," Reliable Distributed Systems (SRDS), 2012 IEEE 31st Symposium on , vol., no., pp.41,50, 8-11 Oct. 2012

[6] Gray, J., Reuter, A., Transaction Processing: Concepts and Techniques, Morgan Kaufmann, San Francisco, CA. 1993.

[7] A. Iosup, S. Ostermann, N. Yigitbasi, R. Prodan, T. Fahringer, D. Epema, “Performance Analysis of Cloud Computing Services for Many-Task Scientific Computing”, IEEE TPDS, Many-Tas Computing, November 2010

Impact of Control Flow Blocks Granularity on Custom Processor Design Time

D. Ivošević, V. Sruk
Faculty of Electrical Engineering and Computing,
Department of Electronics, Microelectronics, Computer and Intelligent Systems
Zagreb, Croatia
danko.ivosevic@fer.hr, vlado.sruk@fer.hr

Abstract - One of the key issues for system level design topic is the design time. This paper describes custom processor design tool as part of C-to-hardware flow and analyses its design time. The flow starts with C code specification and ends with FPGA implementation. The way the C code is processed has impact on the flow execution time. The implemented C code processing results with Control Flow Graph (CFG), and large control flow code blocks severely prolong the overall design time. Between two possibilities for design time improvement, variations in their granularities are chosen over tool internal algorithm and data structures optimizations. For 32-point DCT test case the results show huge design time decrease at the expense on the design quality: implementation resource occupation and execution time.

I. INTRODUCTION

Embedded System Level (ESL) design is an emerging methodology for custom digital system design with focus on higher level design specification. With constant rise of capability of technology and the growth of software and hardware productivity gaps the needs for quality and reliable software support increase, [1].

Software solutions help in area of Electronic Design Automation (EDA) tools development and IP cores production that helps in closing of hardware productivity gap.

This work is motivated by this challenge and deals with the topic of custom processor architecture design from C code specification. The specification in C code and its derivatives, C++ and SystemC, is dominant in High-Level Synthesis (HLS) which usually produces hardware description in RTL code that is ready for logic synthesis targeting FPGA device. The research of this methodology had been the most intensive in academic society in past decades, and, during the time, a number of commercial tools appeared in the market, [2,3].

In this work, one of the production targets is a custom processor architecture model designed for C code specification. The model of architecture is designed according to No-Instruction-Set Computer (NISC) concept, [4]. In this concept, the system designer can arbitrarily choose datapath components having in mind the C application code features. On the contrary, in this paper the task and methodology of automated datapath design is presented. The resulting architecture that contains such automatically designed datapath is in final implemented in FPGA. The methodology assumes no compiler-style optimizations, but fully customizes the architecture. Thus, the emphasis is on optimizations that are applied on datapath level. In such way, the optimizations are closer to the implementation platform while the concept retains processor style of execution familiar to common user. The traditional HLS flow [5] is broken into two stages: processor as the execution engine, and its mapping to the implementation platform.

In previous works custom datapath design for such processor model is analyzed through manual datapath transformations in several iterations [6] and its automated construction [7]. Besides these, we elaborated manual and automatic datapath generation for specific BDD application with highly recursive nature [8] and DCT application code [9].

In the following sections, Section II and Section III, the overview and methodology of implemented design flow is presented. Section IV presents the results for several test cases, and Section V focuses on design time for 32-point DCT case which appeared to be time demanding. Section VI concludes on the presented methodology and results.

II. IMPLEMENTED DESIGN FLOW

Design flow is characterized by several processes, and their inputs and outputs. Globally, there are three major processing steps in the design flow, Fig. 1:

1. C code preprocessing.
   Code is analyzed by its procedures and basic blocks formed by procedures control flows. Resulting notation is Control and Data Flow Graph (CDFG), [10]. SPARK parallelizing compiler is used for initial transformation of C code to CDFG. Further it is altered by our tool to conform the later stages of design flow, [11].
2. Architecture build.
   As the most complex process of the flow it is implemented within our *ArkBuilder* tool. Three separate processes can be identified:
   a. Scheduling of CDFG three-address code statements.

b. Operand and operations usage analysis. Such analysis produces combinations of operands and operations grouped within registers and functional units. Thus, the simplified (or provisional) datapaths consisting only of register files, functional units and their connections are formed for every basic block.
c. Design of final datapath. The design is based on integrating all basic blocks datapath contributions. The result is datapath completed with data memory, multiplexers as arbitral components instanced at other components inputs, and connections interfacing the control unit.

3. Simulation/Implementation.
Specially designed tools *GenCM* and *ProcSynth* generate design RTL description and instruction memory initialization file for core generation. With inclusion of appropriate data memory core the design is synthesized and implemented in FPGA. The behavioral simulation is used for verification purposes. The synthesis, implementation and simulation steps are accomplished with *Xilinx ISE* toolset, [12].

The central point of the flow is the algorithm for final datapath design. It forms the complete architecture that logically corresponds to CDFG 'per basic block' schedules. The datapath contributions of all basic blocks are integrated into final datapath having in consideration basic blocks significance. The significance is defined by their execution cycles shares inside whole application run. Such shares are estimated by profiling of procedures and basic blocks, and basic blocks schedule lengths. Basically, the algorithm for final datapath design is provided with following input information:

- CDFG application description
- Application code profiling information
- Implementation platform description

Here introduced aspects of datapath design are described in following sections using short example to clarify the theoretic information.

III. Processor Design Methodology

The code profiling is performed at two levels of abstraction: procedure level and basic block level. At final, it is flattened at basic block level as number of procedure calls is multiplied with basic block iteration count to get the absolute basic block iteration count.

The code analysis is thus performed for each basic block independently. As the original code is transformed in basic block three-address code notation, its mapping to architecture datapath is straightforward. Functional units with two input ports and one output port correspond to the three-address code notation of basic block statements. Scheduling, allocation and binding tasks are performed on basic block three-address statements. Firstly, the scheduling of statements produce finite state machine with data (FSMD). The analysis of statement operands and operation is performed for every cycle and their non-overlapping usages are noted to allocate registers and

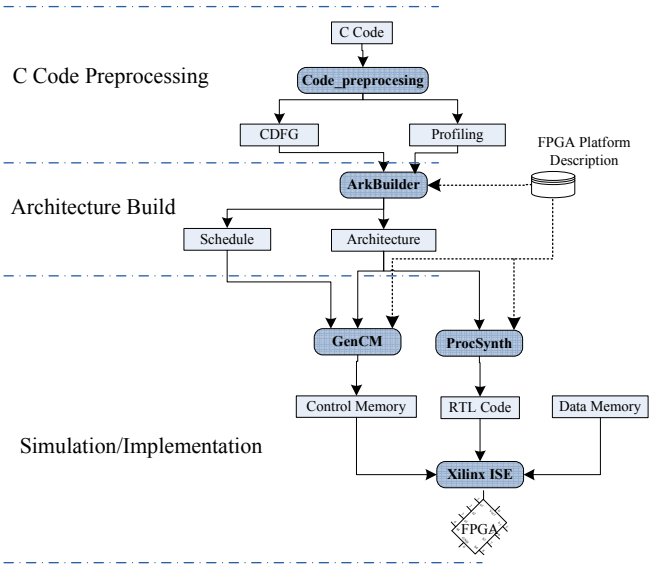


Fig. 1. Design synthesis flow.

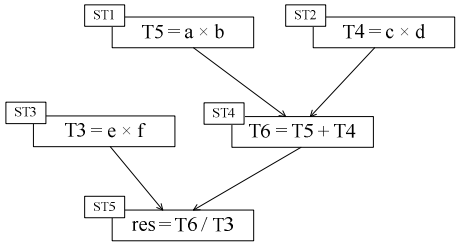


Fig. 2. Data flow graph for expression in (1)

functional units, respectively. In such way, the binding of statements to registers and functional unit is implicit.

A. Basic Block Analysis

As stated in Section II there are three steps in the methodology of architecture build. First two steps, scheduling and basic blocks analysis with forming of simplified datapaths, are performed for every basic block of the application CDFG. The last, final datapath design combines all basic blocks datapaths into unique one and interfaces it to control unit. Here, the methodology is going to be elaborated through simple line of code in (1) assumed to be only basic block content. It is a simple calculation consisting of three multiplications, an addition and a division.

$$g = ((a \times b) + (c \times d)) / (e \times f); \quad (1)$$

Fig. 2 shows data flow graph of (1) consisting of five three-address code statements: ST1 to ST5. Data dependencies are properly extracted to ensure the regularity of calculation. According to those dependencies, the scheduling phase produces finite state machine as in Fig. 3a. Usage analysis of operands and operators notes operand usages and operation activity per state, as in Fig. 3b and 3c.

If one register per operand and one functional unit per operation would be assumes, there would be much

State	Statements	Three-Address Code
S1	ST1	T5 = a × b
	ST2	T4 = c × d
S2	ST3	T3 = e × f
	ST4	T6 = T5 + T4
S3	ST5	res = T6 / T3

a)

Operation	S1	S2	S3
ADD	0	1	0
MUL	2	1	0
DIV	0	0	1

b)

Operand	S1	S2	S3
a	×		
b	×		
T5		×	
c	×		
d	×		
T4		×	
e		×	
f		×	
T3			×
T6			×

c)

Fig. 3. Scheduling (a), and operation (b) and operand usage analysis (c) for (1)

redundancies in the datapath. Therefore, minimization of registers and functional units allocations is applied using compatibility graphs, [13]. For instance, for operands that are not used in same states the priority edges are defined with *i/o* priority notes, Fig. 4a. Value *i* is the number of same operation types for which corresponding operands are input values, and *o* is the number of the same operation types for which they are outputs. For operand used in the same state the incompatibility edges are defined, Fig. 4b. By merging operands with priority edges between, final compatibility graph is derived having only incompatibility edges, Fig. 4c. Operands that are merged to the same node share a register. In the same manner, the combinations of operations are dedicated to functional units. Using this technique, final datapath for expression (1) consists of two functional units and four registers appropriately connected, Fig. 5.

B. Final Datapath Design Algorithm

Algorithm that designs the final datapath integrates contributions of all basic blocks provisional datapaths. The pseudo code that describes it, is following:

$DP = \emptyset$
$FUmax_{op}$ = GET_MAX_INSTANCES(CDFG),
$op \in$ {ADD, SUB, …, ASSIGN}
for $bb_{current} \in CDFG$
  FU_{sorted} = SORT_FU ($bb_{current}$, desc)
  for $fu_{current} \in FU_{sorted}$
    Op= GET_OPERATION($fu_{current}$)
    if $FUinstances_{op} < FUmax_{op}$
      $ACCOMP_{fucurrent}$ = ACCOMP_LOGIC($fu_{current}$)
      $DP \mathrel{+}= fu_{current} + ACCOMP_{fucurrent}$

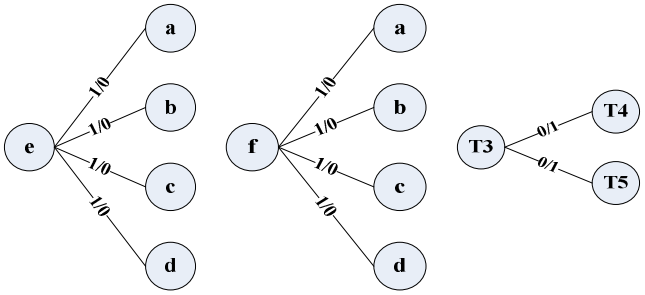
It is based on analysis of maximum needed operation instances for all basic blocks schedules. The contributions of all basic blocks are integrated to the final datapath through their functional units and all components that are connected to those functional units. When number of particular operation instances implemented within existing functional units in the final datapath is exceeded, those types of functional units are not further instanced.

As first, function GET_MAX_INSTANCES() analyses operation usages per basic blocks schedules. As result, maximum numbers of operations per scheduled state is noted - $FUmax_{op}$. Basic blocks are traversed and functional units are sorted by usage frequencies and integrated into final datapath along with their accompanying logic ($ACCOMP_{fucurrent}$); register files, multiplexers and connections. Data memory is instantiated for data arrays used in input C code. Before integration of new functional unit to the final datapath *DP*, number of functional unit instances per operation type is checked against calculated $FUmax_{op}$ value.

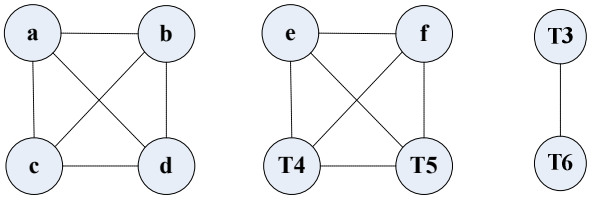
IV. PRELIMINARY RESULTS

Our approach of processor architecture modeling and implementation is tested on following C coded algorithms used inside NISC toolset, [14]:

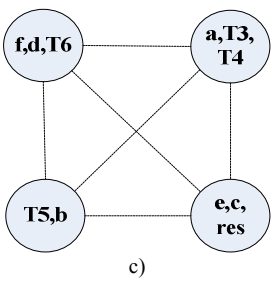
- Discrete Cosine Transform (DCT) on 8×8 matrices in two versions: original (with three nested loops) and unrolled, [6].
- 32-point DCT used in MP3 decoder.
- SHA-1 encryption algorithm, [15,16].



a)



b)



c)

Fig. 4. Initial priority edges (a), incompatibility edges (b) and final compatibility graph (c) for operands

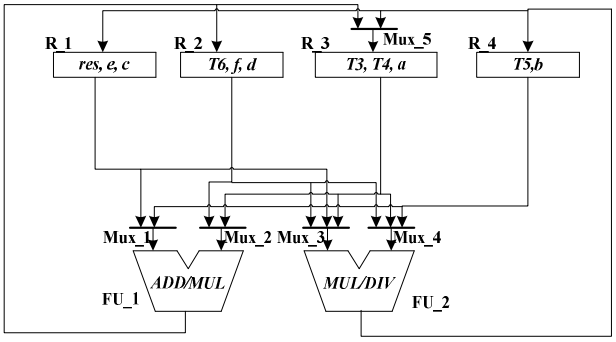


Fig. 5. Final datapath for expression in (1).

TABLE I. TEST CASES CDFG CHARACTERISTICS

Case	CDFG Characteristics			Design Time	Exec. Time / µs
	#blocks	*#statements*	*#operands*		
DCT 8×8	15	28	29	< 1 s	190,0
Unrolled DCT 8×8	5	161	191	1 s	54,9
32-point DCT	1	791	890	25,8 h	8,2
SHA-1	31	158	150	1 s	67,6

Table I shows characteristics of test cases CDFGs: numbers of basic blocks, statements and operands. The first two, DCT and Unrolled DCT represent the same code written in different styles and elaborated in [6].

Table I also shows processor design times and their FPGA implementation execution times. While other three cases have design times of one second or less, 32-point DCT has design time of 25,8 hours. The more detailed analysis shows that the most of it, more than 99%, is spent in stage of operands usage analysis and optimizations, i.e. in handling operand compatibility graphs. Therefore, in next section we focus on minimization of this case design time.

When execution times are compared to those of NMIPS RISC architecture model, the baseline in [6], the execution times are very similar (150,2 µs for DCT 8×8, 51,6 µs for Unrolled DCT 8×8, 11,5 µs for 32-point DCT and 64,3 µs for SHA-1). The implementations of the same applications on embedded Microblaze processor [17] are much slower (516,2 µs for DCT 8×8, 932,5 µs for Unrolled DCT 8×8, 280,4 µs for 32-point DCT and 670,9 µs for SHA-1), and much faster when designed with Vivado HLS tool (9,4 µs for DCT 8×8, 2,0 µs for Unrolled DCT 8×8, 0,5 µs for 32-point DCT and 6,5 µs for SHA-1), [18].

V. CASE STUDY: 32-POINT DCT

32-point DCT used here is optimized version of such transformation, but here it is the largest code without control flow dependencies. Inside only one basic block there are total of 80 additions, 119 subtractions, 80 multiplications and 49 shifts, [19, 20].

There is total of 209 lines of code (LoC) in C code specification. There are eight versions of code granulations undertaken to check the design time change. Starting with original input code (*cdfg0*), the code is granulated from the perspective of the implemented flow user. In *cdfg1* the code is split in two, in *cdfg2* in three parts, etc. The extreme situation is when every line of code is a separate block (*cdfg7*). Table II summarizes features of all CDFG versions in aspects of three-address statements, operands and scheduled cycles per block. More detailed view on granulation points in input C code is illustrated in Fig. 6 which depicts relations between granulation steps.

The comparison of designed datapath and implementation results is presented in Table III. The granulation of code to smaller portions significantly impacts the design time. After two steps of code granulation design time falls to below 20 minutes, and after following two it falls below a minute. The side-effect of such granulation sequence is architecture datapath and FPGA implementation resources occupations growths. This is caused by the fact that final datapath algorithm limits the functional units instantiation, but it is not the case for register files. Therefore, the FPGA resource occupations significantly rise and, in the same time, performance drops. The size problem escalates for *cdfg6* and *cdfg7* for which design could not be synthesize on Virtex-5 SX50T device that was targeted.

VI. CONCLUSION

This paper presents C-to-hardware design flow where processor architecture abstraction level is kept as important abstraction in design representation. The processor is modeled as No-Instruction-Set Computer with fully custom datapath. The datapath is customized according to scheduled three-address code, operand and operation usage analysis and their optimized binding to register files and functional units.

The preliminary results show successful FPGA implementation with performance in a range or better than processor based system design, but worse than high-level synthesis tool. The design time as the key strength in system level design appeared to be too high for test case without control flow dependencies and large number of lines. The processing of such code, 32-point DCT, took

TABLE II. 32-POINT DCT CDFGS FEATURES

Case	#blocks	#statements /block	#operands /block	#cycles /block
cdfg0	1	791	890	418
cdfg1	2	352, 459	445	177, 242
cdfg2	3	256-269	269-355	129-150
cdfg3	5	96-269	100-269	50-150
cdfg4	11	24-159	24-159	14-90
cdfg5	19	24-87	24-159	14-50
cdfg6	46	7-47	7–53	5-26
cdfg7	209	1-20	3–21	1-15

TABLE III. 32-POINT DCT DESIGNS CHARACTERISTICS

Case	Datapath Features		Design Time	FPGA Resource Occupation		Performance	
	#FUs	*#RFs*		*#Slices*	*#DSPs*	*#Cycles*	*Frequency / MHz*
cdfg0	5	29	25,8 h	1577	3	426	52,018
cdfg1	30	98	1,3 h	3353	3	433	46,517
cdfg2	29	117	17,2 min.	4107	3	439	45,486
cdfg3	26	107	6,3 min.	4780	3	454	41,075
cdfg4	14	106	41 s	5052	3	501	40,098
cdfg5	12	104	5 s	5681	3	559	35,088
cdfg6	17	147	2 s	N/A	N/A	N/A	N/A
cdfg7	4	149	5 s	N/A	N/A	N/A	N/A

hours of time. There were two options to shorten the design time: improvements in data structures and algorithms implemented inside tool or reorganization of input C code by breaking it into smaller portions. From the perspective of system designer that uses the implemented flow the latter, i.e. the specification alteration, is only possible impact on design time.

The changes in input code granulation significantly shorted the design time. As the code was granulated to smaller portions, the design time rapidly decreased until the level of only few seconds. The side-effect of such code granulation was in the growth of resulting datapaths and, consequently, the target FPGA device occupation. Also, the performance dropped twice.

Therefore, the future work in sense of shortening design time has following options: better algorithmic manipulation and usage of more appropriate data structures for storing compatibility graphs, and more intelligent C code preprocessing when large CDFGs are produced. Also, the consequential design size growth demands the control over register files datapath instantiations.

ACKNOWLEDGMENT

This work was supported by research grant No. 036-0362980-1929 from the Ministry of Science, Education and Sports of the Republic of Croatia. Special thanks we give to Center for Embedded Computer Systems (CECS) at University of Irvine California as the key support was based on prof. Daniel Gajski's team research and experience.

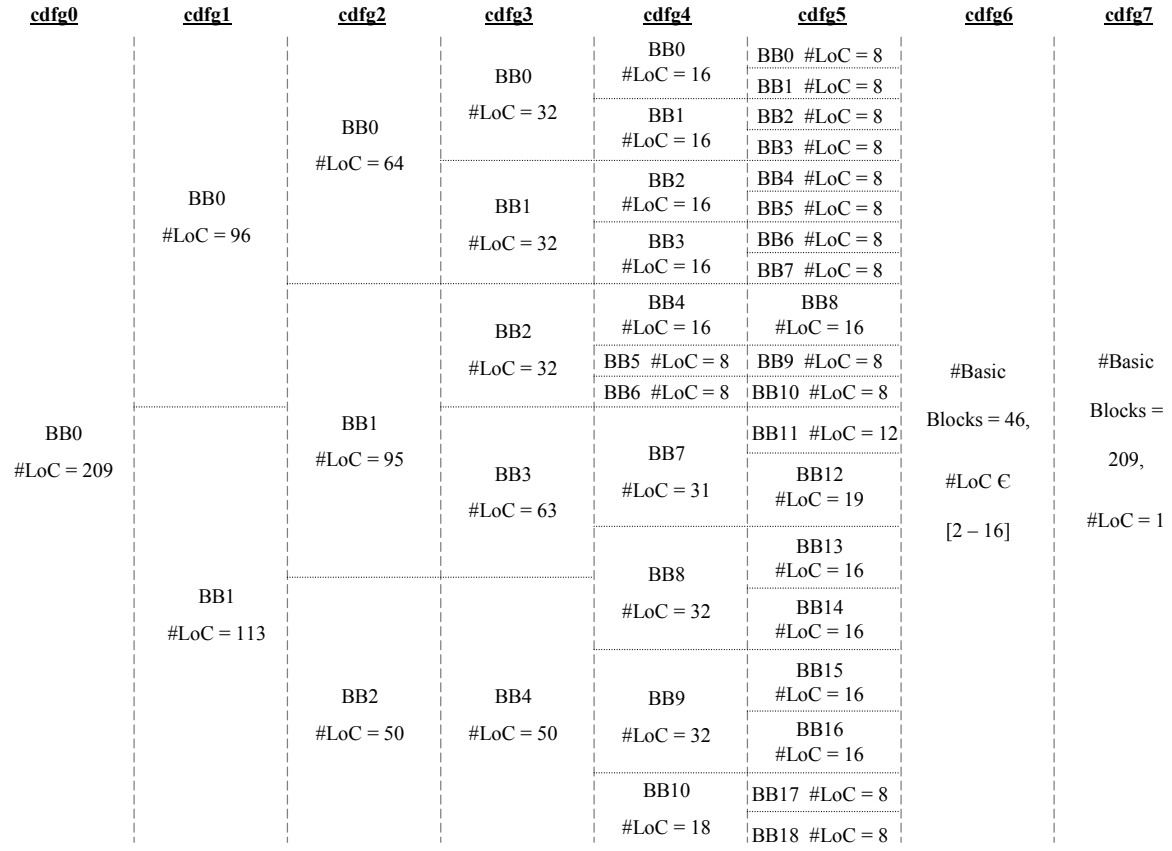


Fig. 6. Overview of 32-point DCT code granulation

REFERENCES

[1] W. Ecker, W. Muller, R. Dömer, Hardware-dependent Software: Principles and Practice, Springer, 2009.

[2] (2013) C to HDL. [Online]. Available: http://en.wikipedia.org/wiki/C_to_HDL/

[3] (2013) Gary Smith EDA. [Online]. Available: http:// http://garysmitheda.com/

[4] M. Reshadi, B. Gorjiara, D. D. Gajski, "NISC Technology and Preliminary Results," Technical Report, University of California, Center for Embedded Computer Systems, Irvine, August 2005.

[5] P. Coussy, D. D. Gajski, M. Meredith, A. Takach, "An Introduction to High-Level Synthesis," in *IEEE Design & Test of Computers*, vol. 26, no. 4, pp. 8-17, July-August 2009.

[6] B. Gorjiara, D. D. Gajski, "Custom Processor Design Using NISC: A Case-Study on DCT algorithm," in *Workshop on Embedded Systems for Real-Time Multimedia*, pp. 55-60, 2005.

[7] J. Trajkovic, D. D. Gajski, "Custom Processor Core Construction from C Code," in *Proceedings of Sixth IEEE Symposium on Application Specific Processors*, Anaheim, California, June 2008.

[8] D. Ivosevic, V. Sruk, "Evaluation of embedded processor based BDD implementation," in Proceedings of the 33rd International Convention MIPRO, pp. 619-623, 2010.

[9] D. Ivosevic, V. Sruk, "Automated modeling of custom processors for DCT algorithm," in Proceedings of the 34th International Convention MIPRO, pp. 762-767, 2011.

[10] A. Orailoglu, D. D. Gajski, "Flow graph representation," in Proceedings of the 23rd ACM/IEEE Design Automation Conference, pp. 503 - 509, 1986.

[11] (2013) SPARK: A Parallelizing Approach to the High-Level Synthesis of Digital Circuits. [Online]. Available: http://mesl.ucsd.edu/spark

[12] (2013) ISE WebPACK Design Software. [Online]. Available: http://www.xilinx.com/products/design-tools/ise-design-suite/ise-webpack.htm

[13] D. D. Gajski, A. Gerstlauer, S. Abdi, G. Schirner, *Embedded System Design*, Springer, 2009, pp. 199-254.

[14] (2013) NISC Technology – Toolset online demo. [Online]. Available: http://www.ics.uci.edu/~nisc/toolset/

[15] (2013) National Institute of Standards and Technology (NIST), "Secure Hash Standards (SHS)". [Online]. Available: http://csrc.nist.gov/publications/fips/fips180-4/fips-180-4.pdf

[16] (2013) C to Verilog. [Online]. Available: http://c-to-verilog.com/howtos.html

[17] (2013) MicroBlaze Soft Processor. [Online]. Available: http://www.xilinx.com/tools/microblaze.htm

[18] (2013) C-based Design: High-Level Synthesis with Vivado HLS. [Online]. Available: http://www.xilinx.com/training/dsp/high-level-synthesis-with-vivado-hls.htm

[19] (2013) mbed: i2s_audio_madplayer. [Online]. Available: http://mbed.org/users/okini3939/code/i2s_audio_madplayer/docs/30b2cf4a8ee2/synth_8cpp_source.html

[20] (2013) libmad - MPEG audio decoder library. [Online]. Available: ftp://ftp.icsi.berkeley.edu/global/pub/speech/software/praatlib-0.3/src/mp3/mad_synth.c

Numerical Relay Protection Coordination using Simulation Software

S. Nikolovski*, M. Havranek , P. Marić*
* Faculty of Electrical Engineering Osijek, Croatia
srete.nikolovski@etfos.hr
mario.havranek@gmail.com
predrag.maric@etfos.hr

Abstract - The paper presents protection coordination of renewable energy source connected to bilaterally supplied distribution network. Multifunctional numerical relays with the switching equipment, renewable energy source facility and the distribution network have been detail modeled in DIgSILENT Power Factory software. Overcurrent, overvoltage, undervoltage, directional overcurrent, underfrequency and overfrequenecy functions have been used for protection coordination. Different relay characteristics have been presented in one time-overcurrent plot. The damage curve and characteristic currents of electrical equipment in the network have been shown to set the relay tripping times and current settings and for a good and thorough protection of the equipment. Protection coordination settings have been checked and verified simulating faults with the real system values.

I. INTRODUCTION

Power system protection is a task of challenge for power engineers. For better understanding such a task, power system simulation tools are used. The paper presents numerical protection adjustment in simulation tool - *DIgSILENT Powerfactory* with own database of numerical relays that contains *time-curent curves* of all protective devices. The *time-current curves* (TCC) can be coordinated with each other and with the short circuit analysis results. Multiple easy-to-use graphical tool for resolving protection devices coordination problems has also included in this simulation tool. The sample case of the numerical protection adjustment is presented on the small power plant (renewable source) connection to the bilaterally supplied distribution network Fig 1. ABB numerical relays and Končar RFX numeric field terminal have been used in simulations.

The *DigSILENT Power Factory* protection modelling features have been implemented with the following philosophy: the protection should be as realistic as possible; the user has to be able to create a new complex protection devices or alter existing ones. Although the protection models may show high complexity, their use must be kept easy. All protection models will act on circuit-breakers [1]. Protection devices are normally stored in the object which they act upon, but they may be stored elsewhere when needed. Recommended, and by default is that protection devices which act upon a single circuit-breaker are store in the cubicle which contains that circuit-breaker (highly recommended). Protection devices which act upon a two or more circuit-breakers connected to the same busbar are stored in that busbar [2]. As a rule, the relay is best stored in the same folder as the e and/or current transformers which it uses.

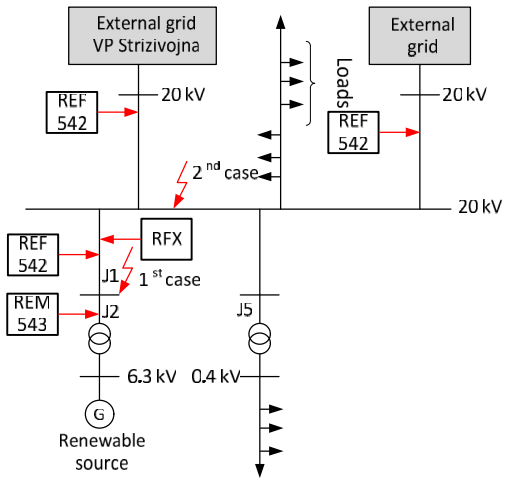


Figure 1. Small power plant (renewable source) connection to distribution network

II. RELAY MODEL CREATION

The simplest way to create the relay model is normally done by right clicking a circuit-breaker and selecting *New Protection device-Relay model,* an empty relay model (*ElmRelay*) will appears. The relay model has a reference to the relay type, the location, the device number and the list of slots. The location is normally set automatically when the relay is defined in the single line diagram by right clicking the cubical [1]. The whole process of creating a specific relay model thus only asks for selecting a relay type from the database, in this case *ABB numerical relays.* Otherwise, a specific unknown relay type (according to the database) can be created using advanced *DIgSILENT Powerfactory* functions such as DSL, or using the database relay type with the same functions and adaptable function settings. The Končar RFX numerical field terminal model has been created using multifunctional ABB REF 54x relay model template from the database applying the overcurrent function settings only. Fig. 1 describes the protection coordination scheme for the modelled system. The ABB REM 543 relay has been installed on J2 field for

generator and block-transformer primary protection (acts on circuit-breaker on J2), the ABB REF 542 has been installed on the J1 field for protection of cable between small power plant and 20 kV switchyard (acts on circuit-breaker on J1), the Končar RFX field terminal has been installed on 20 kV switchyard. Relays installed on External Grids (ABB REF 542) acts on corresponding circuit-breakers of external grid facilities.

A. Overcurrent protection adjustment

The function of this protection is single-phase, two-phase or three- phase overcurrent detection [2]. After the relay model has been defined and all slot elements have been created, the editing of the relay settings may be started. The overcurrent protection includes: the time-overcurrent protection and the instantaneous overcurrent protection. The time-overcurrent protection (relay) allows settings: the time-overcurrent characteristic, the pick up current and the time dial. The time-overcurrent characteristic use the time delay which can be an independent (definite) time or inverse time delay. In this case, the time-overcurrent characteristic is the *definite time characteristic* with an option to time dial set and to choose the time-overcurrent characteristic. Fig. 2 presents the time-overcurrent protection adjustment, while Fig. 3. presents instantaneous overcurrent protection.

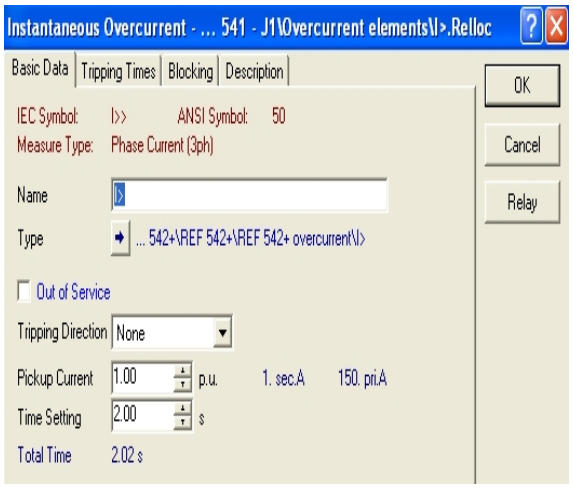


Figure 2. Adjusting the time-overcurrent protection

The pick up current is set to 1.00 p.u secondary which equals 150 A primary and the *definite time* is set to value of 2.00 seconds (Fig. 2) .On the Fig. 3. the pick up current is set to 1.80 p.u secondary which equals 270 A primary. The time setting on Fig. 3 is set by default to 0.00 seconds which corresponds to instantaneous operation.

In the network supplied by two power sources is necessary to use *Directional relay* – a part of the relay model [1]. The directional relay calculates the angle between the polarization voltage or current and the operating current. The polarization current or voltage is rotated for the amount of the expected angle first.

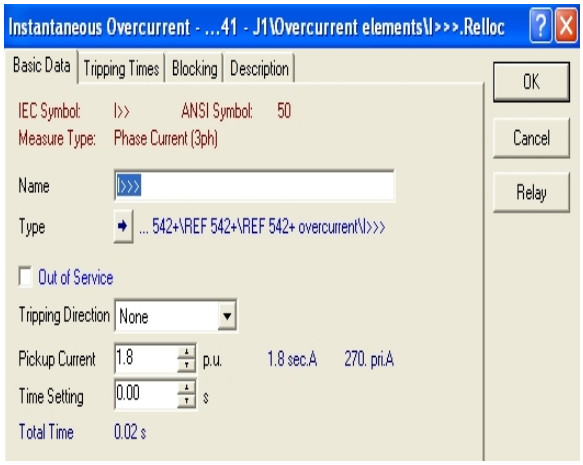


Figure 3. Adjusting the instantaneous overcurrent protection

The relay trips if the remaining angle is smaller than 90° and if both the polarization and the operating voltage/current are large enough [2]. The directional overcurrent protection adjustment has been presented on Fig.4.

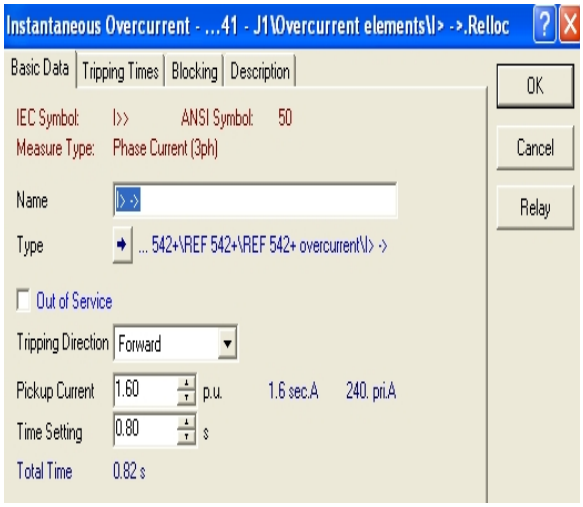


Figure 4. Adjusting the directional overcurrent protection

The tripping direction is an important fact for this type protection adjustment because of the current income during the short circuit.

B. Undervoltage and overvoltage protection adjustment

The voltage drops can occur due to the network overload, the faulty operation of a transformer tap changer and during the short circuit. In this case undervoltage and overvoltage protection monitors any phase to phase voltage and will act on 20 kV circuit-breaker if the voltage drops or rise under/over certain limit that will lead to separate the connection between the power plant and the rest of the system. The adjustment of the undervoltage protection is presented on the Fig. 5 and Fig. 6.

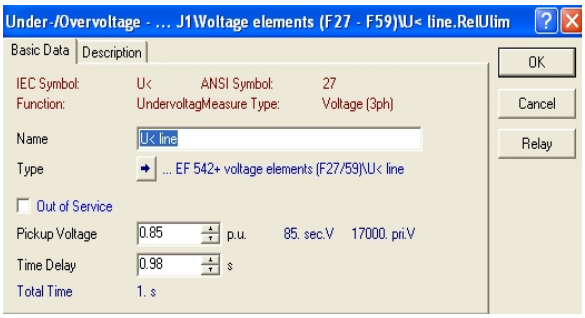


Figure 5. Adjusting the undervoltage protection

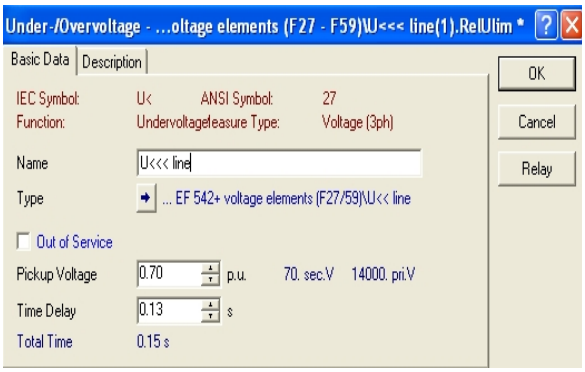


Figure 6. Adjusting the undervoltage protection (fast)

On the other hand, the protection is necessary during the voltage increase over the allowed limit. Adjustment of the overvoltage protection is shown on Fig. 7 and Fig.8.

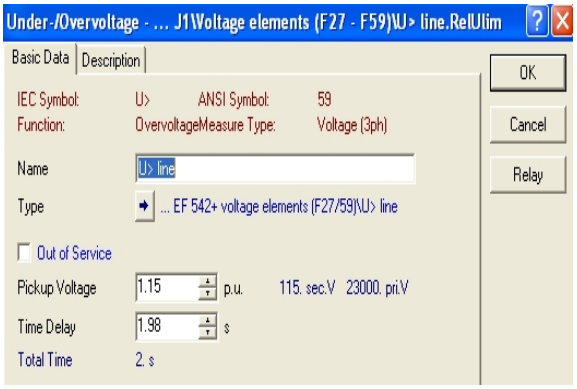


Figure 7. Adjusting the overvoltage protection

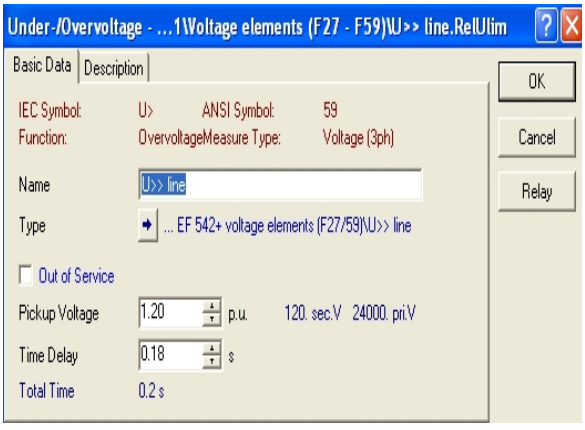


Figure 8. Adjusting the overvoltage protection (fast)

C. *Underfrequency and overfrequency protection adjustment*

Variations in the power supply frequency can occur due to overloads when the network is fed by a limited power source (power plant islanding operation), generator frequency regulator faulty operation, power plant disconnection from the interconnected network [2]. Adjustment of the underfrequency and overfrequency protection is shown on Fig. 9 and Fig. 10.

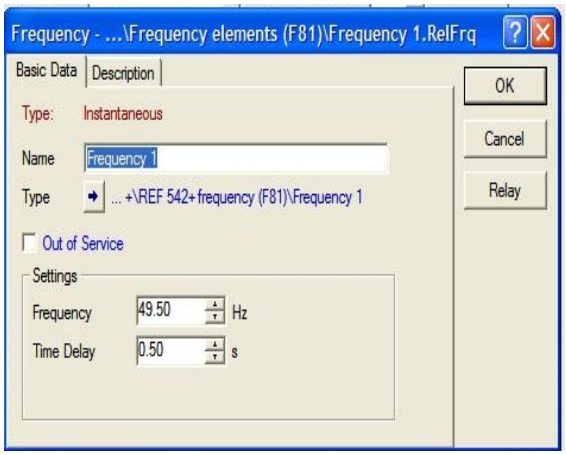


Figure 9. Adjusting the underfrequency protection

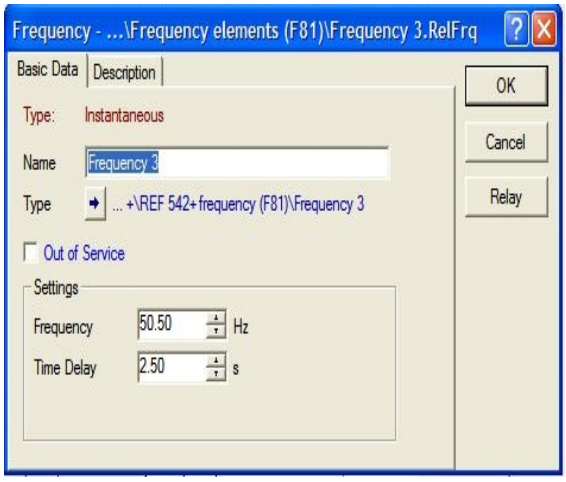


Figure 10. Adjusting the overfrequency protection

III. TIME-OVERCURRENT PLOT CREATING

The plot *VisOcplot* is showing different relay and fuse characteristic in one time-overcurrent plot. Additionally the damage curve and characteristic currents of electrical equipment in the network can easily be shown. This will help to set the relay tripping times and current settings and the selecting of fuses for a good and thorough protection of the equipment.

There are several ways to create a time-overcurrent plot (*VisOcplot*). The easiest way to create and show a VisOcplot is to select one circuit-breaker, where overcurrent relays are installed than right-click the circuit-breaker to open the context sensitive menu. This will show the option *Create Time-Overcurrent Plot* and *Add to Time-Overcurrent Plot. PowerFactory* will then create a

new diagram showing the time-overcurrent plot for all relays selected. It is also possible to create an a user defined "x-value" by right-clicking the graph and selecting the *Set Constant* and *x-value* option. The vertical line will show the values at the intersections of all characteristics. The line movement can be realised by mouse dragging. The intersection of the calculated current with the time-overcurrent characteristic is labelled with tripping time.

The three- phase short circuit is simulated on the cable between small power plant and 20 kV switchyard (the first simulation case - location J1 on Fig.1) The vertical red-coloured curve on the Fig. 11 represents the short circuit current value on the short circuit location. The first relay that will trip is the relay situated in the cubicle of the J1-field -*ABB REF 542* and also the nearest to the short-circuit location. The tripping time for this relay shown on the Fig.11 (red -coloured curve) equals 0.02 s. The second tripping time according to this time-overcurrent plot belongs to the relay situated on the 20 kV switchyard, the RFX numerical terminal filed, shown as dark-blue-coloured curve with the tripping time of 0,04 s. Tripping times of those two relays indicate good protection coordination because the short-circuit location is hereby isolated from the rest of the power system.

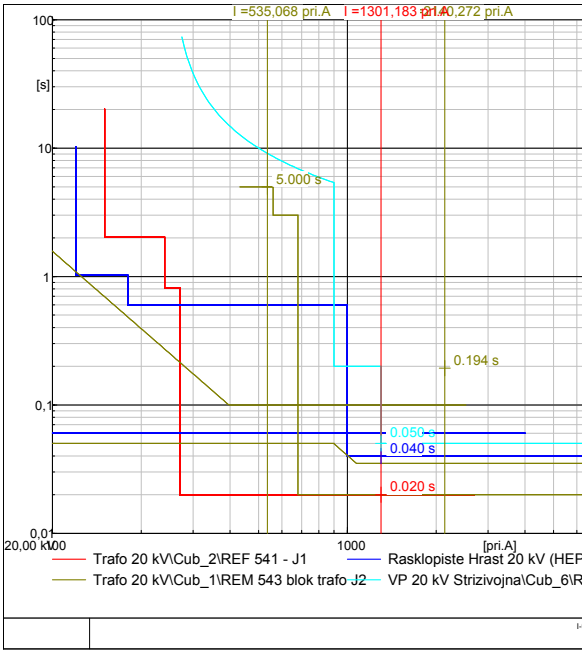


Figure 11. Time-overcurrent plot with tripping times (short circuit on the J1- field)

The time-overcurrent plot on Fig.12 shows the three phase short-circuit simulation results on the power plant place of coupling to distribution network - 20 kV switchyard (the second simulation case – Fig.1). In this case, the network contribution to short circuit is 1302.703 A primary while the contribution of the small power plant (renewable source) is 534.935 A. The relay situated on location – J1 *REF 542* will trip instantaneously with the time of the 0.020 s (red -coloured vertical line on Fig. 12), while the remote relay *REF 542* situated on the external network busbars will trip instantaneously with time of the 0.050 s (light -blue -coloured vertical line on Fig. 12). The rest of the distribution network and the small power plant are both isolated from the short circuit by tripping of those two relays. Such an action indicates good protection coordination.

The time- overcurrent plots may also be used to change the relay characteristic graphically because a relay characteristic is normally the minimum of two or more sub-characteristic [1].

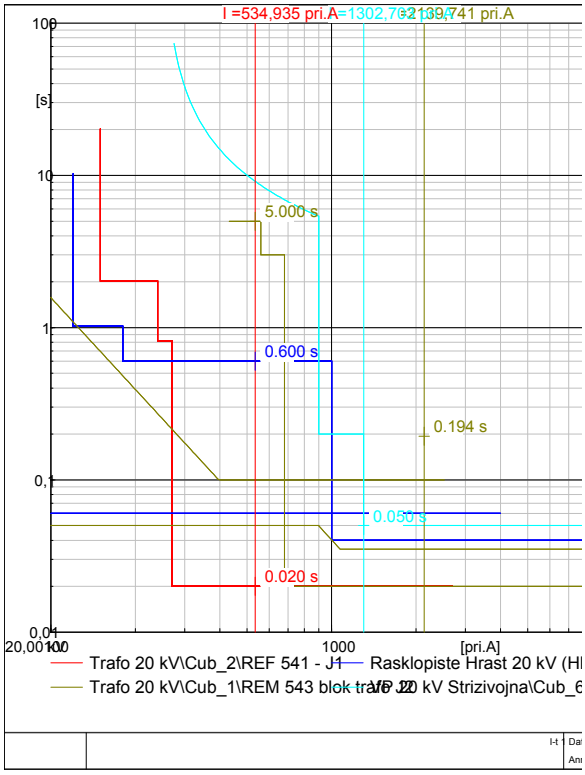


Figure 12. Time-overcurrent plot with tripping times (short circuit-on 20 kV switchyard)

The voltage protection coordination can be actuated disabling the current, the distance and the frequency functions on observed relays. The undervoltage protection acts on generator circuit-breaker in case when the voltage drops under the limited lower voltage value. Typically, the lower voltage value on the generator relay is set to 85% of nominal value. The generator block-transformer relay allows the fast undervoltage tripping function with the time delay of 0,15 s for adjusted 70% of nominal voltage value and the slow undervoltage tripping function with the time delay of 1,00 s for adjusted 85% of nominal voltage value. The overvoltage protection acts on generator circuit-breaker in case when the voltage rises over the limited upper voltage value. The fast tripping function acts with the time delay of 0,18 s when the measured voltage increase above the 120% of nominal value, while the slow tripping function acts with the time delay of 1,98 s when the measured voltage increase above the 115% of nominal value. Islanding operation of small power plants is prohibited according to grid code. The voltage drop on the generator block transformer LV and HV busbars during the three phase short-circuit on the J1 field is shown on Figure 13.

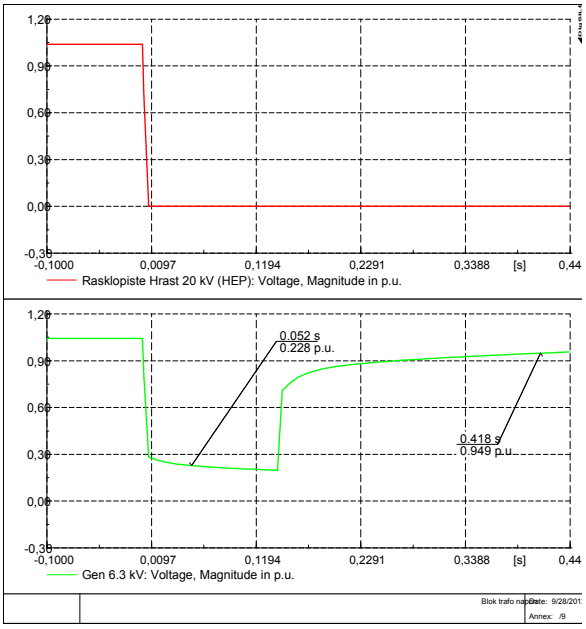


Figure 13. Voltage drop on HV (upper diagram) and LV (lower diagram) generator block-transformer busbar

During the voltage drop on the generator busbars, the undervoltage protection trip and the small power plant is isolated from the rest of the system.

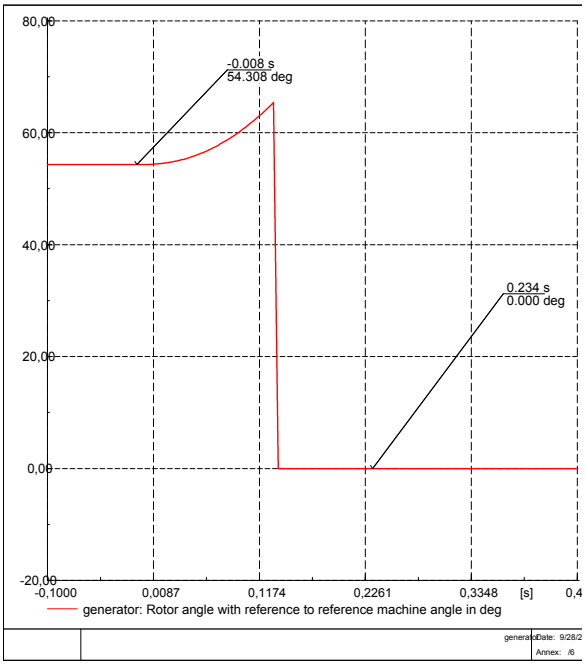


Figure 14. Small power plant generator isolation

IV. CONLUSION

The main purpose of this paper is to present the usage of the simulation tool for numerical protection functions adjustment. *DIgSILENT PowerFactory* simulation software enables detailed modelling and assistance for coordination settings of real digital numerical relays according to calculations of short circuit using method of symmetrical components and the real data for all network components. The numerical protection functions have been tested during the short-circuit simulations on the two network locations that are significantly sensitive to power system protection settings. Protection functions adjustments shown good protection coordination in both analysed cases.

REFERENCES

[1] DIgSILENT Power Factory, User Manual, Gomoringen, 2008.

[2] S. H. Horowitz, A. G. Phadke, Power system relaying, 3rd ed., John Wiley and Son Ltd, England, 2008.

[3] Technical Reference Manual, REF 54_, ABB, 2005.

[4] Pilot Protective Relaying, ABB, Marcel Dekker ,Inc.; USA, 2000.

[5] A.T. Johns and S. K. Salman, Digital Protection for Power Systems IEEE Power Series 15, London ,1995.

[6] M. Sachdev, M. Nagpal, A recursive least error squares alghortihm for power system relaying and measurements applications, IEEE Transanctions on Power Delivery , Vol. 6, No. 3. 1991, pp. 1008-1015

[7] ANSI/IEEE C37.110 Standard : IEEE Guide for the application of current transformers used for protective relaying purposes

Engine Fault Detection for Piston Engine Aircraft

Dubravko Miljković
Hrvatska elektroprivreda, Zagreb, Croatia
dubravko.miljkovic@hep.hr

Abstract - Digital engine monitor can record vast amount of data in form of engine parameters from the aircraft piston engine. By analyzing these parameters it should be possible to detect majority of current or impeding engine problems. Statistical description of engine parameters together with rule based pattern recognition of catalogued graphic engine fault patterns shown as bar graphs on the engine monitor display may help detect abnormal engine conditions. Method for automatic analysis of engine monitor data and labeling potential problems is described.

I. INTRODUCTION

Almost all general aviation aircrafts (with the exception of business jets) are powered by engine–propeller combination. In most single-engine light aircraft, the power plant is a four-stroke reciprocating engine with a direct drive to a propeller. Aircraft piston engines are relatively reliable devices. Engine failures are rare, but do happen. In case of engine failure it is possible to land an aircraft, but this is very risky event, particularly if engine failure happen over inhospitable terrain, in IFR (instrumental) conditions or at night.

II. PISTON ENGINE AND PROPELLER COMBINATION

Thrust necessary for flight is generated by engine-propeller combination.

A. Piston Engine

Piston engine is an economical source of power for small (general aviation) aircraft due to its power output, price and fuel consumption at cruise speed of a typical general aviation aircraft. Essentially it is a heat engine that uses one or more reciprocating pistons to convert pressure into a rotating motion. Most common aircraft piston engine are horizontally opposed, air-cooled four-, six- and eight-cylinder engines. Engines have old fashion but reliable fixed-timed dual magneto ignition systems (no use of electronics).

B. Propeller

Propellers are generally directly driven by an engine, although gearbox speed reductors exist on some higher power engines. Most training aircraft employ fixed-pitch propeller while complex aircrafts and trainers use constant speed propeller. Constant speed propeller is necessary to cope efficiently with this with variations in speed range and engine power. As aircraft speed increases, so does propeller efficiency, up to a peak. However, at faster speeds, efficiency reduces. By varying the pitch, it is possible to extend this maximum efficiency over a greater speed range. Also, as airplane climb higher, the density of the air decreases. To maintain speed, a coarser blade angle must be taken.

C. Engine-Propeller Power Management

Aircraft engine-propeller combination (with constant speed propeller) is controlled by three levers: thrust, propeller pitch and mixture. Thrust controls manifold pressure (MAP), propeller pitch controls RPM (revolutions per minute) and mixture controls fuel pressure and indirectly fuel flow. Manifold pressure (MAP) is the actual pressure in the inlet manifold in inches of mercury (Hg), measured by a sensor downstream of the throttle plate.

III. ENGINE MONITOR

The main source of engine information available to pilot are several gauges indicating engine rotational speed (RPM, tachometer), oil pressure, oil temperature, exhaust gas temperature (EGT) and fuel flow. These gauges give very basic information about the engine condition. More advanced solutions exist today in form of engine monitors. Such engine monitors cover much more engine data then basic gauges in a cockpit (about dozen of parameters that are also recorded and can be analyzed later), Fig. 1. Adequate skill is needed for correct engine monitor data interpretation. Beside engine condition monitoring, these engine monitors can be used for improved engine operation (fuel economy). Piston engine is not particularly efficient and only a small portion of the energy from combustion produces movement of the piston during the power stroke. The greatest part of energy passes into the exhaust pipe as hot gasses. By monitoring the temperature of exhaust gasses it is possible to assess the quality of the combustion process. Diminished efficiency of the

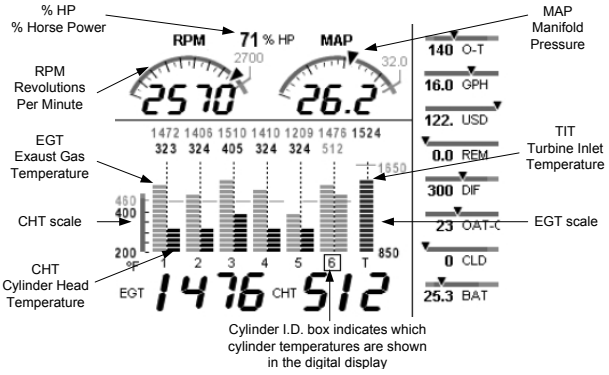


Figure 1. Engine monitor display (EDM 830)

TABLE I Monitored engine parameters

Parameter	Description
EGT	Exhaust Gas Temperature
CHT	Cylinder Head temperature
OIL TEMP	Oil Temperature [1]
OIL PRES	Oil Pressure [1]
TIT 1	Turbine Inlet Temperature 1[1]
TIT 2	Turbine Inlet Temperature 2 [1]
OAT	Outside Air Temperature
CDT	Compressor Discharge Temperature [1]
IAT	Intercooler Air Temperature [1]
CRB	Carburetor Air Temperature [1]
CDT - IAT	Intercooler cooling
RPM	Rotations Per Minute
MAP	Manifold Pressure
% HP	% Horse Power
CLD	CHT Cooling Rate [2]
DIF	EGT Span [3]
FF	Fuel Flow [1]

[1]optional, [2]fastest cooling cylinder, [3]difference between the hottest and coolest *EGT*

combustion process that generates power indicates engine problems like low compression, non-uniform fuel distribution, faulty ignition, and clogged injectors, [1]. Engine parameters are collected from engine monitor probes installed on an engine. Engine monitor used in an experiment (EDM 830) records and calculates numerous engine parameters that are listed in Table I and shown in form of bar graph and digital display, Fig. 1. Parameters are displayed and recorded at the programmed interval (default setting is every 6 seconds). Depending on the monitored engine not all parameters are available. Other parameters (battery voltage, remaining fuel etc.) collected by the monitor are collected but are not relevant for consideration here. Example of engine monitor log is shown in Fig. 2. Graphical representation for main engine parameters through the duration of one whole flight is shown in Fig. 3 (upper curves represent *EGT*s and lower curves *CHT*s), [2]. Engine diagnosis charts supplied with the engine monitor describe engine fault patterns (shown

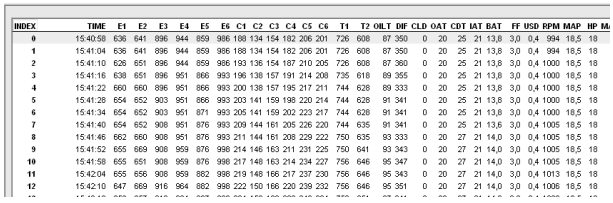


Figure 2. Engine monitor log

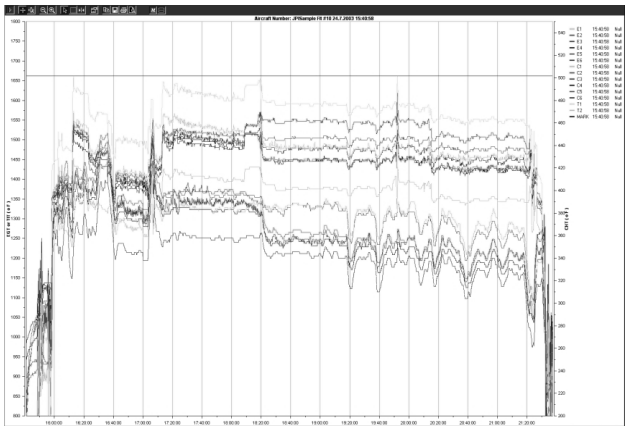


Figure 3. EzTrends engine parameter plots

in terms of bar graph on a display) can help diagnose various engine faults, [1]. There are 15 general patterns that indicate particular faults.

IV. Method for Fault Detection

Proposed method is combination of statistical and pattern recognition approach. It preserves default engine monitor alarm limits but adds new alarm limits corresponding to current engine working regime and augments it all with rule based fault pattern recognition, Fig. 4. Method is intended for parsing engine parameters log after the flight. Great problem with advising fault detection method is due to very reliable aircraft engines. It is very difficult, with exception of large manufacturer and overhaul services, to obtain sufficient large sample of failed engines. On the other hand artificial failures can be produced (failure injection), but this process could harm the expensive engine (some failures would require destructive testing with high price tag), yet it will not cover all problems.

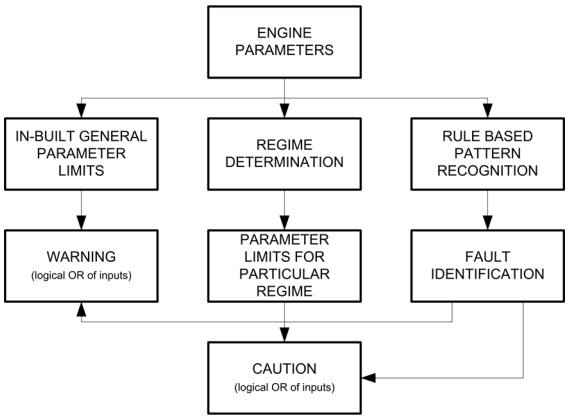


Figure 4. General description of the method

V. Statistics of Engine Parameters

Statistical analysis of engine parameters logs (included with EzTrends software: Flt#10 of duration 5.58 hours and Flt#11 of duration 3.69 hours, Continental TSIO-550?) was performed and Box-Wiskers plot was used for graphical representation of key values from summary statistics. Values represented in the summary are the mean, std. dev., minimum, maximum, 1th percentile and 99th percentile:

$$[\mu_{p_i}, \sigma_{p_i}, \min(p_i), \max(p_i), P1, P99] \quad (1)$$

The mean value for parameters p_i is determined by

$$\mu_{p_i} = \frac{1}{N_R}\sum_{j=1}^{N_R} p_{i,j} \quad (2)$$

and the standard deviation is determined by

$$\sigma_{p,i} = \sqrt{\frac{1}{N_R}\sum_{j=1}^{N_R}\left(p_{i,j} - \mu_{p,i}\right)} \quad (3)$$

where N_R is the number of records.

Percentile is the value of a variable below which a certain percent of observations fall. For example, the 99th percentile is the value below which 99% of the observations may be found. Percentiles are very suitable for exploring the distribution of number sets using various

TABLE II Engine parameters present in a log of analyzed engine

Parameter	Description
EGT 1-6	Exhaust Gas Temperature[1]
CHT 1-6	Cylinder Head temperature[1]
OIL TEMP	Oil Temperature[1]
TIT 1	Turbine Inlet Temperature[1]
TIT 2	Turbine Inlet Temperature 2[1]
CDT	Compressor Discharge Temperature
IAT	Intercooler Air Temperature
CDT - IAT	Intercooler cooling
RPM	Rotations Per Minute
MAP	Manifold Pressure
% HP	% Horse Power[2]
CLD	CHT Cooling Rate
DIF	EGT Span
FF	Fuel Flow

[1]included in a statistical summary (gray), [2]histogram of values determined

exploratory data analysis graphs (like Box-Wiskers plots) Parameters that have been present in available engine logs and are here considered are presented in Table II.

A. Analysis of the Whole Engine Log

Statistical results of engine parameters for the log that encompass duration of whole flight are presented in Table III and Fig. 5 (please note wide temperature scales and corresponding temperature ranges for various engine parameters). Maximal values correspond quite close to built-in (but modifiable) default alarm levels encompassing all engine regimes during a flight (one fits all approach), as shown in Table VI.

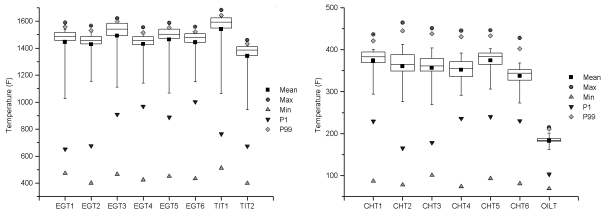


Figure 5. *EGT1-6*, *TIT1*, *TIT2*, *CHT1-6* and *OILT* for all regimes

TABLE III Statistical summary of engine parameters encompassing all engine regimes

Param.	Mean μ	Std.Dev.σ	Min	Max	P1	P99
EGT1	1445,3	167,9	473	1588	651,2	1558
EGT2	1427,4	146,3	400	1566	675	1531
EGT3	1491,4	145,9	466	1620	908	1599
EGT4	1429,5	109,1	424	1554	968,2	1515
EGT5	1463,7	139,1	451	1587	887	1553
EGT6	1442,6	111,6	433	1558	1002	1522
TIT1	1540,2	174,5	511	1682	764,4	1642,8
TIT2	1340,9	143,7	399	1459	673	1433
CHT1	373,6	35,3	87	436	229	421
CHT2	360,1	45,3	78	464	165,2	445
CHT3	356,1	42,5	101	451	178,4	438
CHT4	351,7	33,5	74	445	236	431
CHT5	374,1	33,2	93	446	240	433
CHT6	337,2	29,9	81	428	230	402
OILT	182,5	15,4	69	215	103,2	211
FF	10,17	3,41	1,3	27,8	3	22,5

B. Multiple Engine Regimes

In ideal case one would determine distribution of engine parameters for each engine working regime corresponding to various flight phases: engine run up, take-off, climb and full throttle operation cruise and descent. Engine, propeller and mixture combinations for various flight phases are presented in Table IV, [3].

TABLE IV Example of RPM/MAP power setting combinations for a typical GA aircraft

Flight phase	MAP	FF	Mixture	RPM range	Pitch Lever
Take off	28-30“	high	rich	2650-2700	fine
Climb	26-28“	high	BPM[1]	2600	fine
Cruise	20-24“	med.	BPM/ BEM[2]	1800-2400	medium -coarse
Appr.	18-22“	low	rich	1800-2400	med or out of CSU[3] range
Go around	as for takeoff				

[1]BPM – Best Power Mixture, [2]BEM – Best Economy Mixture, [3]CSU – Constant Speed Unit

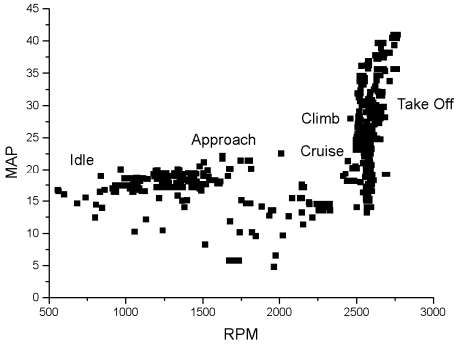


Figure 6. *RPM*/*MAP* combination for various flight phases

The concept of switching between multiple regimes is introduced to separate engine parameter statistics for each engine regime. One way to determine engine working regime would be from clustered *RPM*/*MAP* combinations, Fig. 6. However, in usual operations only small subset of these combinations is used, yet it may happen that for some reason pilot select unusual *RPM*/*MAP* combination lacking historical data in available engine logs (e.g. outside clustered areas). Because engine operation is of central consideration here, engine regimes are determined from calculated percent of the maximal horse power (% *HP*). This is simple and logical choice instead of using more complex multivariate clustering based techniques (e.g. using *RPM*, *MAP*, *OAT* and *FF*). % *HP* is already calculated by engine monitor from *RPM*, *MAP*, *OAT* and *FF*. Resulting percentage of Horse Power (% *HP*), is used as a proxy for determining engine regime. Fig. 7 illustrates calculated % *HP* as a function of RPM and MAP values. Histogram of statistical distribution for % *HP* is shown in Fig. 8. Same data presented in six bins corresponding to six engine regimes is shown in Fig. 9. Engine regime selection is shown in Table V and Fig. 10.

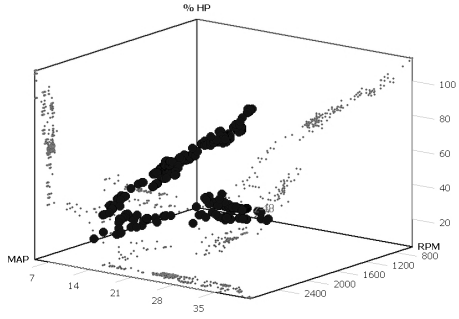


Figure 7. % *HP* as a function of *RPM* and *MAP*

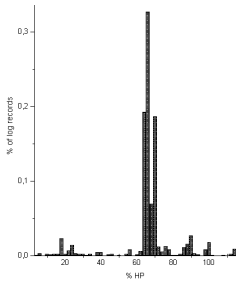


Figure 8. % *HP* histogram

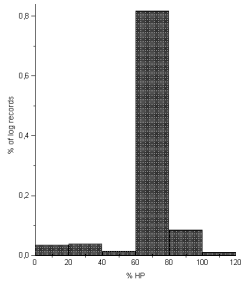


Figure 9. % *HP* histogram in 20% bins

TABLE V Engine regime selection by % *HP* value

% *HP*	Engine regime
0 – 19	1
20 – 39	2
40 – 59	3
60 – 79	4
80 – 99	5
100 - 119	6

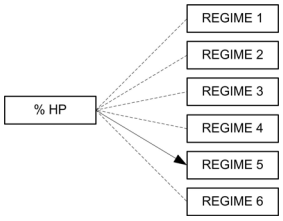


Figure 10. Engine regime selection using % *HP*

To counteract effect of short application of power bursts by engine throttle that don't have immediate effect on engine temperatures (that change more slowly), moving average of several samples (e.g. N=5) may be applied to calculated % *HP* values P_i as a form of low pass filter giving averaged power $\hat{P}$ and consequently regime r:

$$\hat{P}\frac{1}{N}\sum_{i=1}^{N}P_i \tag{4}$$

$$r_i = \left\lceil \frac{\hat{P}}{20} \right\rceil + 1 \tag{5}$$

Statistical plots engine parameters for various engine regimes are shown in Fig. 11-16. Dependence of engine parameters on calculated % *HP* (P_i) is clearly evident (please note different auto scales for temperature axis).

VI. Alarm Limits

The EDM has several programmable owner-programmable exceedance settings for all parameters. Exceedance warnings are both visual and aural. When a parameter falls outside normal limits, the display flashes its value and acronym. Once the parameter value returns within its normal limits, the flashing stops.

A. Default Alarm Limits

Default alarm limits are conservatively set (by JPI) below engine manufacturers (Lycoming and Continental) recommendations, Table VI. If a parameter gets out of its normal limits, the digital display will blink indicating the value and abbreviation of the problematic parameter. Because the temperature values of *EGT*s can assume different ranges depending on the current flight phase (climb, cruise) or engine run-up, monitor doesn't provide

TABLE VI Default Engine Monitor Alarm Limits

Measurement	Default Low Limit	Default High Limit
CHT		450 °F 230 °C
OIL	90 °F 32 °C	230 °F 110 °C
TIT		1650 °F 900 °C
CLD		-60 °F/min -33 °C/min
DIF		500 °F 280 °C
MAP		32 inch Hg

alarm limits for individual *EGT*s, it calculates the *DIF* parameter instead. The value for *DIF* is the difference between the hottest and coolest *EGT*s. This EGT span is important for monitoring the values of *EGT*s, [2]. Default alarm limits are set to encompass all flight regimes ("one fits all"). Engine monitor used in an experiment has default settings for limits shown in Table VI. Simple limit checking is used for exceedance warnings. Two limit values, thresholds, are present, a maximal value Y_{max} and a minimal value Y_{min}. A normal state is when

$$Y_{\min} \le Y(t) \le Y_{\max} \tag{6}$$

Big advantage of limit checking is its simplicity and reliability, [4]. Maximal value Y_{max} and a minimal value Y_{min} are determined from parameter statistics

$$L_{L,i} < p_i < L_{H,i} \tag{7}$$

where

$L_{L,i}$ is low limit for parameter p_i
$L_{H,i}$ is high limit for parameter p_i
p_i is engine parameter i

Most parameters don't need lower limit and only high limit is used (e.g. temperature too high). Oil temperature and fuel flow need both limits (low oil viscosity al low temperature and abnormal fuel consumption). Alarm levels for warning alert (that require immediate crew awareness and corrective action) commonly set at the engine monitor are universal for all phases of flight and provide detection of serious faults.

B. Engine Regime Dependent Alarm Limits

Determination of the appropriate threshold could be quite difficult task [5, 6]. In this method it is supposed that record form the engine log should be closer examined if the value of engine parameter falls above value of 99% percentile or below value of 1% percentile if lower limit is used for that parameter (both rare events). This choice is experience based, considering the tradeoff in accuracy, [6]. Fine detection suitable for caution alert (corrective action may be required) is achieved by statistical analysis of engine parameters within each engine regime. Parameter range that is acceptable for one engine regime may be different for other regimes. Engine parameter statistics is collected and analyzed. Alarm limits are determined from extracted statistical data supplied in engine monitor logs. Limit values (1 and 99 percentiles) for parameters p_i and regime r are shown in Table VII. *OILT* and *FF* use both limits. Just upper limit is used for *EGT*, *TIT* and *CHT*.

$$L_{L,i,r} < p_i < L_{H,i,r} \tag{8}$$

$$p_i < L_{H,i,r} \tag{9}$$

where $L_{L,i,r}$ and $L_{H,i,r}$ are low and high limits for regime r.

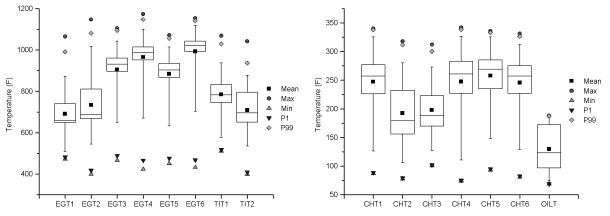


Figure 11. *EGT1-6*, *TIT1*, *TIT2*, *CHT1-6* and *OILT* for regime 1

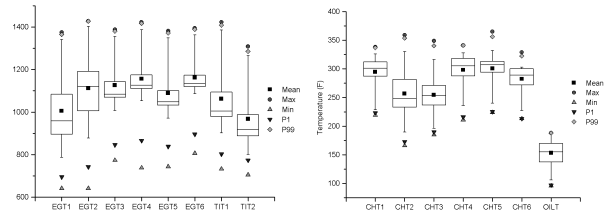


Figure 12. *EGT1-6*, *TIT1*, *TIT2*, *CHT1-6* and *OILT* for regime 2

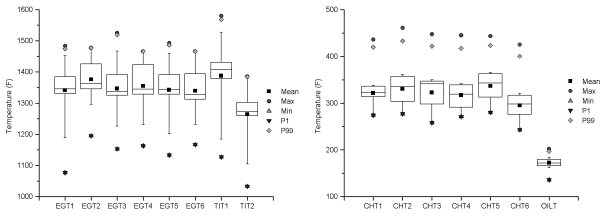


Figure 13. *EGT1-6*, *TIT1*, *TIT2*, *CHT1-6* and *OILT* for regime 3

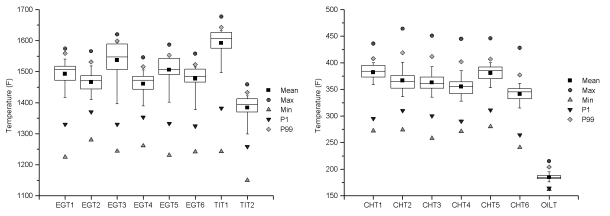


Figure 14. *EGT1-6*, *TIT1*, *TIT2*, *CHT1-6* and *OILT* for regime 4

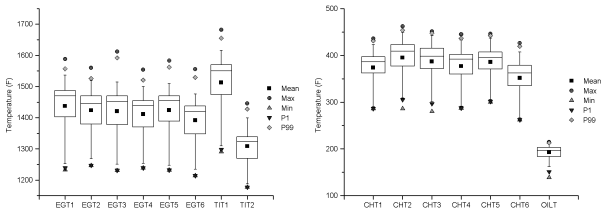


Figure 15. *EGT1-6*, *TIT1*, *TIT2*, *CHT1-6* and *OILT* for regime 5

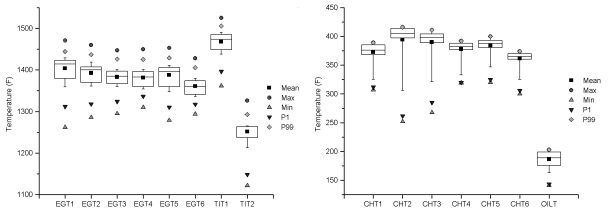


Figure 16. *EGT1-6*, *TIT1*, *TIT2*, *CHT1-6* and *OILT* for regime 6

VII. ENGINE FAULT PATTERNS

Engine monitor is capable of displaying EGT-CHT patterns suitable for fault detection. Patterns consist of bar graphs, darker bars represent EGT and lighter bars CHT values. Each pattern corresponds to one or more engine problems. Proposed pattern recognition technique employed in this method is rule based. This is due to rather precise fault pattern descriptions available in pilot's guide that comes with the engine monitor, [1]. There is also a lack of numerous real world patterns that would otherwise justify use of some statistical pattern recognition techniques. Rules are defined as descriptions in English language, usually one sentence and catalogued in manual. Following parameters (measured and calculated) are used in rules that describe conditions for fault pattern: EGT_{max}, EGT_{min},CHT_{max}, CHT_{min}, *DIFF* and *RPM*. Here is a list of catalogued patterns (six cylinder engine), but now with derived simple mathematical description (conditions) suitable for program implementation. Resulting conditions are described in relations (10) - (21).

TABLE VII 1% and 99% percentile values for various engine parameters and engine regimes (used limits are shaded)

Param. p_i	Regime r											
	1		2		3		4		5		6	
	L_L	L_H	L_L	L_H	L_L	L_H	L_L	L_H	L_L	L_H	L_L	L_H
EGT1	482	991	695	1365	1077	1475	1330	1559	1239	1557	1312	1445
EGT2	418	1081	742	1427	1195	1477	1370	1531	1247	1526	1318	1437
EGT3	488	1094	846	1380	1153	1519	1330	1599	1231	1592	1324	1426
EGT4	466	1147	866	1417	1163	1466	1353	1516	1239	1521	1336	1425
EGT5	476	1056	838	1372	1133	1487	1333	1553	1232	1563	1310	1429
EGT6	468	1141	896	1389	1167	1465	1325	1523	1214	1530	1317	1406
TIT1	518	1030	802	1408	1127	1568	1382	1643	1298	1655	1396	1506
TIT2	409	936	773	1285	1033	1385	1258	1433	1177	1427	1148	1293
CHT1	88	338	223	337	274	415	295	408	286	431	312	389
CHT2	79	312	173	354	277	433	310	419	305	454	262	416
CHT3	102	300	190	340	258	422	300	412	297	448	285	411
CHT4	75	339	216	341	271	417	290	402	287	437	320	392
CHT5	95	333	225	356	280	423	311	407	302	442	325	400
CHT6	83	327	213	323	243	400	264	377	262	420	306	374
OILT	69	188	97	188	136	197	164	204	151	213	143	203
FF	1,4	4,5	3,4	9,2	6,5	11,7	8,2	13,4	10,9	22,3	21,7	27,8

1. 75° to 100° EGT rise for one cylinder during flight

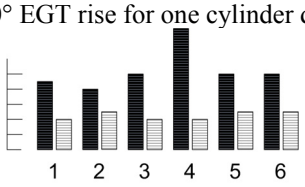


Figure 17. Pattern 1

CAUTION if abs(EGT_{max} - EGT_i) > 75 for any i=1,…,6 (10)

2. EGT increase or decrease after ignition system maintenance
   – not implemented (external information about maintenance is needed)
3. Loss of EGT for one cylinder

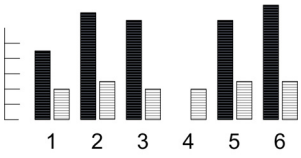


Figure 18. Pattern 3

CAUTION if EGT_{min} < 600 (11)

4. Loss of EGT for one cylinder; no digital EGT
   – not implemented, imprecise log entry definition
5. Decrease of EGT for one cylinder

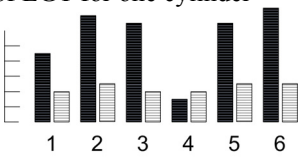


Figure 19. Pattern 5

CAUTION if 600 < EGT_{min} < 1000 (12)

6. Decrease of EGT for one cylinder at low RPM

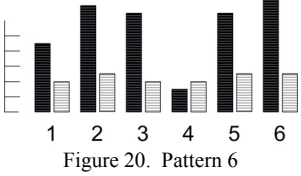


Figure 20. Pattern 6

CAUTION if *DIFF* > 500 AND *RPM* < 1500 (13)

7. EGT and CHT not uniform (injection engines only)

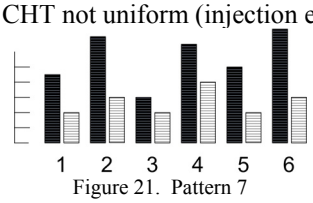


Figure 21. Pattern 7

WARNING if (EGT_i/CHT_i)<3 OR (EGT_i/CHT_i)>6 for any i=1,...,6 (14)

8. Decrease in EGT for all cylinders

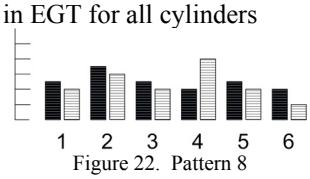


Figure 22. Pattern 8

WARNING if $EGT_i < 1200$ for all i=1,...,6 (15)

9. Slow rise in EGT, low CHT

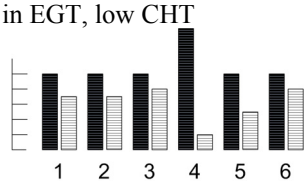


Figure 23. Pattern 9

WARNING if $EGT_i > 1600$ and $CHT_i < 300$ for any i=1,...,6 (16)

10. High CHT on cylinders on one side of engine

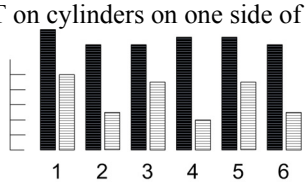


Figure 24. Pattern 10

WARNING if abs(CHT_{left}–CHT_{right}) > 100 *left*=2,4,6 *right* =1,3,5 (17)

11. Rapid rise in EGT/CHT of one cylinder

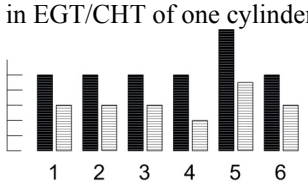


Figure 25. Pattern 11

WARNING if $EGT_i > 1650$ and $CHT_i > 400$ for any i=1,...,6 (18)

12. Sudden off scale rise for any or all cylinders

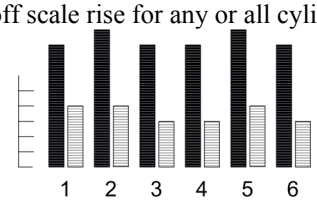


Figure 26. Pattern 12

WARNING if $EGT_{max} > 1650$ for any i=1,...,6 (19)

13. Loss of peak EGT - not implemented, leaning process
14. Decrease in peak or flat EGT response to leaning process - not implemented, leaning process (mixt. adj.)
15. Bellow 10,000 ft full throttle causes EGTs to rise – not implemented, 10,000 ft info needed
16. CHT more than 500º, EGT normal. Adjacent EGT may be low

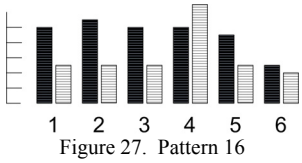


Figure 27. Pattern 16

WARNING if $CHT_{max} > 500$ AND $EGT_{max} < 1600$ (20)

17. Large DIFF at low RPM

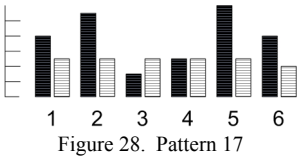


Figure 28. Pattern 17

WARNING if *DIFF* > 500 AND *RPM* < 1500 (21)

All records in the engine log could be checked for fault patterns using procedure and derived conditions:

```
for record=1 to last
{
 find EGT_max, EGT_min, CHT_max and CHT_min, get DIFF and RPM
 for i = 1 to N_patterns
   check conditions for pattern(i)
}
```

VIII. Conclusion

Proposed method combines default engine monitor limits, statistical analysis of engine parameters for different engine working regimes and rule based pattern recognition. Based on graphic presentation of available engine data percent of the maximal horse power (% *HP*) was chosen as a variable for regime selection. Engine parameter statistics were determined for all engine regimes and presented using exploratory data analysis Box-Wiskers plots. Method is primarily intended for parsing engine log after the flight. Build in default engine monitor alarm limits are preserved for the detection of severe engine problems and issuing warning alerts. Tighter limits are imposed on engine parameters for each particular engine regime and for detecting finer engine problems and issuing caution alerts. Upper limits are 99 percentile of engine parameters for a particular regime (for oil temperature and fuel flow additional 1 percentile is used for lower limits). Mathematical description for most fault patterns is produced from linguistic description and expert opinion. Such mathematical descriptions in terms of pattern conditions are suitable for rule-based pattern recognition. With larger scale statistics (including more flights) and determination of reliable thresholds (adding minimal parameter deviation periods) the method could be applied in real time with indications in a cockpit.

References

[1] *Pilot's Guide, Engine Data Management, EDM-730, EDM-830 and EDM-740 Experimental*, JPI, California, USA, 20010, Available: https://www.jpinstruments.com/wp-content/uploads/2012/06/PG-EDM-730_830-Rev-A-02-JULY-2009.pdf, [18 March 2012]

[2] *EzTrends* (Download and Plotting Software), JPI, California, USA, Available: http://www.jpitech.com/softwaredownload.php [13 March 2012]

[3] Variable pitch propeller brief, RAF Lyneham, Available: http://www.lynehamaviation.co.uk/documents/VPFirefly.ppt [13 March 2013]

[4] D. Miljković, "Fault Detection Methods: A Literature Survey", *Proc. MIPRO 2011, Vol. III, CTS & CIS*, pp. 110-115, Opatija 23-27 May 2011

[5] D. Miljković, "Engine Fault Detection in a Twin Piston Engine Aircraft", *Proc. MIPRO 2012, Vol. III, CTS & CIS*, pp. 1022-1027, Opatija 21-25 May 2012

[6] D. Miljković, "Review of Novelty Detection Methods", *Proc. MIPRO 2010 - Vol. III: CTS & CIS*, pp. 27-32, MIPRO, Opatija 24-28 May 2010

Active Noise Reduction for General Aviation Aircraft Using Conventional Headsets

Dubravko Miljković
Hrvatska elektroprivreda, Zagreb, Croatia
dubravko.miljkovic@hep.hr

Abstract - Active Noise Reduction (ANR) headsets are quite efficient at reducing noise in a cabin of general aviation aircraft, however they are generally bulky, uncomfortable and expensive. If compromise is accepted in terms of achieved lower noise attenuation for passenger, acceptable noise reduction could be achieved using conventional headsets that are much more comfortable, better suitable for listening to the music and could be wireless. In this paper solutions are presented that use synchronized waveform generator where noise canceling waveform is dependent on engine rotational speed and seat position within a cabin of a single engine general aviation aircraft.

I. INTRODUCTION

It is very difficult to achieve global noise reduction within aircraft cabin. This would require large number of secondary sources, reference and error microphones together with DSP hardware. Active Noise Reduction (ANR) headphones become very popular solution instead. ANR headphones use one or more microphones placed near the ear, and electronic circuitry which uses the microphone signal to generate an "antinoise" signal. Such headphones are convenient for pilots because of the high importance of securing communication, however they are quite bulky, heavy and its frequency range is often limited to a range necessary for the communication with the Air Traffic Control (ATC). Difference in frequency and weight range for some typical ANR and conventional headsets is shown in Table I. For greater passenger comfort weight and frequency range 20-20,000 Hz may be of interest. This goal may be accomplished with conventional headsets by accepting lower noise reduction than achieved by professional ANR headsets. Such headsets can be supplied by predetermined noise canceling signal specific to seat position. Passenger head movement within a cabin of the general aviation aircraft is generally quite small so the predetermined signal may fulfill canceling requirement quite well, particularly when considering that noise frequencies needed to be actively suppressed are low and appropriate wavelengths quite sizable in comparison to the head movements.

TABLE I Comparison of some ANR and conventional headsets

Headphones	Frequency range	Weight
David Clark H10-13X (ANR)	100-5,500 Hz	510g
Telex Stratus 30 (ANR)	50-10,000 Hz	403g
Shure SRH241A	20-20,000 Hz	238g
Pioneer SE-MJ751	8-28,000 Hz	200g

II. PRINCIPLE OF ACTIVE NOISE REDUCTION

Active noise reduction (ANR) is a reduction of the unwanted noise by generating an acoustic signal that interferes with the noise. It is based on principles of superposition and destructive interference, [1, 2]. Noise reduction is achieved by generating a signal identical in sound pressure level but exactly reverse in phase to the offending noise, Fig. 1. As a result these two sounds cancel each other. Consider the case when noise source on is sinusoidal:

$$p_1 = A\sin\omega t \tag{1}$$

The sound pressure from secondary source is:

$$p_2 = -(A+a)\sin(\omega t+\theta) \tag{2}$$

where a and θ represent the amplitude and the phase error between two waves.

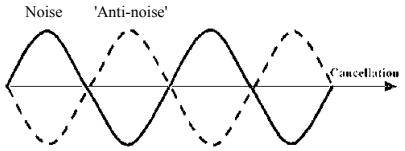


Figure 1. Active vibration control is based on destructive interference between vibrations

Resultant noise is given by relation:

$$p_1 + p_2 = p = A\sqrt{4\left(1+\frac{a}{A}\right)\sin^2\frac{\theta}{2}+\left(\frac{\alpha}{A}\right)^2} \tag{3}$$

Level of the resultant depends on amplitude and phase errors. Great attenuation can be achieved if amplitudes and phases are closely matched as shown in (4) and Table II, [1]. There is a great need for precision when generating “antinoise”.

$$\Delta = 10\log\left[4\left(1+\frac{a}{A}\right)\sin^2\frac{\phi}{2}+\left(\frac{a}{A}\right)^2\right] \tag{4}$$

TABLE II Attenuation (in dB) as a function of amplitude and phase errors between original noise and canceling signal („antinoise“)

a/A	ϕ				
	0	1	3	10	20
0		35,0	25,5	15,2	9,2
0,005	46	33,7	24,4	15,1	9,2
0,05	26	25,7	22,6	14,2	9,0
0,10	20	19,8	18,6	13,6	8,6
0,20	14	14,0	13,6	11,0	7,6

In real world noise source is of more complex waveform that can be represented as a sum of harmonically related components, each with its frequency and phase angle. Efficient ANR system should produce the control signal that will address all this components.

III. General Aircraft Noise

Interior noise levels depend a lot on the aircraft, but on average the values are 80 dB and above, up to 110 dB in case of some piston (e.g. Cessna 210) and turboprop aircraft during some phases of flight (takeoff) and some general aviation aircraft do not provide adequate acoustic protection. Measured noise levels for two stationary Cessna 172 airplanes, together with corresponding articulation indexes (measure of the speech intelligibility in a noise environment) are shown in Table III, [3].

TABLE III Interior noise for aircrafts with two and three blade propeller and different rotational speed (RPM)

RPM	Cessna 172N Two blade propeller	
	Noise Level L_{Aeq} dB(A)	Articulation Index (AI)
900	69,5	0,45
1500	82,3	0,28
2400 TO/GA	93,3	< 0.20 (0.15)*
RPM	Cessna FR172F Three blade propeller	
	Noise Level L_{Aeq} dB(A)	Articulation Index (AI)
900	69,0	0,45
1500	83,0	0,28
2400 TO/GA*	91,0	< 0.20 (0.17)*

*TO/GA - Take Off/Go Around power

A. Noise Contributing Sources

Aircraft noise contains the following main components: engine noise, propeller noise, airframe noise and structure borne noise (as a particular kind of airframe noise), [4]. Aircraft interior noise is combination of all mentioned components that, with various degrees, penetrate into the aircraft cabin. In case of multiengine aircraft when propellers are not synchronized noticeable beats interference may exist.

B. Engine Noise

Piston engine noise is the result of pressure pulses on intake and exhaust (engine exhaust noise) during engine four cycles, as well as the noise coming into the cabin through the firewall. The engine exhaust noise originates at the exhaust tailpipe openings and is transmitted through the cabin walls, firewall, and nose gear bay into the cockpit. In the case of piston engine, noise spectrum is dependent on rotational speed (*RPM*) and number of cylinders.

Cylinder firing rate (*CFR*) is dependent on the rotational speed:

$$CFR = \frac{RPM}{60} \tag{5}$$

Engine firing rate (*EFR*) is dependent on *CFR* and the number of cylinders (*N*):

$$EFR = N \cdot CFR \tag{6}$$

For all cylinders to fire two engine revolutions are needed. This may translate in two full periods before noise waveform repeats itself.

A. Propeller Noise

In a single engine aircraft propeller noise enters the cabin through front window in the form of pressure pulses. Propeller noise is composed of tonal and broadband components. The tonal component contains basic frequency and its harmonics. The basic frequency f_1 or *BPF* (blade pass frequency) is the product of the propeller rotation speed and the number of propeller blades:

$$BFR = \frac{N_R N_B}{60} \tag{7}$$

where is:

BPF basic frequency of tonal propeller component
N_R propeller rotation speed (rotations per minute)
N_B number of propeller blades

Blade pass frequency for an aircraft with two and three blade propeller at three rotational speeds is shown in Table IV.

TABLE IV Blade Pass Frequency (BPF) in Hz for various rotation speeds (RPM) and number of blades

RPM (N_R)	2 blades	3 blades
900 – Idle	30	45
1500	50	75
2400 – TO/GA*	80	120

*TO/GA - Take Off/Go Around power

Beside the base frequency also there are also harmonic components:

$$f_N = f_1 N \tag{8}$$

where is

f_n frequency of *n-th* harmonic
f_1 basic tonal frequency (*BPF*)
N number of particular harmonic

B. Airframe Noise

Airframe noise is the result of air flow (wind around airframe). It is of the broadband flow mixing type except where a resonant cavity is formed (e.g. at control surface gaps). Its main characteristic is a great dependence on aircraft speed. Noise intensity is related to aircraft speed with the following equation:

$$I = kv^n \tag{9}$$

where *v* is the speed of an aircraft and the exponent *n* varies between 5 and 6 and is dependent on the shape of fuselage.

C. Structure Borne Noise

Engine and propeller vibrations excite fuselage that resonate at its vibratory modes that excite acoustic modes and subsequently radiate noise. Reduction of the structure borne noise could be achieved using vibration dampers.

At low RPM the noise is dominated by exhaust noise from the reciprocating engine. At higher RPM influence of propeller noise become considerable. In flight, the noise generated aerodynamically gradually increases with the aircraft speed.

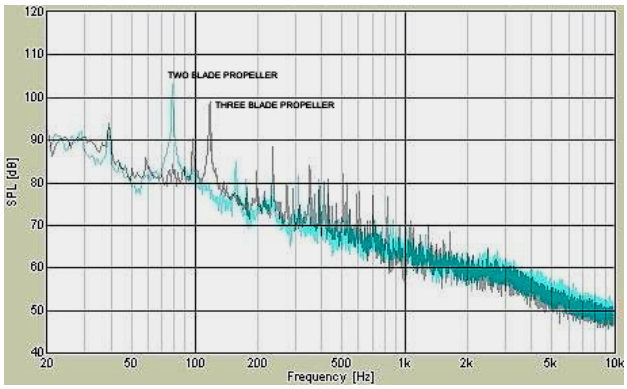


Figure 2. Spectral representation of interior noise for aircraft with two and three blade propeller– Takeoff/go around power, static

Tonal components are present in all flight phases. At higher speeds influence of aerodynamic noise become greater with a corresponding increase of relative importance of broadband noise in noise spectrum. An example of noise spectra for general aviation aircrafts with two and three blade propeller extracted from static noise recordings performed in a cabin of Cessna 172N (two blade propeller) and Cessna FR172F Reims Rocket (three blade propeller) is shown in Fig. 2. Noise waveforms are to an extent dependent on the engine RPM. Examples of noise waveforms for four different RPM are shown in Fig. 3-6. The waveform shape and amplitude fluctuates sometimes quite significantly. In overall noise there are deterministic and stochastic components. Engine and propeller produce mostly repetitive and deterministic sound. Airframe noise is by its nature stochastic sound. Structure borne sound is a combination of deterministic and stochastic sound.

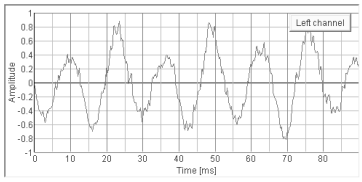


Figure 3. Signal c172_rpm1 - 1130 RPM

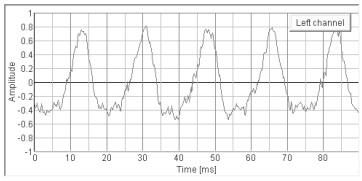


Figure 4. Signal c172_rpm2 – 1705 RPM

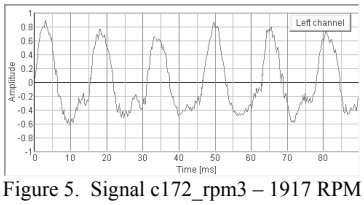


Figure 5. Signal c172_rpm3 – 1917 RPM

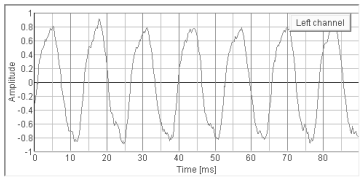


Figure 6. Signal c172_rpm4 – 2299 RPM

IV. Positional Difference

The cabin arrangement of a typical general aviation aircraft (Cessna 172) is shown in Fig. 7. Noise measurements for one particular GA aircraft (older, slightly bigger and more noisy Cessna 182E) is shown in Table V, [5]. Note slight noise differences across seat positions.

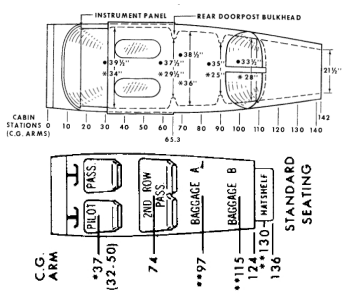


Figure 7. Cessna 172 cabin with seating

TABLE V Different noise levels are measured at various seat positions (NASA) and various flight phases

Measurement position	Noise level dBA				
	NCL[*5]	BRC[*6]	C65[*7]	C75[*8]	DES[*9]
M1[*1]	94.7	94.8	92.1	92.9	91.9
M2[*2]	93.5	93.4	91.3	93.2	90.9
M3[*3]	92.4	91.6	91.1	93.2	90.0
M4[*4]	90.8	89.9	89.7	91.0	90.1

[*1]Pilot seat, [*2]Copilot seat, [*3]Passenger - rear left, [*4]Passenger - rear right, [*5]NCL – Normal Climb, [*6]BRC – Best Rate Climb, [*7]Cruise at 65% power, [*8]Cruise at 75% power, [*9]Descent

The impulse response measurement using TSP (Time Stretched Pulse) method and decay curve for a cabin is shown in Fig. 8, (Cessna 172S). Some reflections exist within a cabin, acoustic modes are present but heavily attenuated due to high absorption and low reflection of quality interior furnishing materials and seats.

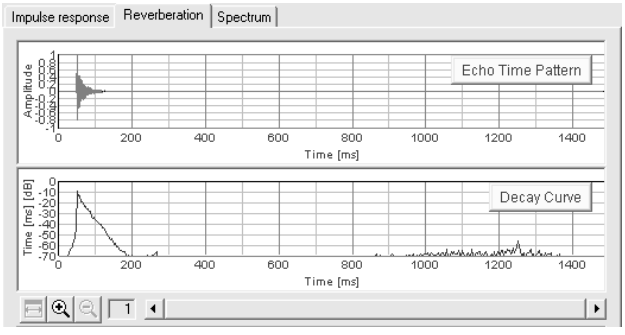


Figure 8. Impulse response measurement and decay curve for the cabin

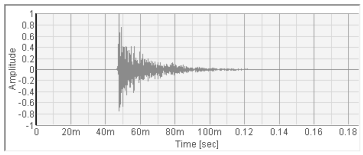


Figure 9. Impulse response measured at front seats

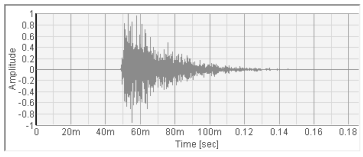


Figure 10. Impulse response measured at rear seats

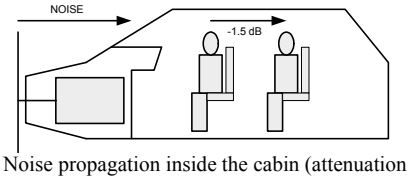


Figure 11. Noise propagation inside the cabin (attenuation and delay)

Distance *s* between the front and rear head positions is for Cessna 172 around 90 cm, and with the speed of sound *v* of 330 ms^{-1} adds calculated 27 ms sound delay Δt:

$$\Delta t = \frac{s}{v} \tag{10}$$

Due to the reflections of the noise within a cabin, as illustrated in Fig. 9 and Fig. 10 slightly different noise waveform is present at the front and rear seat positions. Attenuation of around 1.5 dB and small distortion of waveform due to the increased reflections in rear part of a cabin is present, Fig 11.

V. Analog ANR Headphones

ANR headphones employ local active noise control within small space of the earcup. Local ANR is very effective within local area of the 1/12 wavelength around destructive interference point of undesired noise source. ANR headsets combine passive and active attenuation. Passive attenuation is suitable for higher frequencies whilst active for lower frequencies. Core parts of ANR headphones are a noise-sensing microphone, some processing electronics and an anti-noise speaker. Analog ANR headphones exist as open or closed-loop system, [6].

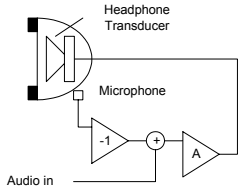


Figure 12. Open-loop system ANR headphones

The open-loop system has an external microphone and the advantage of simplicity, Fig. 12. The closed-loop system is more accurate as microphone is placed inside the ear cup and is physically close to the place where actual noise cancellation has to take place, Fig. 13. ANR in headsets is always combined with the passive reduction

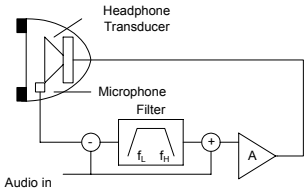


Figure 13. Closed-loop system ANR headphones

that becomes more effective at higher frequencies. Noise reduction of one ANR headset used in Cessna 172 aircraft is shown in Fig. 14. Three curves depict noise spectrum before cancellation, after passive cancellation only and remaining curve after the addition of analog ANR. Most headphones that are currently on the market use analog electronics for feedback, although digital models exist too.

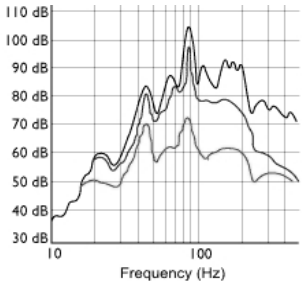


Figure 14. Passive, active and combined noise reduction

VI. ANR Using Conventional Headsets

Conventional headsets provide some degree of passive attenuation. Headset's earcups are able to block out some high frequency noise. Closed-back (or sealed) styles have the back of the earcups closed. Such closed and circumaural design is preferred choice for passive attenuation of ambient noise. Typically conventional closed-back headphones (not intended for noise reduction) block 8 to 12 dB. Depending on the model some may block even much more of ambient noise (32 dB for Sennheiser HD 380 Pro). However, the passive noise attenuation is not efficient at lower frequencies. Active attenuation is necessary by supplying signal to the headset with assumed "antinoise" waveform. Naturally this doesn't provide as efficient attenuation as with ANR headsets designed and optimized specifically for the task, but can still yield acceptable results, whilst retaining comfort and frequency range of conventional headsets. Also, wireless headset models can be used.

A. Description of the Method

Method uses two waveform generators, synchronized with tacho signal related to the engine rotational speed, RPM. Stored waveforms are position and RPM dependent. Front and rear seats have stored different waveforms that are appropriate for that seat position. Shape of noise waveform is dependent on RPM and it is not convenient to use just one waveform across whole RPM range. It considers time delay and noise attenuation between front and rear seats and stores two waveforms, one for front seats, other for rear seats. Because it is almost impossible to very accurately guess volume needed for "antinoise" signal in headphones, trimmers for the individual volume setting are provided (for initial and rare subsequent adjustments) that work together with cabin (general) ambient noise level sensing. Block diagram of the method is shown in Fig. 15. As the method is suitable for both software and hardware implementation, block diagram is of general description. Key elements are tacho generator, clock, control unit, waveform memory, D/A converters, low pass filters and output amplifier. Tacho generator detects engine rotations (e.g. inductive combined with acoustic) and resets address counter pointing to waveform samples. Clock signal (with sample rate f_s that equals the recording rate) pushes waveform samples to D/A converters (8-bit suffice due to normalized waveforms). Control unit generates appropriate addresses for the waveform memory. Waveform periods based on RPM are calculated from the tacho period *T*:

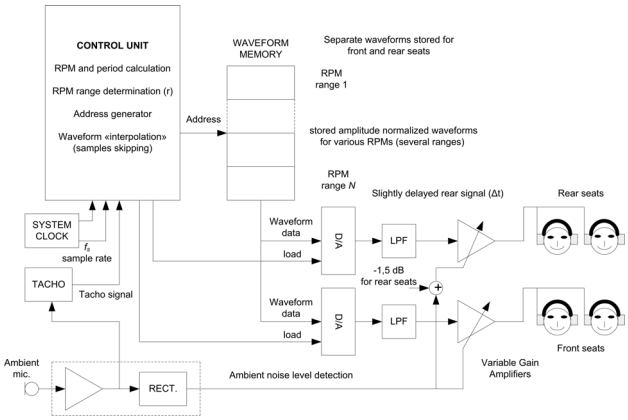


Figure 15. Block diagram of the method

$$RPM = \frac{60}{T} \tag{11}$$

Depending on the calculated RPM, the appropriate RPM range is determined and representative waveform is selected from the memory. Because calculated RPM is very rare at exact RPM range boundary, stored waveform is modified by skipping some of stored waveform samples. Frequency and phase response of typical conventional headphones is flat, Fig, 16. Phase response at lower frequencies (important for ANR) doesn't introduce significant phase error in canceling signal, doesn't interfere with noise cancellation and no additional phase correction is needed.

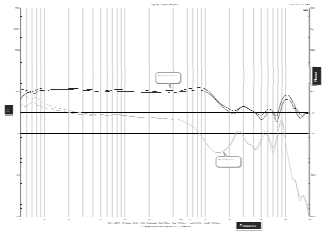


Figure 16. Frequency (upper curve) and phase response (lower curve) of typical conventional headphones

B. Preparation of Wavefoms

Waveforms are recorded once for different RPMs. They are somewhat dependent on the aircraft type. All major flight phases should be included (idle, take off, climb, cruise and descent). Current RPM range *r* is determined from:

$$r = \frac{RPM - RPM_{idle}}{W} \tag{12}$$

where RPM_{idle} is lowest engine speed and *W* is the range width. Setting RPM_{idle} to 600 and *W* set to 400 leads to following five RPM ranges: 600-1000, 1000-1400, 1400-1800, 1800-2200 and 2200-2600 RPMs. Slightly different and delayed waveform is supplied to headsets at front and rear seat positions. For that purpose noise waveform recorded closely behind the windshield is passed using auralization process involving measured two impulse responses (with measurement microphone placed at head level at front and rear seats). Alternative and better solution would be precise noise measurements at front and rear (or even all) seat positions. Recorded waveforms fluctuate slightly between successive periods. To remove these fluctuations, synchronous averaging is

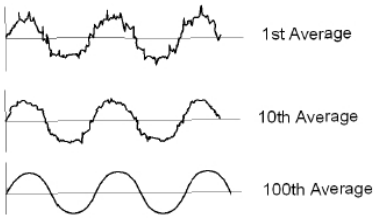


Figure 17. Synchronous averaging

performed on a recorded signal and averaged waveform is stored in waveform generator.

Input noise signal captured by the recording microphone, *f(t)* has a deterministic signal component *S(t)* and a noise signal component (in this case fluctuating waveform component) *N(t)*:

$$f(t) = S(t) + N(t) \tag{13}$$

$$f(iT) = S(iT) + N(iT) \tag{14}$$

Arithmetic averaging of periodic signal after averaging of *m* (good value for *m* is 50) periods is described by:

$$\frac{1}{m}\sum_{k=1}^{m} f(iT) = \frac{1}{m}\sum_{k=1}^{m} S(iT) + \frac{1}{m}\sum_{k=1}^{m} N(iT) \tag{15}$$

After numerous iterations noise (random) components of sum cancel each other and remains only averaged waveform with periodic components as shown in Fig. 18.

C. Fine Time Adjustments

Waveforms are stored for different RPMs. To cope with RPM changes between RPM limits of the particular RPM range (that is represented with the single waveform recorded at the lower limit of the RPM range), fine time adjustment is achieved by skipping one waveform sample after putting out *k* samples when producing output waveform for the particular RPM. For example if stored wavefrom for the particular RPM range consists of 250 samples and current RPM is 2% higher than lower value of appropriate RPM range, than it will skip 5 samples or one in 50 samples (50=250/5). This way system gets time compressed waveform of appropriate length regardless of the RPM value within one particular RPM range, Fig. 18.

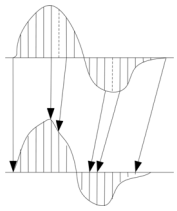


Figure 18. Producing new waveform by skiping every *k* samples

$$p = \frac{RPM}{RPM_{i,L}} - 1 \tag{16}$$

where *p* is ratio of samples that should be skipped and

$$k = \frac{1}{p} \tag{17}$$

where every *k*-th sample is skipped.

Low pass filtering is applied to analog signal after the D/A converters as we are dealing with first three or four harmonic frequencies with significant acoustic levels.

D. Ambient Level Sensing

Estimate of the current noise level in a cabin is sensed by one microphone placed forward in a cabin. It is accomplished by rectifying amplified signal form a sensing microphone (kind of a volume meter). This noise level modifies amplitude of stored amplitude normalized waveforms using variable gain amplifiers.

VII. RESULTS

Passive attenuation is dependent on a particular headset. Simulation was performed on available recorded noise signals. Achieved overall active noise reductions are shown in Table VI. It is suitable for low frequencies that include first few engine-propeller harmonics with highest noise energy content. Impulse response was included in simulation to distort the noise waveform due to cabin reflections. Spectral difference between original noise and active attenuation is show in Fig. 19. Great attenuation can be noted at discrete frequencies corresponding to blade pass frequency and its harmonics (>25 dB at some BPFs).

TABLE VI Achieved overall active noise reduction in 0-10 kHz range* (ideal case - no positional error and no A or C frequency pondering)

Signal	RPM	BPF (Hz)	Attenuation (dB)
c172_rpm1	1130	37,6	9,9
c172_rpm2	1705	56,8	13,3
c172_rpm3	1917	63,9	13,1
c172_rpm3	2299	76,7	21,4

*across the whole range, not at particular discrete components like BPF

Despite moderate overall active attenuations, they are meaningful addition to passive attenuation. By lowering the noise floor, the listening dynamic range of the headphones is increased by an additional 20 dB. Lowest attenuation was achieved for test signal c172_rpm1. It could be noted from Fig. 3 that waveform is repeating with period of two engine rotations vs. one rotation for the rest of test waveforms. This is due to dominating engine noise at that low RPM. Engine noise has a period of two rotations and stored waveform containing just one rotation is not suitable for this low RPM values (fortunately with low noise levels too). At higher RPMs with increasing contribution of propeller noise better active reductions are achieved. Zone of silence according to λ/12 criterion (30º phase error between noise and "antinoise") is shown in Table VII. Even with three blade propeller there is still enough room for occasional small head movements.

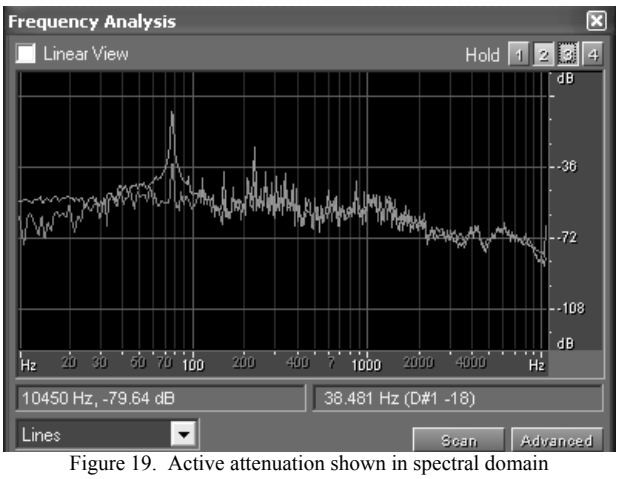


Figure 19. Active attenuation shown in spectral domain

TABLE VII λ/12 at cruise setting (2200 RPM)

Frequency	λ/12 (m)	
	two blade propeller	three blade propeller
BPF	0,375	0,333
2BFP	0,188	0,167
3BPF	0,125	0,111

VIII. CONCLUSION

Achieved active attenuations provide useful reduction in subjective loudness (one half for each 10 dB). Active attenuation is used for lower frequencies while attenuation of higher frequencies depends on noise attenuation properties of conventional headphones. In combination with active attenuation and some cheap microcontroller or ASIC hardware that produce "antinoise" signal for all four or six seats (suitable for intercom) this paper proposed and analyzed one approach. Advantage of the method is its suitability for periodic noise and simultaneous generation of canceling signal for all seat positions. Disadvantage lies in difficulties of accurately capturing all influential factors (ambient noise level, tacho signal, waveform shape for particular RPM range, attenuation and waveform change between front and rear seats) in simple and effective way necessary to achieve high level of active noise reduction. Assumptions about single engine propeller general aircraft noise (repetitive periodic noise) are built into the method. Method is not suitable for twin engine aircraft due to more complex cabin noise coming from two engine/propeller combinations, particularly when engine synchronizer and synchrophaser are switched off (if present on aircraft) during the take off and landing. Improvements are possible when using conventional headphones with integrated microphone (for intercom). This microphone can be used to detect ambient level more precisely than one used for global cabin ambient level. Also, necessary time delay for waveform could be more precisely determined during long periods when accompanying microphone is not used for talking. Noise waveform shape and amplitude has variations even at constant RPM Synchronous waveform generator with stored averaged waveform samples can't follow these fine fluctuations, hence attaining somewhat lower active attenuation results. Alternative could be the use of fixed (predetermined) digital filters for modifying picked ambient noise and introducing necessary time delay and amplitude change.

REFERENCES

[1] H. G. Leventhall and L. Wong, A Review of Active Attenuation and Development of an Active Attenuator 'Open Refuge', HSE Contract Research Report No. 4/1988

[2] D. Miljković, M. Maletić and B. Somek, "Active Noise Control", *Proceedings MIPRO 2007*, Vol. III, CTS & CIS, 21-25 May 2007, Opatija, Croatia

[3] D. Miljković, J. Ivošević and T. Bucak, "Two vs. Three Blade Propeller - Cockpit Noise Comparison", *Proceedings 5th Congress of Alps-Adria Acoustics Association*, 12-14 Sept 2012, Petrčane, Croatia

[4] D. Miljkovic, M. Maletić and M. Obad, "Comparative Investigation of Aircraft Interior Noise Properties", *Proceedings of the 3rd AAAA*, 27-28 September 2007, Graz, Austria

[5] J. F. Unruh and P. D. Till, General Aviation Interior Noise: Part I, Source/Path Identification Technology, MASA/CR-2002-211665, May 2002

[6] Chu Moy, Active Noise Reduction Headphone Systems, 2001, Available: http://headwize.com/?page_id=601 [17 March 2013]

Reducing Off-road Vehicle Seat Vibrations Using Pneumatic Active Suspension and Fuzzy Logic Based Controller

Aida Salihbegovic, Almir Salihbegovic, Maja Hadziselimovic and Zikrija Avdagic
Department for Automatic Control and Electronics
Faculty of Electrical Engineering, University of Sarajevo
Sarajevo, Bosnia and Herzegovina
aida.salihbegovic@gmail.com

***Abstract*—This paper describes the capabilities of fuzzy logic based controllers in the process of pneumatic active suspension of a heavy vehicle seat vibrations. The pneumatic active suspension system is introduced to solve conflicting requirements of comfort and handling. The air spring is used as an active element and the damper is used as passive element for reducing vibrations. The usage of the active air spring enhances passenger comfort by comparison with passive, semi-active and active hydraulic suspension, during long time drives at a very bumpy road. Description of the seat, including its mechanical characteristics, is given in the paper. The mathematical model was created according to the physical setup of the vehicle seat at the testing laboratory. MATLAB and Simulink are used as tools for developing the simulation model of the driver seat. Control system description and implementation at the experimental setup using dSPACE module, are also explained. The SEAT value is used for the validation of control quality. The obtained simulations show that the developed road adaptive suspension controllers provide superior driver's comfort for most difficult types of road.**

***Index Terms*—active suspension, reducing vibrations, air spring, hydraulic cylinder, damper, fuzzy logic, fuzzy controller, nonlinear systems, SEAT value**

I. Introduction

Driving an off-road vehicle for a long time period can cause health disorders on a driver's body [1], [2]. Low frequencies of the heavy vehicle vibrations are responsible for health disorders, such as low back pain, fatigue and stress. The human body is most sensitive to vibrations at frequencies from 1 Hz to 80 Hz, according to the ISO 2631 standard [1].

Classical control theory, based on an explicit mathematical model description, has been previously used for reducing vibrations of the heavy vehicle seat [3]–[5]. The main limitation in the traditional control methods are the difficulties to provide exact mathematical description of some physical phenomenons. Furthermore, the mathematical equations may be too complex, or the user is unable to collect all the needed information about the system. The development of soft computing techniques has made a great progress in solving mentioned problems.

By proposing a new nonlinear air-spring model, improvement of passengers ride comfort in rail vehicles is described in [6], [7]. The results showed that passengers comfort is improved and ride comfort index is reduced by modifying the air suspension parameters. In [8] is disscused the automotive suspension system capable of independent control of stiffness, damping and ride-height. The application of this suspension for improvement of vehicle dynamics is addressed in [9]. A novel design of a pneumatic automotive suspension system capable of instantaneous stiffness change with no power input, and no ride-height change due to stiffness change is proposed in [10]. In [11] it is shown that the H_∞ controller provides good trade-off between ride quality, suspension packaging and road holding constraints. Experiments are done only with a front wheel suspension system. Paper [12] explores the state of the art in fuzzy inference systems, neural networks, genetic algorithms and their combination for suspension control issues. Also, the paper focuses the practical implementation problems raised by nonlinear and uncertain control properties. In order to overcome uncertain conditions caused by rules or membership functions in fuzzy controllers, type-2 fuzzy sets are described in [13]. The proposed approach integrates the IT2 membership functions, T-S fuzzy model, Wu–Mendel uncertain bounds, and further-optimization algorithms into one control framework.

Off-road vehicle seat vibrations were reduced using neural networks [14]–[16], fuzzy logic [17] and hybrid modes [18] to control passive and semi-active elements in the under-seat construction. The properties of the passive suspension components can not be changed while driving heavy vehicle. In semi-active suspension systems, the hardness of the damper and/or the stiffness of the spring can be adjusted, depending on the specific situation. Active suspension represents the most advanced technique for vibration rejection, which implies energy consumption by the suspension components.

This paper introduces the air-spring as an active element, that can achieve higher level of suspension performances in comparison with passive and semi-active suspension elements in the under-seat construction [14]–[18]. Also, it will be shown that the active air spring enhances passenger comfort comparing it to the active hydraulic piston [19], during long time drives at a very bumpy road.

The acceleration of the seat and the deflection have been

used for validation of the control system performances. In this paper, pneumatic active suspension will be used to solve conflicting requirements of comfort and handling. The active suspension elements allow design of the controllers that focus on minimizing the acceleration of the seat and, consequently, the force on the driver, in a broader range of the deflection. The physical setup of the vehicle seat is assembled at the testing laboratory in order to develop a control system for vibrations reduction. Also, a mathematical model has been created for simulation of the system behaviour. Then, a control concept for the suspension elements in the under-seat construction, which minimizes the impact of the vibrations on a driver's body, is obtained. The proposed control provides robustness, low-cost and light-weight implementation to the real system. The experiments are conducted by acting on a vehicle seat, both simulated and real model, with the test signal that represents the toughest conditions of the bumpy road.

The results of vibrations reduction, that can be achieved for very difficult types of road, using a fuzzy logic based controller and the combination of active and passive control elements, are also discussed. There are several advantages of the used approach. The exact dynamics of the whole system does not need to be known in order to design the fuzzy logic based controller. It is possible to generate forces in both directions with the usage of the air-spring as an active element. This is a great progress comparing to the passive and semi-active suspension elements previously used [14]–[18]. It will be shown that the designed controller gives superior results, even for the most difficult types of road.

The paper is organized as follows. The model of the vehicle seat is described in Section II. The limitation for the control concept and the design of road adaptive suspension controllers are subject of Section III. Implementation of the control algorithm on the experimental setup is briefly discussed in Section IV. The simulations and experimental results are presented in Section V. Concluding remarks are given in Section VI.

II. Model description

A physical model of the heavy vehicle seat is given in Fig. 1. It consists of the seat, its suspension system and 75 kg dummy, which represents the driver's body. The suspension system contains the active air spring and the passive damper and, that are fixed to the vehicle's cabin floor. The mentioned elements are used for vibrations reduction and achived performances will be compared with control concept described in [19].

A simulation of the cabin floor movements is done using a hydraulic cylinder, which generates the vibrations. The vibrations need to be isolated in the under-seat area, so they should not be transferred to the driver. It can be seen in Fig. 1 that the seat is equipped with 5 sensors, which enable collecting all the needed data.

The mathematical model of the vehicle seat can be described by a second order differential equation:

$$m_S \ddot{z}_S = F_{LF} + F_D - F_R - m_S g, \tag{1}$$

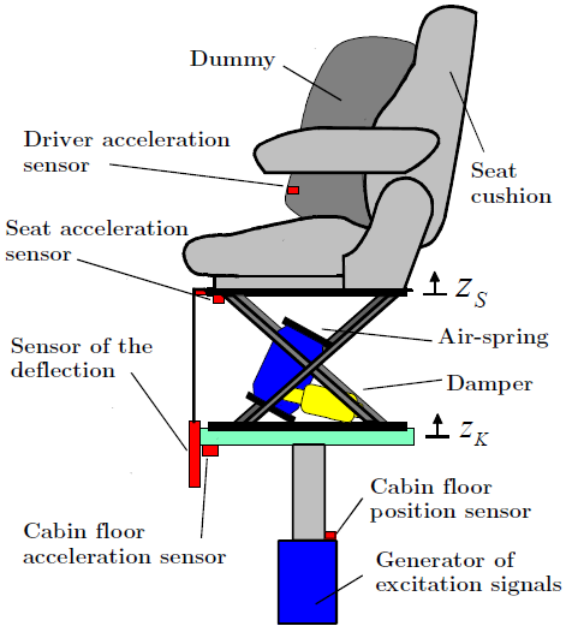


Fig. 1. Model of the off-road vehicle seat

$$F_{LF} = K\left(z_K - z_S\right), \tag{2}$$

$$F_D = C\left(\dot{z}_K - \dot{z}_S\right), \tag{3}$$

where:

z_K – the road disturbance,
z_S – the seat position,
m_S – the total mass of the seat and the dummy,
C – the damping coefficient,
K – the air-spring stiffness,
F_{LF} – the air-spring force,
F_D – the oil damper force, and
F_R – friction force in the under-seat construction.

The total mass of the seat m_S includes the mass of the seating area, the cushion mass and the mass of the dummy.

Body of the air-spring is made of the leather body with variable volume. There are two ON-OFF valves mounted to the air-spring. The first one connects the air spring to the compressor with constant pressure of 9 bars. When this valve is opened, compressed air enters the spring and the pressure increases. The second valve enables decreasing the air-spring pressure to the atmospheric pressure. The force generated by the air-spring is directly proportional to the spring stiffness. The nominal stiffness of the air-spring can be described with:

$$K = \gamma A^2 \frac{P}{V}, \tag{4}$$

where A, P and V are values of the cross-cut surface, pressure and volume of the air-spring, respectively, and γ

is heat capacity ratio of the air. Adiabatic compression and expansion of the air are assumed. It can be noticed in (4) that stiffness of an active air-spring depends on the air volume in the spring. These changes need to be instantaneous and height of the air-spring needs to be constant during this operation. This is partialy achieved with the passive damper, which is used as auxiliary suspension element. As a result the active air-spring has high stiffness at high vibration's frequencies and low stiffness at low frequencies. Damping control is achived with the damper, which is passive element and in parallel with the active air-spring arrangement.

It should be noticed that the complete and detailed mathematical model of the seat (including the model of the passive damper, the active air-spring and other physical processes) is far more complex [20]–[22] then the presented one, and it is not included in this paper for the sake of simplicity.

III. Control System Description

A. A limitation and validation of the control concept

There are two important measures of quality that are used for the control system development: the deflection and the SEAT value [20]–[22]. Deflection is defined as the relative movement of the seat to the cabin floor, and evaluated as a difference between the movement of the seat z_S, and the movement of the cabin floor z_K:

$$z_A = z_S - z_K. \tag{5}$$

The deflection of the seat needs to be less than 5 cm: $|z_A| < 5$ cm. Excessing this value would result in damage of the under-seat construction. The second criterion that is used to measure the control quality is the SEAT value. Smaller SEAT value means that unwanted vibrations are better reduced. The SEAT value is calculated as the ratio between the values of the effective seat acceleration, and effective cabin floor acceleration:

$$SEAT = \frac{\ddot{z}_{S_{eff}}}{\ddot{z}_{K_{eff}}}. \tag{6}$$

Fig. 2 presents a scheme that is used to calculate the SEAT value. Before the SEAT value evaluation, the measured values of the accelerations (seat and cabin floor) are filtered. The filtering is done using the VDI 2057 German engineers' recommendations and ISO standard [1]. The filter is designed with sensitiveness of a human body to specific frequency of the vibrations in mind.

Driving heavy vehicles is usually done at a very bumpy roads. As a result, high vertical vibrations are carried to a driver through the seat construction. The excitation signals are the representation of a bumpy road. All the experiments were done using ISO em3 signal, because it is the one that simulates the toughest conditions of the road.

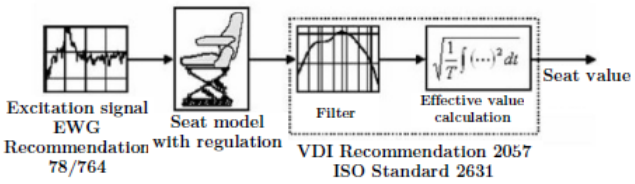


Fig. 2. SEAT value evaluation scheme

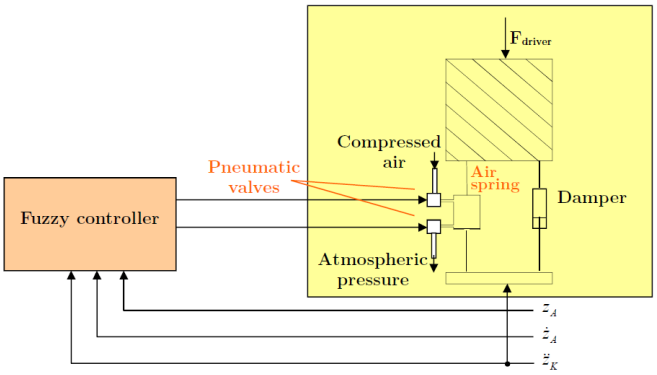


Fig. 3. Fuzzy controller design

B. Fuzzy controller design

In this paper, a Sugeno fuzzy controller was used. It uses less processing time, and shows better performance when used in a real-time applications. Designed fuzzy controller uses three input variables: the deflection of the seat z_A, the velocity of the deflection $\dot{z}_A$ and the acceleration of the cabin floor $\ddot{z}_K$ (see Fig. 3). The valves position is used as an output variable of the fuzzy controller. The force value generated by the air-spring depends on the valve position. In the process of generating the fuzzy controller rules, it was also necessary to compensate the influence of the passive damper on the system. Triangular and trapezoidal membership functions are used to describe input variables of the controller. The best system performances were achieved using these two types of membership functions.

While designing the rules of the fuzzy controller, it was noticed that the rules which had good control performance concerning the SEAT value, had a negative impact on the values of the seat deflection. Thus, writing the appropriate set of rules was actually equal to finding the balance between decreasing the SEAT value and assuring that the deflection stays in the allowed range of ± 5 cm. This is one of the contributions of this paper.

IV. Experimental Setup Implementation

The designed controller is created and adjusted using MATLAB/Simulink. After getting and improving the results in the simulation environment, the testing was done at the experimental setup, installed in the laboratory. The dSPACE card is used as an interface between the experimental setup and the computer. As shown in Fig. 4, the signals that are acquired from the sensors are transmitted to the dSPACE card, through

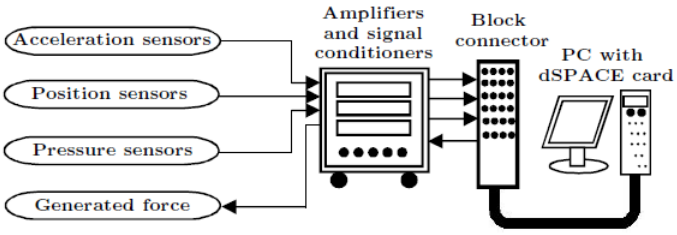


Fig. 4. Shematic view of the experimental equipment

a series of the amplifiers and signal conditioners. Likewise, output signals of the dSPACE card are through amplifiers transferred to the under-seat construction. MATLAB is used as a sofware interface between dSPACE module and the seat model.

V. Simulation and Experimental Results

All simulations and experiments on the physical model, were conducted in a duration of $25\,s$, what is considered to be enough time for the unwanted transients to subside.

A. Simulation results

According to section II, a MATLAB/Simulink model was developed for simulation. The best result of the simulation is SEAT $= 0.35$, with $em3$ test signal and satisfied limitation of the seat deflection. This is improved result comparison it to the other suspension techniques applied to the off-road vehicle seat [14]–[19]. A diagram of the cabin floor acceleration and acceleration of the seat obtained with pneumatic active suspension, is shown in the Fig. 5. It is easy to notice that most of the time the acceleration of the seat is much lower than the acceleration of the cabin floor, which means that designed control is quite good. The main characteristic of the system is its high sensitivity. Particularly, if the fuzzy rules are not precise designed, small changes in control force tend to augment deflection out of its limits. Fig. 6 and Fig. 7 show comparison of pneumatic and hydraulic active suspension concept. Improvement of the SEAT value about 5% with pneumatic concept is obtained. Significantly lower forces generated by the active air-spring imply less power consumption and more comfortable driving. With the designed fuzzy controller the deflection values do not overstep the allowed range (see Fig. 8). The highest value of the seat deflection is $4\,$cm. When the deflection value is not near the limit, the main goal is reducing the acceleration of the seat as close to zero as possible. This means that the resulting forces need to be small. But in the critical situations, high intensity of the force has to be chosen to keep the deflection value within its range and has a negative impact at the SEAT value. In Fig. 8 it can be noticed that the deflection amplitudes obtained with pneumatic suspension are lower over the time, what improves trip comfort.

B. Experimental results

The result of the control algorithm applied to the physical setup was not as good as in the simulation, as expectable. The

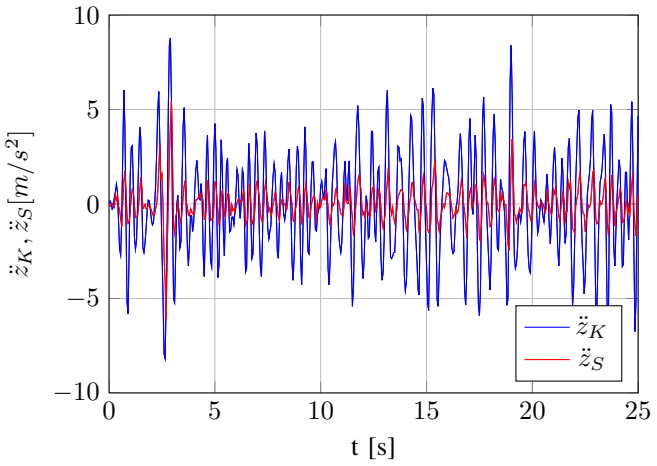


Fig. 5. Acceleration of the seat and the cabine floor with pneumatic active suspension

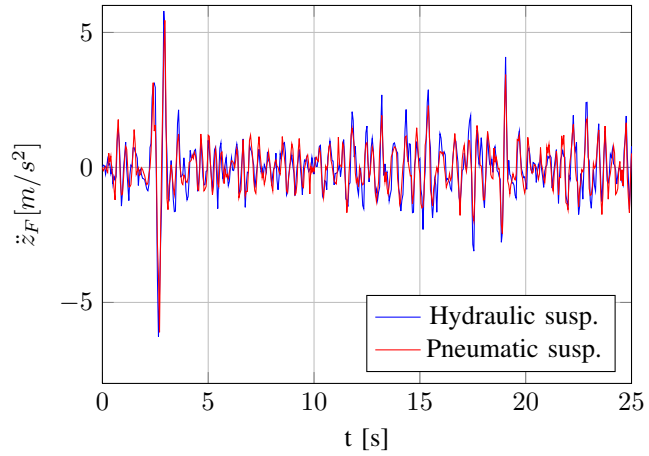


Fig. 6. Comparison of the driver accelerations obtained with the pneumatic and hydraulic active suspension

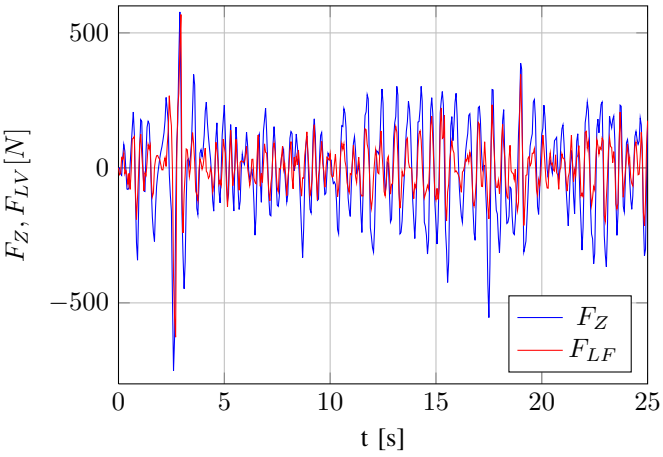


Fig. 7. Comparison of the driver forces obtained with the pneumatic and hydraulic active suspension

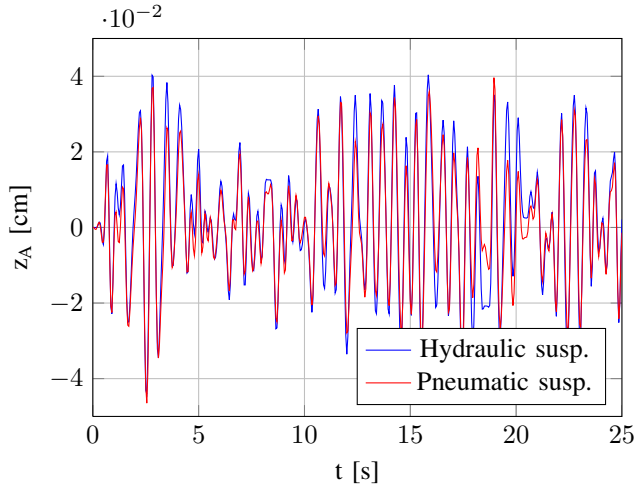


Fig. 8. Comparison of the deflection obtained with the pneumatic and hydraulic active suspension

results differ mostly because of the unmodelled dynamics in the simulation mode (e.g. exact friction forces and noise are very difficult to model). Hence, the seat value SEAT $= 0.46$ is obtained, which is improved about 6% comparing to the result provided by the hydraulic suspension.

VI. Conclusion

The low-frequent kind of a heavy vehicle vibrations are mainly responsible for the driver's comfort. Different types of suspension are known: passive, semi-active and active suspension systems. These concepts differ in complexity and power requirements. In this paper, active suspension system is described with active air spring and passive damper mounted to the under-seat construction. Proposed fuzzy controllers guarantee significant vibrations reduction and robust stability for the toughest conditions of the bumpy road. The validity of the proposed control structure is verified through both, simulation and experiment. The deflection is always in the predefined limits, even at the real model. Also, in order to maintain the deflection in its limits, in [14]–[18] SEAT value is sligthly increased, which is not an issue for the control system described in this paper. Lower energy consumption and more comfortable trip are achived comparing to the hydraulic active suspension system [19]. Ongoing work involves embedding of this active suspension system in an off-roade vehicle.

Acknowledgment

All the experimental work was conducted at the Department for Automatic Control of the Friedrich Alexander University in Erlangen – Nürnberg, and supported by DAAD.

References

[1] I. O. for Standardization ISO 2631-1, "Mechanical vibration and shock - Evaluation of human exposure to whole body vibration - Part 1. General requirement," pp. 1 –31, 1997.

[2] I. O. for Standardization ISO 7096, "Earth-moving machinery - Laboratory evaluation of operator seat vibration," pp. 1 –21, 2000.

[3] C.-H. Chuang and Q. Wang, "Active controllers for reducing vibration amplitude," in *American Control Conference*, jun. 1992, pp. 1611 –1615.

[4] A. Connolly, M. Green, J. Chicharo, and R. Bitmead, "The design of LQG and H_∞ controllers for use in active vibration control and narrow band disturbance rejection," in *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, vol. 3, dec. 1995, pp. 2982 –2987.

[5] I. Fialho and G. Balas, "Road adaptive active suspension design using linear parameter-varying gain-scheduling," *Control Systems Technology, IEEE Transactions on*, vol. 10, no. 1, pp. 43 –54, jan. 2002.

[6] H. Sayyaadi and N. Shokouhi, "Improvement of passengers ride comfort in rail vehicles equipped with air springs," *World Academy of Science, Engineering and Technology 53*, pp. 827 –833, 2009.

[7] ——, "New Dynamics Model for Rail Vehicles and Optimizing Air Suspension Parameters Using GA," *Sciencia Iranica Journal, Transactions B*, pp. 496 –512, 2009.

[8] H. V. Deo and N. P. Suh, "Axiomatic design of a customizable automotive suspension," in *Axiomatic Design, Proceedings of ICAD 2004. International Conference on*, jun. 2004.

[9] ——, "Variable stiffness and variable ride-height suspension system and applications to improved vehicle dynamics," in *Transactions of the SAE 2005. World Congress*, apr. 2005.

[10] ——, "Pneumatic suspension system with independent control of damping, stiffness and ride-height," in *Axiomatic Design, Proceedings of ICAD 2006. International Conference on*, jun. 2006.

[11] T. T. Nguyen, V. G. Ngyuen, and S. B. Kim, "Control of active suspension system by using h_∞ theory," vol. 2, no. 1, pp. 1 –6, mar. 2000.

[12] J. Cao, H. Liu, P. Li, and D. J. Brown, "State of the art in vehicle active suspension adaptive control systems based on intelligent methodologies," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 9, no. 3, pp. 392 –405, sept. 2008.

[13] J. Cao, P. Li, and H. Liu, "An interval fuzzy controller for vehicle active suspension systems," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 11, no. 4, pp. 885 –895, dec. 2010.

[14] A. Zikrija, C. Senad, and K. Samim, "Reducing vibration of the seat with semi-active damper by using the artificial neural networks," in *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, aug. 2007, pp. 125 –130.

[15] Z. Avdagic, S. Konjicija, and S. Ruhotina, "Use of artificial neural networks in different approaches for reducing vibration of driver's seat," in *Industrial Informatics, 2009. INDIN 2009. 7th IEEE International Conference on*, jun. 2009, pp. 313 –318.

[16] O. Tanovic, S. Huseinbegovic, and B. Lacevic, "Road type recognition using neural networks for vehicle seat vibration damping," in *Signal Processing and Information Technology, 2008. ISSPIT 2008. IEEE International Symposium on*, dec. 2008, pp. 320 –323.

[17] S. Huseinbegovic and O. Tanovic, "Adjusting stiffness of air spring and damping of oil damper using fuzzy controller for vehicle seat vibration isolation," in *Control and Communications, 2009. SIBCON 2009. International Siberian Conference on*, mar. 2009, pp. 83 –92.

[18] O. Tanovic and S. Huseinbegovic, "Hybrid fuzzy-neural network structure for vehicle seat vibration isolation," in *Control and Automation, 2009. ICCA 2009. IEEE International Conference on*, dec. 2009, pp. 2354 –2359.

[19] A. Salihbegovic, E. Sokic, A. Salihbegovic, and M. Hebibovic, "Reducing off-road vehicle seat vibrations using hydraulic piston and fuzzy logic controller," in *ELMAR, 2012 Proceedings*, sept. 2012, pp. 133 –136.

[20] G. Meier, G. Roppenecker, and C. Wurmthaler, "Regelungskonzepte für Fahrzeugsitze mit semi-aktiver Dämpfung," *Verein Deutscher Ingenieure Berichte*, no. 1672, pp. 53 –59, 2002.

[21] A. Kühnlein and C. Wurmthaler, "Aktive Federung eines Fahrersitzes mittels Idealmodell," *Verein Deutscher Ingenieure Berichte*, no. 1971, pp. 785 –792, 2007.

[22] A. Kühnlein, "Regelung eines aktiven Nutzfahrerzeugsitzes mittels Idealmodell," *Verein Deutscher Ingenieure Berichte*, no. 2002, pp. 521 –535, 2007.

Comparative analysis between methodologies and their software realizations applied to modeling and simulation of industrial thermal processes

B.Hr. Dimitrov*, H. B. Nenov ** and A.St. Marinov***
* Technical University of Varna/Department of electrical engineering, Varna, Bulgaria
** Technical University of Varna/Department of computer science, Varna, Bulgaria
*** Technical University of Varna/Department of electronics and microelectronics, Varna, Bulgaria
bdimitrov@processmodeling.org, ico762001@yahoo.com, a.marinov@tu-varna.bg

Abstract - The current paper presents a comparative analysis between numerical computer based methods applied towards modeling of thermal processes. The processes that are analyzed concern equipment based on electrical resistance heating - such as chamber resistive, blast and through furnaces. The study applies computer based modeling, using the finite element method (FEM) and the system of ordinary differential equations (SODE) method in order to solve different problems involving technological thermal processes such as: controlled heating for quenching, melting, forced cooling and others. Modeling capabilities for temperature control trough hysteresis and PID regulation are shown. Results from using FEM and SODE for solving those different tasks are combined to form a uniform comparative study that formulates recommendations for selecting the most appropriate modeling method for the specific thermal process.

I. INTRODUCTION

Physical processes in electro thermal devices, their analysis and the methods for modeling are presented in detail in various literature sources [5,6]. Despite the advances in this area, the design of some particular devices such as chamber resistive furnaces (CRF), which are the object of discussion in the current paper, could benefit from application of specialized computer based modeling methods such as FEM and SODE. Those methods can be applied to: analyzing transient processes of heating in the thermal load - when: quenching; regulating the temperature based on hysteresis relay or PID control; controlled forced cooling for normalizing; and others.

The present paper aims at suggesting a comparative analysis and evaluation of the two modeling methods - FEM and SODE, applied to modeling of thermal processes in CRFs. The obtained results will be used to form a set of suggestions for applications of the aforementioned methods directly into operating equipment. This will allow analyzing processes in thermal equipment and thermal loads for which there is no information, experimental data and others. Based on such analysis regulators for the equipment can be adjusted with a minimum of experimental data.

The suggested study is based on multiple experiments and dedicated models of CRF for different applications. The two modeling methods FEM and SODE are compared based on experimental data. The complete analysis includes a variety of thermal equipment (example is shown on figure 1) - high and middle temperature furnaces with different construction for laboratory and industrial applications. The experimental study is not only limited to chamber furnaces but also includes: blast, trough and other furnaces. Experiments are conducted using an infra red (IR) thermal camera and thermocouples. Figure 2 shows a sample measurement.

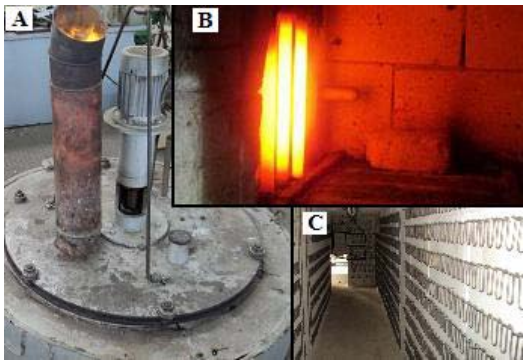


Figure 1. CRF, object of the suggested study: A - blacst furnace ;B - high teperature chamber furnace with sillicon carbide heaters; C - trough furnace with continiuous operation;

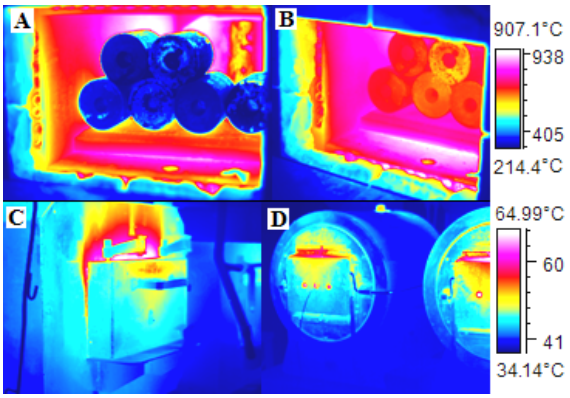


Figure 2. Experimental study with an IR thermal camera , used to confirm results of the developed models. A, B – respectivly - beging and end of the heating proces in a CRF loaded with non-heated blanks.; C, D – thermal distribution on the housing and door of a idustrial (C) and laboratory (D) furnaces

Bulgarian scientific fund

II. ANALYSIS

A. *Modeling of CRF using Finite Elements Method (FEM)*

Modeling the thermal processes with FEM is discussed in various sources and is well backed by different software products [3,7,8,9]. Thus essentials of modeling with FEM are not presented in detail in the current paper. Sample model developed based on FEM is shown on figure 3. The FEM models allow the analysis of the distribution of thermal fields, as the main elements subjugated to examination are: the surface of the studied CRF; the walls and housing of the CRF; thermal insulation.

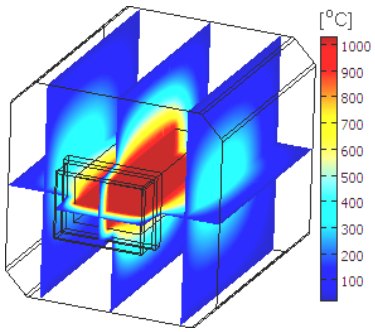


Figure 3. Model of a CRF developed based on FEM

B. *Modeling of CRF using systems of ordinary differential equations (SODE)*

When modeling thermal processes with SODE, equivalent circuits based on analogy between electrical and thermal processes are used. For the purposes of the current paper the SODE methodology is implemented in MATLAB.

It has to be noted that using equivalent circuits is a dated method and its application should be avoided. Despite this its usage for the current study and thermal equipment could be justified. Several points compared to FEM could be indicated:

- It is possible to model the regulation process of the temperature inside the working space of the furnace. This requires that the equations describing the regulator (hysteresis or PID controller) are initialized in the workspace environment of the software product that secures the mathematical calculations - for example MATLAB.
- It is possible to model two dimensional (2D) and three dimensional (3D) models that can reflect the thermal process. In this case MATLAB offers a rich mathematical device that allows working with SODE, such as - ode15, ode45, ode 23 [8]. The computing procedures in the current work are done using Runge-Kutta methods and their modifications [2].
- Acceptable accuracy compared to FEM.
- Simplified computed procedures, especially when working with 3D models, respectively lesser hardware demand.

An example for a linear equivalent circuit for modeling thermal processes in CRF is shown on figure 4. Analogically a 2D and 3D model can be devised. In the case the model shown in figure 4 reflects a chamber furnace (figure 1B) constructed from a high temperature refractory and insulation layer. The thermal load that is subjected to heating can be a monolith blank or a container filled with blanks. The model is devised into two layers and a center, which allows monitoring the process of heating in depth and determining the thermal differences between the surface and the center.

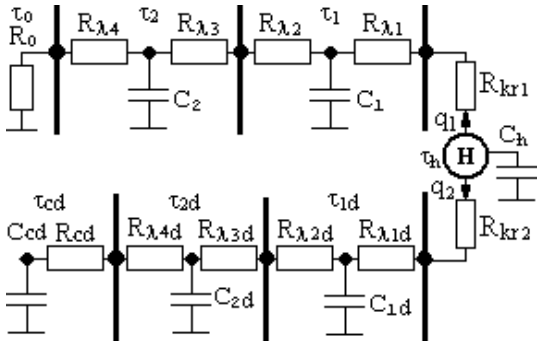


Figure 4. Equivelent circuit used to model thermal procesess in complex model of a furnace and thermal load.

The symbols in figure 4 represent:

- $R_{\lambda 1 \div 4}$ - resistance of the conductive heat transfer trough the walls of the furnace. They conduct the heat flow of the losses q_1 - form the working space to the outside environment.
- R_0, R_{kr1}, R_{kr2} - resistances of convection and radiation, respectively from the housing of the CRF towards the outside environment and from the heater H towards the surface of the refractory layer and the heated thermal load.
- $R_{\lambda 1 \div 4d}$, R_{cd} - resistances of the conductive heat transfer in the heated thermal load. The useful heat flow q_2 is conducted trough them.
- C_1, C_2, C_{1d}, C_{2d}, C_{cd}, C_h - heat capacity, respectively of the two layers of the furnace and the two layers and the center of the heated thermal load and the heater.
- τ_h, τ_1, τ_2, τ_0, τ_{1d}, τ_{2d}, τ_{cd} - temperatures, respectively: of the heater; of the two layers of the furnace; the outside environment; the two layers and the center of the heated thermal load. Temperatures are given as initial values and are determined on each iteration of the computing procedure. Only the temperature of the outside environment τ_0 is excluded from this rule - it is set as constant. In this way a constrain from third order is set.

All resistances and capacitors are variable, their values are determined at each iteration of the computing procedure. In this way an account of the functional relation between the coefficients of the heat transfer - $\lambda = f(\tau)$ and their specific heat capacity $c = f(\tau)$ for materials and temperature can be made.

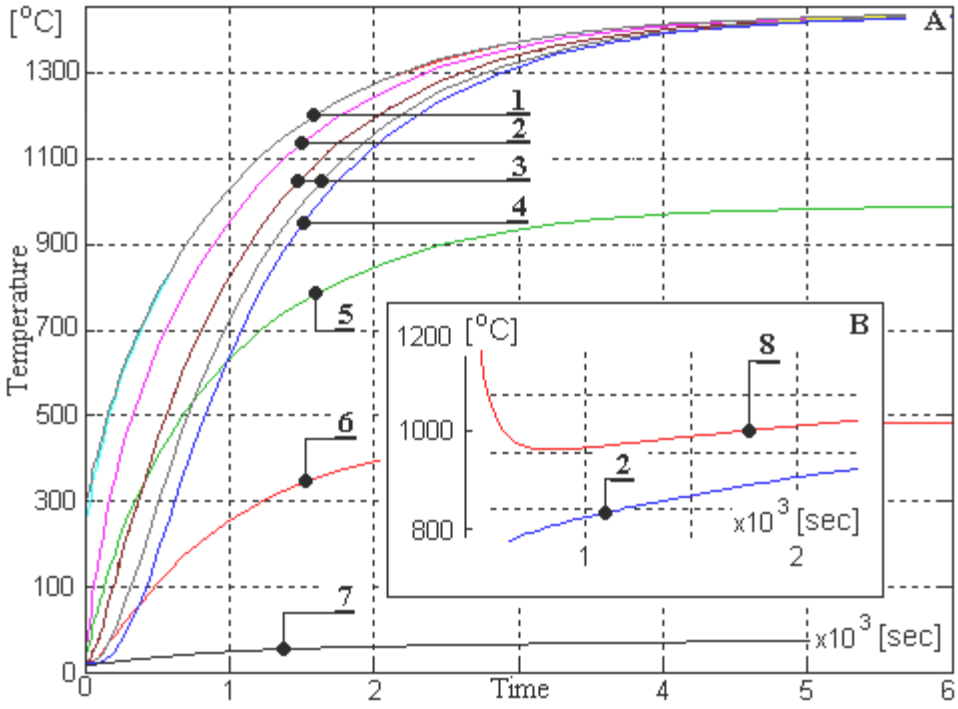


Figure5. Transient process of heating a CRF - obtained trough SODE modeling: A – Steady-state of the temperature in the high temperature furnace. B – loading of the heated furnce with a thermal load with enviroment temperature. Used indexing: 1 – temperature of the heater; 2,3,4 - respectivly - suface, indepth layers and the center of the thermal load; 5,6, 7 – respectivly - temperature of the refractory and the thermal isulation layers of the furnace and its housing; 8 - temperature in the furnace space - detrmined by a thermo couple included in the model.

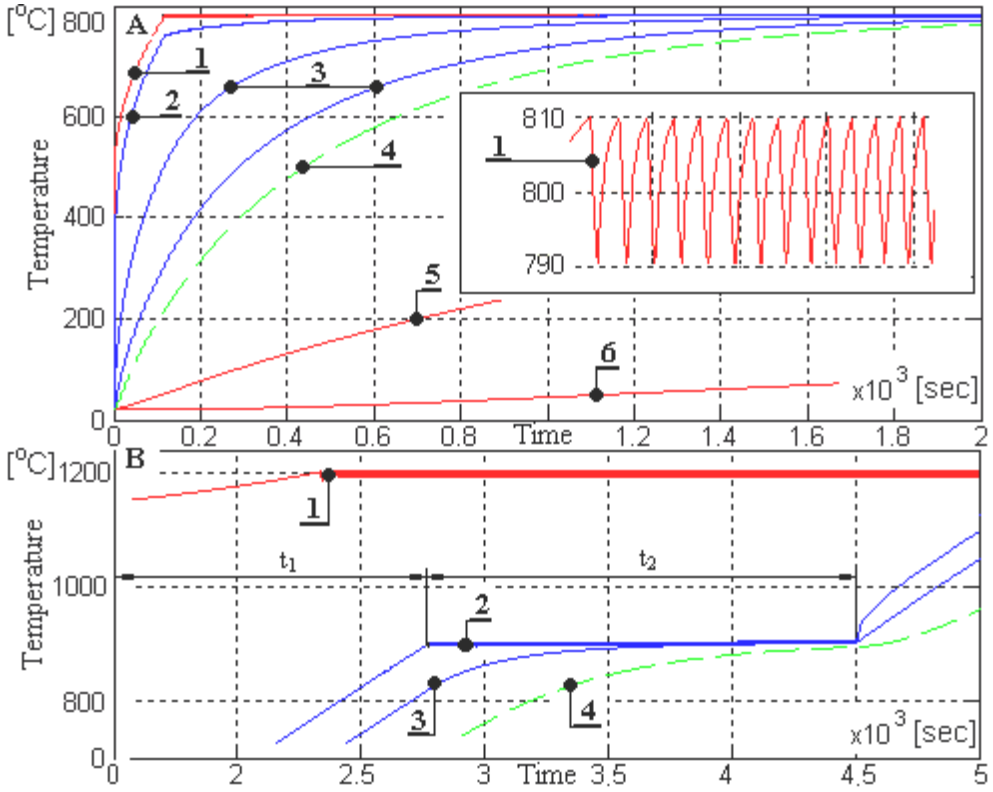


Figure 6. Simulation of the work process in the multitude - furnace-thermal laod-thermoregulator. A – process of regulation with set temperature in the thermal load. B – Melting process. Used indexes: 1 – temperature of the heater; 2,3,4 - respectivly - suface, indepth layers and the center of the thermal load; 5,6 – respectivly - temperature of the refractory and the theroisulation layers of the furnace

Based on the equivalent circuit a set of equations that describe the thermal process can be drawn, as follows:

$$
\begin{aligned}
&q = q_1 + q_2 = \frac{\tau_h - \tau_1}{R_{kr1} + R_{\lambda 1}} + \frac{\tau_h - \tau_{1d}}{R_{kr2} + R_{\lambda 1d}} \\
&C_1 \frac{d\tau_1}{dt} + \frac{\tau_1 - \tau_2}{R_{\lambda 2} + R_{\lambda 3}} = \frac{\tau_h - \tau_1}{R_{kr1} + R_{\lambda 1}} \\
&C_2 \frac{d\tau_2}{dt} + \frac{\tau_2}{R_{\lambda 4} + R_0} = \frac{\tau_1 - \tau_2}{R_2 + R_3} \\
&C_{1d} \frac{d\tau_{1d}}{dt} + \frac{\tau_{1d} - \tau_{2d}}{R_{\lambda 2d} + R_{\lambda 3d}} = \frac{\tau_h - \tau_{1d}}{R_{kr2} + R_{\lambda 1d}} \\
&C_{2d} \frac{d\tau_{2d}}{dt} + \frac{\tau_{2d} - \tau_{cd}}{R_{\lambda 4d} + R_{cd}} = \frac{\tau_{1d} - \tau_{2d}}{R_{\lambda 3d} + R_{\lambda 4d}} \\
&C_{cd} \frac{d\tau_{cd}}{dt} = \frac{\tau_{2d} - \tau_{cd}}{R_{cd} + R_{\lambda 4d}} \\
&\tau_h = \frac{q + \dfrac{\tau_1}{R_{kr1} + R_{\lambda 1}} + \dfrac{\tau_{1d}}{R_{kr2} + R_{\lambda 1d}}}{\dfrac{1}{R_{kr1} + R_{\lambda 1}} + \dfrac{1}{R_{kr2} + R_{\lambda 1d}}}
\end{aligned}
\tag{1}
$$

The software representation of the equations is a system of equations that can be given as:

$$
\begin{aligned}
&C_1 \frac{d\tau_1}{dt} = -\left(\frac{1}{R_{kr1} + R_{\lambda 1} + R_{kr2} + R_{\lambda 1d}} + \frac{1}{R_{\lambda 2} + R_{\lambda 3}} \right).\tau_1 + \\
&\left(\frac{1}{R_{\lambda 2} + R_{\lambda 3}} \right).\tau_2 + \left(\frac{1}{R_{kr1} + R_{\lambda 1} + R_{kr2} + R_{\lambda 1d}} \right).\tau_{1d} + \\
&+ \frac{q}{1 + \left(\dfrac{R_{kr1} + R_{\lambda 1}}{R_{kr2} + R_{\lambda 1d}} \right)} \\
&C_2 \frac{d\tau_2}{dt} = \left(\frac{1}{R_{\lambda 2} + R_{\lambda 3}} \right).\tau_1 - \left(\frac{1}{R_{\lambda 2} + R_{\lambda 3}} + \frac{1}{R_{\lambda 4} + R_0} \right).\tau_2 \\
&C_{1d} \frac{d\tau_{1d}}{dt} = -\left(\frac{1}{R_{kr1} + R_{\lambda 1} + R_{kr2} + R_{\lambda 1d}} + \frac{1}{R_{\lambda 2d} + R_{\lambda 3d}} \right).\tau_{1d} \\
&+ \left(\frac{1}{R_{kr1} + R_{\lambda 1} + R_{kr2} + R_{\lambda 1d}} \right).\tau_1 + \left(\frac{1}{R_{\lambda 2d} + R_{\lambda 3d}} \right).\tau_{2d} \\
&+ \frac{q}{1 + \left(\dfrac{R_{kr2} + R_{\lambda 1d}}{R_{kr1} + R_{\lambda 1}} \right)} \\
&C_{2d} \frac{d\tau_{2d}}{dt} = \left(\frac{1}{R_{\lambda 2d} + R_{\lambda 3d}} \right).\tau_{1d} - \\
&- \left(\frac{1}{R_{\lambda 2d} + R_{\lambda 3d}} + \frac{1}{R_{\lambda 4d} + R_{cd}} \right).\tau_{2d} + \left(\frac{1}{R_{\lambda 4d} + R_{cd}} \right).\tau_{cd} \\
&C_{cd} \frac{d\tau_{cd}}{dt} = \left(\frac{1}{R_{\lambda 4d} + R_{cd}} \right).\tau_{2d} - \left(\frac{1}{R_{\lambda 4d} + R_{cd}} \right).\tau_{cd}
\end{aligned}
\tag{2}
$$

Results from the computation procedures are realized as 3D models, and are shown of figure 5 and figure 6. Figure 5A shows a transient process of the heating in a CRF. The simulation allows to determine the temperatures: of the heater (figure 5A - 1); the in-depth distribution in the load (figure 5A- 2, 3, 4); the walls of the furnace (figure 5A -5,6) and its housing (figure 5A -7).

The process of loading the already heated furnace is shown at figure 5B. Figure 5B -8 shows the decrease in the temperature in the heating space resulting: due to the placements of the thermal load that has to be processed - the temperature the element corresponds to one of the environment; the losses due to the opening of the furnace's door. This model allows to determine the thermal loading (thermal shock) in the blank that will be processed.

Figure 6A shows the process of temperature control using a thermal regulator. The model allows determining the required tuning of the regulator that will secure the desired technological process. As an example, when working with a PID regulator, the simulation procedure can help determine the coefficients of the proportional, the differential and the integration elements.

The process of melting, based on simulation is shown on figure 6B. The model allows determining the time required to reach the melting temperature - t_1 and the melting time - t_2.

C. Comparative analysis between FEM and SODE, when modeling CRF.

The analysis aims at determining the simulation errors when modeling CRF using FEM and SODE - based on experimental data. The experiments are conducted with several industrial chamber furnaces that heat blanks with different parameters.

Figure 7A shows a comparison between 3D models based on SODE (Figure 7A - 1), FEM (Figure 7A -2) and experimental data (Figure 7A - 3). The accumulated error is unacceptably high; this is due to the idealization of the developed model, done so the simulation can fit the existing computation equipment. The models exclude some depreciation defects - this introduces an extra error. This requires an additional specification to the models based on experimental data. The corrections are made on the functional relations $\lambda = f(\tau)$, $c = f(\tau)$ and the radiation coefficients in the chamber. Thus in the working range in medium and high temperature furnaces (above 500^{o}C to nominal) an accuracy of 5 to 10% for both methods can be achieved (figure 7B). As it can be seen in the same figure in the beginning of the heating process the error can exceed 100%. This is due to the model of convection heat transfer inside the furnace. In order to simplify computing procedures the convection heat transfer is set as zero, because its main effect is in the lower temperature region. This can be deemed acceptable, as in the furnaces that are objected to the current analysis low temperature processes do not take part.

The suggested approach is not firm, but in the case can be considered acceptable. In this way an satisfactory accuracy can be achieved, that can be used to model the process of control using a thermal regulator. Figure 7C shows results obtained through the application of SODE model (Figure 7C - 1,2), in the given case used to determine the coefficients of a PID regulator. The

aforementioned are used to control the experimental process described in Figure 7C -3.

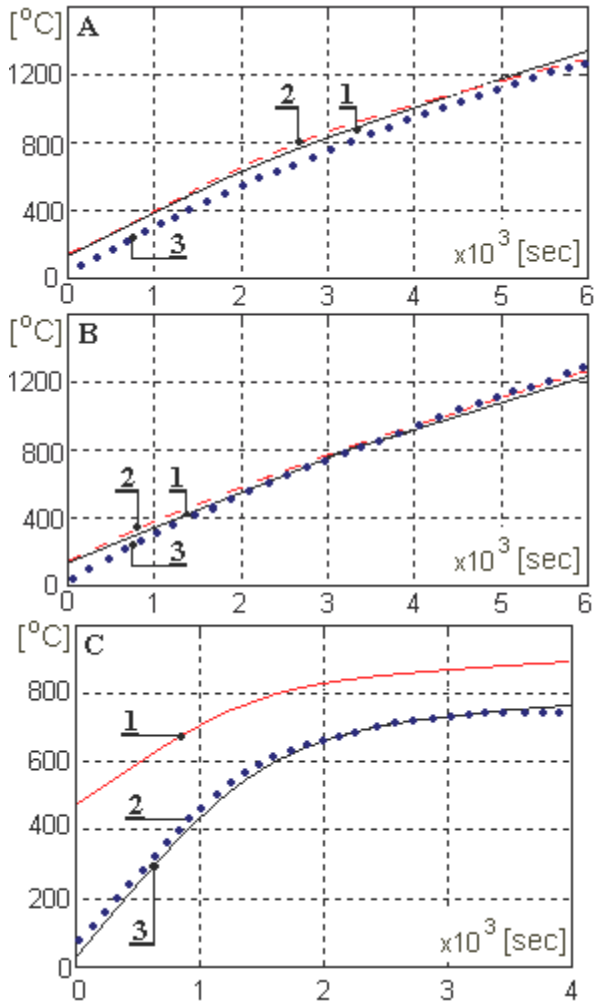


Figure 7. Comparison between results using models developed using FEM and SODE and experimental data; A, B: 1 – SODE; 2 – FEM; 3 – Experimental data; C – control usign PID regulator: 1, 2 – temperature

D. Formulating suggestions for the usage of FEM and SODE when modeling CRFs.

Models can be included in the documentation of the furnace that they are developed for - this will make their usage more flexible and efficient.

Based on the results obtained trough modeling and experiment, the following suggestion when working with FEM and SODE for describing electrical resistance furnaces can be made:

- Models developed using FEM provide visual representation of the distribution of the thermal fields and flows in different equipment. Those models are suitable on design and scientific level and require significant hardware computation recourses.
- For practical application models developed using SODE can be used. They offer the capability to analyze thermal transients in different technological processes. SODE models require significantly less computation recourses compared to FEM
- Exploitation of furnaces is linked to a certain depreciation of the used equipment. This subjects the model to a periodical check of its adequacy and correction of its parameters.

III. CONLUSION

The results from the conducted study for the considered equipment show that both approaches - using SODE and FEM are applicable to analyzing thermal processes in CRF.

Computing procedures realized using SODE require relatively less hardware resources. This makes them significantly applicable in practice where computer systems with lesser hardware capabilities compared to laboratory equipment are used. This allows different analyses on thermal processes to be conducted using general computational equipment. The analyses can be executed for new thermal loads where information and data is not available. Based on data acquired from the analysis a fine tuning on regulators can be achieved. This reduces the initial experiments for adjusting the regulator to a minimum, which leads to increase in energy efficiency.

ACKNOWLEDGMENT

This paper is developed in the frames of project “Increasing energy efficiency and optimization of electrotechnological processes and devices”, № МУ03/163 financed by the National Science Fund

REFERENCES

[1] Bradley L., Numerical Solution of Differential Eqations. Department of Computer Science, University of Colorado 1998.

[2] Gerstmayr J., Stangl M., Dibold M., High-Order Implicit Runge-Kutta Methods for Discontinuous Multibody Systems. Martin-Luther-University Halle-Wittenberg, Germany, 2004.

[3] Розин Л. А., Метод конечных элементов. Санкт-Петербургский государственый университет. 2000 год

[4] Мэтьюз Дж. Г., Куртис Д. Ф. Численные методы изпользование MATLAB. “Вильямс” Москва 2001 год

[5] Димитров Б. Анализ на изчислителния процес за математичен модел на съпротивителни електропещи. Списание Електротехника и електроника Е+Е 11-12 2006 страници 78-83.

[6] Димитров Б. Анализ и моделиране на електотермични процеси и устройства /камерни електросъпротивителни пещи с периодично действие/. Докторска дисертация, ТУ-Варна 2008.

[7] http://www.comsol.com/

[8] http://www.mathworks.com/

[9] http://www.ansys.com/

Twenty years of monitoring and control of electricity consumption in RTB Bor, Serbia

Viša Tasić*, Vladimir Despotović**, Darko Brodić**, Marijana Pavlov*, Dragan R. Milivojević*
*Mining and Metallurgy Institute, Bor, Serbia
** University of Belgrade, Technical Faculty in Bor, Bor, Serbia
e-mail: visa.tasic@irmbor.co.rs

Abstract - This paper presents a historical overview of development and application of a distributed system for monitoring and control of electrical energy consumption in the Copper Mining and Smelting Complex Bor (RTB), which is one of the greatest consumers of electrical energy in the Republic of Serbia. Realization of the real time monitoring and control system started in early 1990s. The main characteristics and performances of three generations of the system are described in this paper.

I. INTRODUCTION

Tracking electrical energy consumption in real time is one of the most important issues for companies those are large power consumers in order to control energy usage and avoid peak pricing. Monitoring system that collects data about active and reactive power and energy makes it easier for companies to detect potential power concerns. This enables the reduction of energy consumption per unit of production and makes it more predictable.

The overall electrical energy costs consist of two main components: the energy costs and the power costs. The power costs imply the peak load, which is the maximum average power load over a designated interval of time (15 minutes) during one month. It is apparent that, under certain circumstances, 15 minutes of unusually high-energy import resulting from concurrent start-up of power consumers may drastically increase power costs for an entire month. This can be often avoided with only a minimal technical effort. Peak load costs can be significantly reduced by just a few short, targeted interventions at the right point in time. Potential savings in this area remain unexploited by many companies. This short-term redistribution of imported energy is usually executed automatically using optimization systems for peak load [1].

The Copper Mining and Smelting Complex Bor (RTB) is one of the largest electric energy consumers in the Republic of Serbia. The consumption in its production plants (mines, flotation, smelter and refining plants) is over 50 GWh per month, which amounts approximately 2000000 EUR/month. Only modest savings in electrical energy consumption can lead to substantial cost savings. This is achieved using the monitoring and control system designed and implemented in RTB in early 1990s [2].

The system is constantly upgraded and improved until nowadays, both in terms of hardware devices and software support. Because of the complexity, a large number of transformer stations and their spatial dislocation (over 80 km), system includes only main transformer substations.

This paper gives a historical overview of three generations of the system for monitoring and control of electrical energy consumption. The first two generations were designed only for monitoring, whereas the latest generation has a control possibility for peak load reduction.

II. THE FIRST GENERATION (1990-1996)

The first generation of the system for monitoring and control of electrical energy consumption in RTB Bor was designed and implemented at the beginning of 1990 [2]. The implemented system had the following structure (as shown in Figure 1):

- Distributed measuring stations (MS) installed at each transformer station,
- Data's concentrators and remote workstations (RWS) installed at Control Centers in Bor and Majdanpek,
- Host computers installed at Computer Centers in Bor (ICL 2958D) and Majdanpek (ICL ME29).

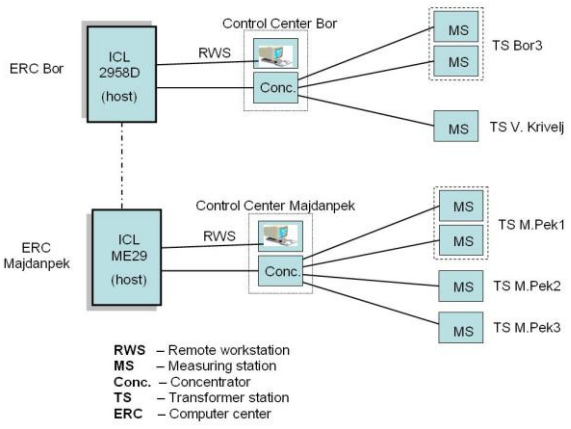


Figure 1. The configuration diagram of monitoring system

The integral monitoring system was organized as a three-level MAN network. Data's concentrator (master node) was connected via base-band modems and leased lines with the measuring stations (slaves) at one, and host computer, at the other side. Communication protocol, BSCP (Binary Synchronous Communication Protocol) [3] was specially designed and applied in order to support real-time data transfer between all network nodes.

Figure 2. The measuring station (first generation) with I/O modules and power transducers

The preparation of the transformer stations was carried out in such a way that the power transducers were installed into the measuring circuit of secondary cells of all transformers. Their basic characteristics were: Aron's measuring methods, input voltage 3x100 V symmetrically, output signals 4-20 mA, unlinearity < 1% and overloads up to 20%. Output signals from the power transducers were supplied to the inputs of the measuring station, which carried out the measuring process. Power transducers, as well as measuring stations, were designed and developed at the Mining and Metallurgy Institute Bor, Department of Industrial Informatics (Figure 2).

Main characteristics of measuring stations were: microprocessor Motorola M6800, 48 (40) KB for data (RAM), 16 (24) KB for software (EPROM), 8-bit A/D converter, 64 analog inputs, 512 digital state signals (input/output) with mutual point, baseband modem for synchronous communication through leased telephone lines (9600 bps) and functional keyboard.

The master network node was data concentrator. It was actually an 8-bit microprocessor system (modified ISKRA DATA 1680 shown in Figures 3 and 4) which was composed of workstation, floppy drive and supported a dot matrix printer [2]. The measurements of the input analog signals were carried out in one-second intervals, while the recording of digital input states was performed in 100 ms intervals. Data's concentrator collects data from measuring stations, performs data processing and creates the outgoing messages about the current state and trends of changes in the electrical energy consumption. Complex calculations and data analysis related to the behavior of the power grid was performed by host computers, ICL 2958D in Bor, and ICL ME29 in Majdanpek [4].

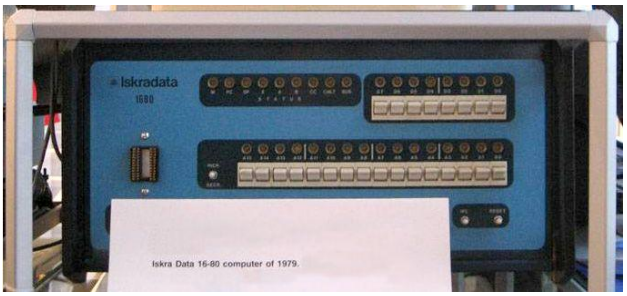


Figure 3. Computer Iskra Data 1680

Figure 4. Workstations at Control Center in Bor (1991)

From a technical point of view, the system showed acceptable stability and reliability. The measurement error was less than 3%. However, the first generation of this system did not include automatic control of peak load, which required organizational measures, i.e. "manual demand management" to prevent concurrent operation of several large power consumers. The inflexibility of this type of system became apparent if an important order suddenly needed to be processed quickly. Despite all, in the early days of application, the system has shown positive financial effects. The peak power costs in Bor3 transformer substation were reduced approximately 800000 USD in 1991 [5].

III. The Second Generation (1996-2006)

During the 1990s, personal computers (PCs) were introduced as the incoming technology, and the company's various plants showed a great interest in a transition from mainframe host computers to PCs. The Control Center purchased its first PCs in the middle of 1990s: an ICL M30 and M45. They supported synchronous communications and could be connected directly to the data concentrators (as shown in Figures 5 and 6). As the data concentrators and measuring stations were not changed, the main improvements were made on the software side. Moreover, the system for monitoring and control of electrical energy consumption became an integral part of the more general supervisory control and data acquisition (SCADA) process-control system, which covered the technological processes in the most important plants of RTB. The real-time application was developed in Borland Turbo Pascal 7.0 development environment and run under the MS DOS operating system (Figure 7). It was

designed to perform data transfer, analyses and real-time visual interpretation of the measurements. The program included three basic modules: communication module for data transfers from the measuring stations, operational module for the real-time data processing and their visual presentation, and module for off-line data processing and archiving. Interaction with the process was obtained under the control of operational program, and it was reflected in forced warning messages on a screen. These messages informed on the limit or maximum power overloads (peak) at each transformer station, or some technological or organizational part of the company. As a reaction to the warning message, power consumers could be switched off or switched on, if it was technically possible and economically justified. An example of off-line application used for data analysis and reporting, that was working in MS Windows environment in the control center in 1996 is shown in Figure 8.

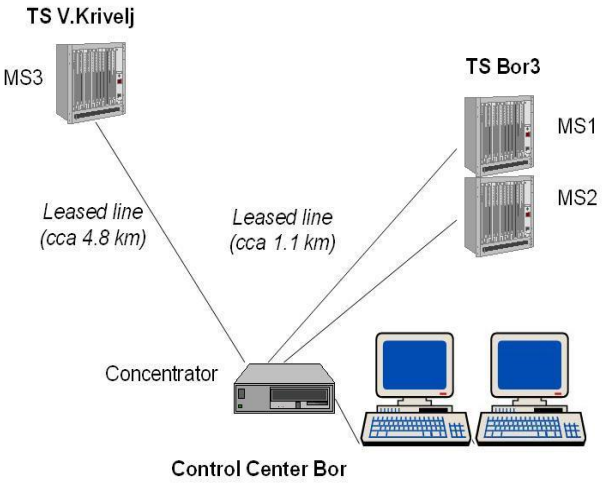


Figure 5. The monitoring system configuration diagram in 1996 in Bor

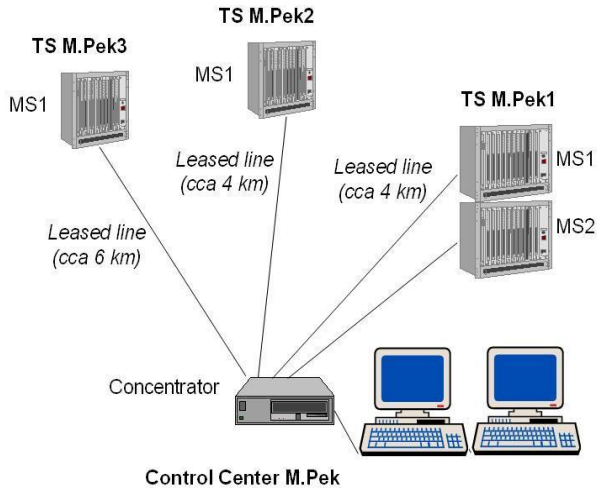


Figure 6. The monitoring system configuration diagram in 1996 in Majdanpek

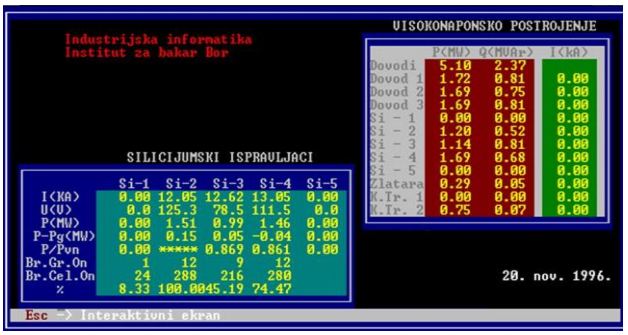


Figure 7. Example of the real-time application working in MS DOS environment in the control center in 1996

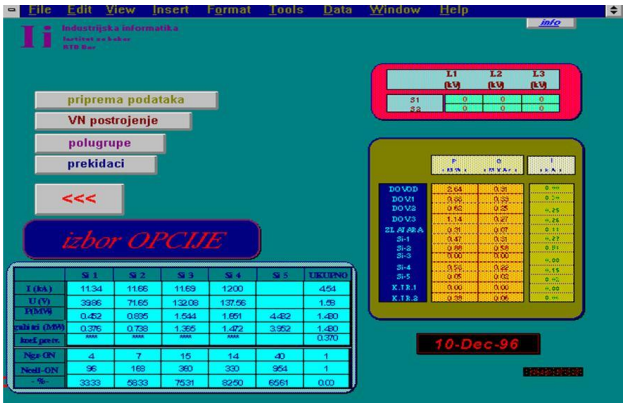


Figure 8. Example of the off-line application working in MS Windows environment in the control center in 1996

IV. The Third Generation (2006-2013)

The next step in modernization of the system for monitoring and control of electrical energy consumption took place in 2006. The newly developed process control system was based on dedicated industrial computers, which could collect the information about the process parameters and execute actions in order to keep or change relevant process conditions. Microprocessor measuring station (MMS), fully designed and developed in the Department of Industrial Informatics of the Institute of Mining and Metallurgy, was used as a programmable logic controller (PLC). The master PC was added as an interactive workstation, forming an entity which was a core of the complex distributed control and monitoring system, as shown in Figure 9.

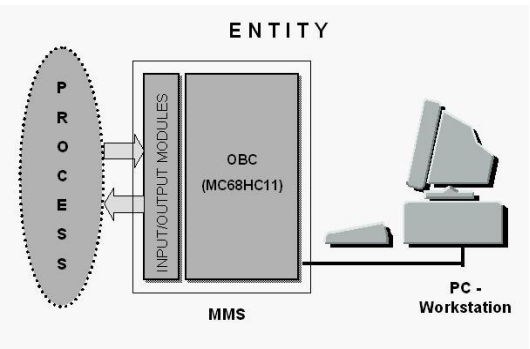


Figure 9. Simplified diagram of a simple control system with MMS

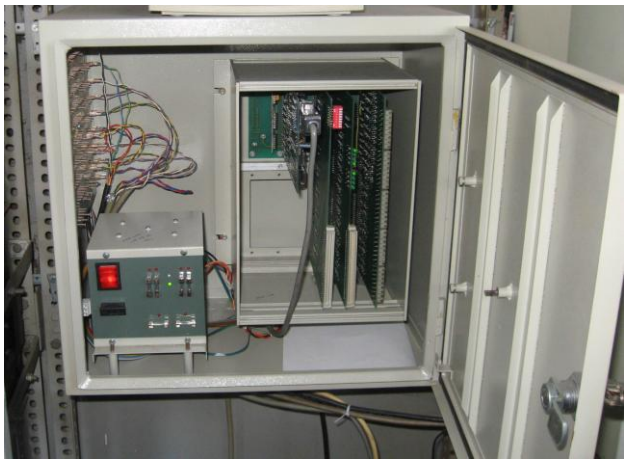


Figure 10. MMS implemented in the transformer station

In functional point of view, MMS (Figure 10) was an autonomous unit, which was used with the aim of data acquisition and command execution. The CPU module was the one-board computer based on MC68HC11 microcontroller [6, 7, and 8]. Main characteristics of the MMS were: microcontroller Motorola 68HC11E, internal eight channel, 8-bit A/D converter, 56 analog inputs, 64 + 64 digital state signals (input + output) with mutual point (or independent), RS232 communication port, 48 (56) KB for data (RAM), 16 (8) KB for software (EPROM), LC display and functional keyboard.

In hardware point of view, MMS contained the serial communication interface with direct back-to-back connection to standard PC RS232 serial port. In order to support the data transfer in both direction's appropriate communication protocol named Asynchronous Serial Protocol (ASP) was developed [6]. The connection between PC and MMS was permanent. The two-node network runs according to the master-slave principle. The PC as a master node was starting the session, maintaining the data transfer and regular end of the session. The MMS had to respond to every PC demand.

The MMS performed measurements according to the algorithm developed for this purpose. The output current signals (4-20 mA) from the power transducers were led to analog inputs of the MMS. It calculated average minute values of electrical power, which were compared to limits given by the algorithm. In this way, it was possible to predict potential exceeding the limit value of the peak load in any 15-minute interval and react accordingly by temporarily switching-off some power consumers.

The special SCADA real-time application, operating in MS Windows environment (Windows 98/2000/XP), was developed in order to support MMS functions and to perform data transfer, analyses and real-time result's interpretation. It was based on client/server architecture running on both master and remote stations, enabling integration in a complex distributed control and monitoring system. SCADA was developed using Microsoft Visual C++ 6.0 development kit [9].

To cover the main transformer stations, it was necessary to realize a complex industrial network [7, 10].

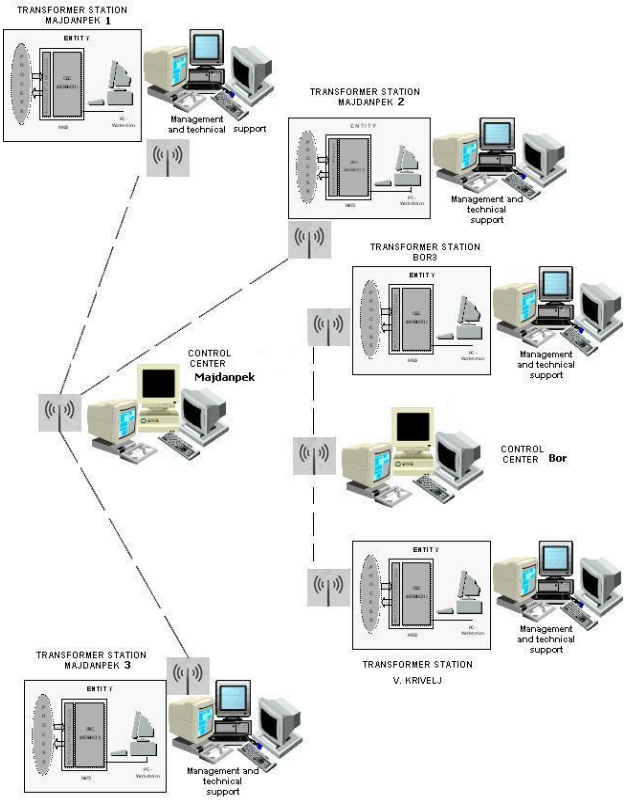


Figure 11. Restructured configuration diagram of monitoring system

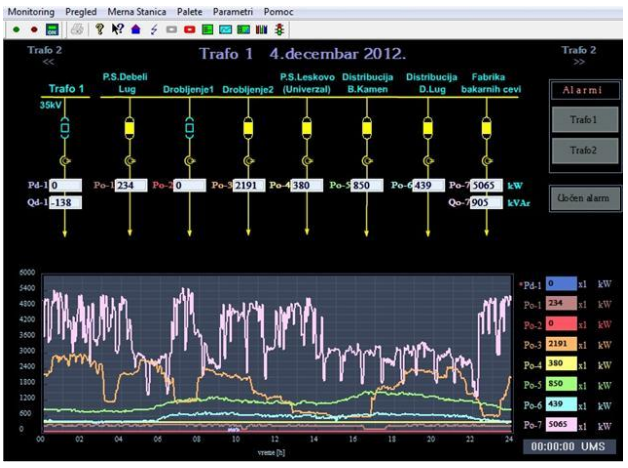


Figure 12. Sinoptic screen of the transformer station

The configuration of the realized network is shown in Figure 11. It consists of heterogeneous nodes, PC workstations and MMS. The workstations act as servers in network configuration, running the appropriate SCADA application (as shown in Figure 12). Remote workstations, running the client versions of the SCADA application, were used as the monitoring terminals (without the control function) for process monitoring on distant locations. One server application could run exclusively on one server workstation. However, all the client applications could run in multitasking at the same workstation. This gives a possibility of monitoring processes locally and at remote plants at the same time.

The system is still in use and covers all main transformer stations of the RTB complex in Bor (Bor 3 and Veliki Krivelj) and Majdanpek (Majdanpek 1, 2 and 3), forming a distributed industrial network in a wide geographical area (within a diameter of 80 km).

CONCLUSION

The system for monitoring and control of electrical energy consumption has been in use for over 20 years in RTB complex Bor. Apart from the favorable price/performance ratio, it has shown to be very reliable and especially resistant to poor communication conditions, thanks to the solid transfer quality control. Three generations of the system were realized in this period, making improvements both in hardware and software points of view.

The presented system can be used for local control, or as integral parts of the distributed control system. The main advantage of given solutions is their modularity, ease of use, low price, and its applicability in different production processes.

An especially important part of the system is the peak load control. Our results have shown that the investment in such a solution can be less than 10% of the yearly benefit due to peak load reduction for large-scale systems.

ACKNOWLEDGMENT

This work was partly funded by the Grant of the Ministry of Education, Science and Technological Development of Republic of Serbia, as a part of Project No. TR33037 “Development and Application of Distributed System for Monitoring and Control of Electrical Energy Consumption for Large Consumers."

REFERENCES

[1] V.Tasić, D.Milivojević, M.Pavlov, V.Despotović, D. Brodić, “Microcontroller Based Systems for Peak Load Reduction,” Proceedings of 35th International Convention, MIPRO 2012, pp. 919-923, Opatija, Croatia, 21.5.-25.5.2012.

[2] M. Radojković, D. Milivojević, G. Jojić-Blagojević and M. Cajić, “System for Monitoring and Control of Electrical Energy Consumption in RTB Bor,” Proceedings of VII Scientific Convention on Microcomputers in Process Control Systems, MIPRO’88, pp. 5-31 – 5-35, Rijeka, Croatia, 18-20.05.1988.

[3] D. Milivojevic, M. Radojkovic, G. Jojic Blagojevic, Dj. Simon, S. Lalovic, “Communication Subsystem of DSKP Systems,” Proceedings of XIV International Conference of Information Technology, Book 1, pp. 173.1-173.5, Sarajevo-Jahorina, BiH, 1990.

[4] D.Milivojević, M.Pavlov, V.Despotović and V.Tasić, “Half a Century of Computing in the Serbia Copper Mining and Metallurgy Industry,” IEEE Annals of the History of Computing, no.3. vol. 34, pp. 34-43, July-September 2012.

[5] J. Petrovic, D. Grubisic, D.R. Milivojevic, M. Radojkovic, G. Jojic Blagojevic, Dj. Simon, S. Lalovic, “The Effects of Control System for Electrical Energy Cosumption in RTB Bor in 1991,” Procedings of 28th October Convention on Mining and Metallurgy, pp. 114-121, Donji Milanovac, Serbia, October 1-3, 1992.

[6] D.R. Milivojevic and V.Tasic, “MMS in Real Industrial Network,” Information Technology and Control, vol. 36, pp. 318-322, 2007.

[7] V. Tasic, D. R. Milivojevic, M. Pavlov and V. Despotovic, “Reducing the Costs of Electricity in the Copper Mining and Smelting Complex RTB Bor (Serbia) by Using of the Distributed Control System,” Proceedings of 33rd International Convention on Information and Communication Techology, Electronics and Microelectronics MIPRO 2010, pp. 664-667, Opatija, Croatia, 24.05.-28.05.2010.

[8] M68HC11 Reference Manual, Motorola Inc., 1990.

[9] B.Stroustrup, The C++ Programming Language , Addison-Wesley Pub Co, 3rd edition , 2000.

[10] D.Milivojević, V.Tasić, M.Pavlov and V.Despotović, “Synthesis of DCS in Copper Metallurgy”, Proceedings of XLII International Scientific Conference on Information, Communication and Energy Systems and Technologies ICEST 2007, Ohrid, FYR Macedonia, Book 2, pp. 629-631, June 2007.

Remote control of model vehicles using Android mobile devices

Bruno Vunderl, Martin Žagar and Danko Basch
Faculty of Electrical Engineering and Computing, Zagreb, Croatia
bruno.vunderl@fer.hr

Abstract - One of the main principles of the smartphone and tablet operating systems is simplicity and accessibility. This feature was usually not available on the embedded systems and devices. By connecting these two types of devices, user interface and accessibility can be simplified. A myriad of new possibilities can be added to the embedded systems while giving smartphone and tablet users control over the external devices. The goal of this project is to create and build a fully functional model of a remotely controlled boat controllable through a mobile application. Connection is established through standard Bluetooth protocol and users are presented with the readings of the sensors found on the model. Boat is controlled by rotating the mobile device laterally and longitudinally and has the ability to change velocity or steer away from the obstacles.

I. INTRODUCTION

In the recent years smartphone and tablet sales indicate an exponential growth worldwide. In 2012 surveys estimated 44% of US citizens owning a smartphone or a tablet. Compared to 2011 data, annual growth was 13% or over 41 million new users just in United States of America [1]. On the other hand, automation in home applications shows a steady linear growth but has still limited availability to general public [2]. These two branches of industry share a common bright future and can successfully cooperate to yield new products and applications.

Connection between the two can be established using wired or wireless technology. Wired connection has its advantages and special usages but limits the portability of the system itself. Since smart mobile devices evolved from regular cell phones, most of them implement some wireless communication modules. This article is therefore focused on methods of wireless data transfer and possibilities derived from them.

Systems depending on communication between mobile and embedded devices have both industrial and commercial use. Current implementations can be found in car industry, mobile payment systems, digital watches and home automation systems [3].

The goal of this project is to design and build a fully functional product based on interaction of an embedded system and a smart mobile device. In this case implementation will be a remotely controlled boat. Platforms used in this project will be an Arduino Uno on the embedded system side and an Android application on the side of a smart mobile device. Intent behind that particular selection is to prove that it is possible and inexpensive to design such a system. Both of platforms are in wide use and are available to the general public. Also, both platforms have open source code and can be easily modified to a particular use.

II. SYSTEM LAYERS

A. Embedded system

An embedded system is a computer system designed to control specific functions within a larger system. It is embedded as part of a complete device, often including hardware and mechanical parts. Embedded systems contain processing cores that are either microcontrollers or digital signal processors dedicated to handle a particular task.

Most of the systems used in this project require some part of computing directly on the physical layer of the product. Therefore, when designing a system, engineer is presented with a choice between designing complicated digital circuitry or choosing an existing of-the-shelf microcontroller. Usually, the latter is better option because it yields better flexibility and can be easily modified [1]. A microcontroller is a small computer on a single integrated circuit containing a processor core, memory, and programmable input/output peripherals.

In this particular case microcontrollers serve as proxies and adapters from mobile devices to the electronic circuitry and electromechanical implementations of the product. They gather and redirect control words and data from both ends of the system and manipulate the data if it is needed. Reviewed microcontrollers have a way of connecting to a wireless module or have one built into the circuit board.

Specifically, Arduino UNO is selected for its versatility, availability, low cost and good documentation. Board is powered by the ATMega328 microprocessor complimented with 2KB SRAM and 32KB of programmable memory. Clock is set to 16MHz. The configuration suits very well to this particular use but can be limiting for more complex systems.

Connection to the functional peripherals is usually acquired through the direct link between the microcontroller and the peripheral device. But there are many ways of communication and control. Digital devices are commonly controlled using parallel input and output, I2C or SPI buses [2]. On the other hand, analogue devices are usually controlled using pulse-width modulation (PWM) or DA converter while AD converter and PWM can be used to gather data from analogue sources.

Most commonly used peripheral devices are keyboards, motors, indicators, displays and various sensors. The number of devices that can be simultaneously connected to the microcontroller depends on the microcontroller but is practically limited just by power consumption and complexity of wiring.

B. Smart mobile device

Mobile phones are embedded systems with a specific purpose of making phone calls and exchanging text messages on the go. Over the time, cell phones also became a platform with the ability to capture photos, play music and videos, connect to the Internet and run various third party applications [6]. This features required hardware and software upgrades and at one point a fully featured operating system. Mobile phones and tablets with such operating system are called smart mobile devices. What makes these devices unique and perfect for this project are portability, easy user interface and a myriad of different sensors and connection types built in one small package.

In this particular system, smartphones and tablets serve as a handle to a broad range of users while complimenting the system with new sensors and more processing power. Also, it is a good practice to remember that most of these devices can be connected to the Internet and receive data and control from the remote sources. All of the main operating systems have open software development kits (SDK) and it is fairly easy to start developing applications. SDK provides good support for using embedded sensors and communication methods. Applications can easily be transferred to a single mobile device or published on market specific to the chosen operating system.

The first thing user notices when starting an application is the user interface (UI). It is an important scope of the application and defines its quality. Modern mobile operating systems have separate files and packages for UI design. There are three ways to design the interface. The most complicated one is to dynamically create the content by directly programming it in the SDK programming language. Benefit of this way is flexibility but the time consumed by development process is significantly higher than in the two other ways. Second way is programming UI using markup language used by the SDK. It usually consists of layouts and widgets defined by the properties following the specified widget or layout tag. When compiling or building application, markup language is read out by the compiler and translated to the binary xml. At the end, it is joined by the rest of the program and the application properties in a single package. The last way is to use graphical user interface when designing the UI for the mobile application. This inceptional method is just an upgrade over the markup language style and therefore yields a markup language file that can also be edited as such.

Functions of the application are realized using programming language specified by the SDK. Most of the languages used in modern systems are based on the object oriented paradigm. It is also possible to write the program or some part of it in native assembly language using native development kit (NDK). The first option is used more often while NDK is used only if the application requires low level optimization.

Most of the contemporary smartphones and tablets have multiple sensors [7]. Commonly found are gravitational sensor, digital camera, digitizer for touch input, ambient light sensor, proximity sensor and gyroscope. Some additional data can also be retrieved from system properties. Such are processor temperature, battery voltage and wireless connection quality.

Data handled by the mobile device usually consists of sensor readings and controls to the embedded system. Both can temporarily be stored in internal application memory or outputted to a database or a file.

C. Communication

Communication between the embedded system and smart mobile device can be wired or wireless and is used to transport control words and various data. It can be one way or duplex depending on the requirements of the system. Packages used for communication can be of standardized type or programed by the developer. An important aspect of the whole system is to have a common interface so both sides can successfully exchange data.

This requires implementing a stream of data between the embedded system and mobile device. Communication between them can be synchronous or asynchronous. If the communication channel is asynchronous, data is usually received using listener thread. Otherwise synchronization protocol is required by the system. Outgoing data is transmitted through the output stream wirelessly or uploaded to a remote server.

III. POSSIBLE IMPLEMENTATIONS

A. Embedded systems implementation

Embedded systems consist of control logic and functional parts of the system (input devices, sensors, indicators...). Since the latter depend on the special functions of the system, the digital logic and microcontrollers will be in the focus of this paper.

Commonly used microcontrollers are based on Reduced Instruction Set Computing (RISC) or Complex Instruction Set Computing (CISC) architecture. Most popular general use microprocessors are designed by ARM. When there is a need for a lot of computing, x86 architecture can be used. The drawback of x86 architecture is usually high power consumption and complexity. Programs for the microcontrollers can be stored in the internal or external ROM or can be obtained from the network and temporarily stored in RAM. The first option is more common as embedded system programs don't get edited often. In case of using System-on-a-Chip (SoC) microprocessor with built-in memory programming is done by connecting the entire microcontroller to the PC and uploading the program. When using external memory, the entire memory chip can be disconnected from the microcontroller and the program can be uploaded separately. This is a preferred option if the microcontroller can't be removed from the rest of the electronics.

There are many different off-the-shelf development platforms ranging from basic 8-bit microprocessors on a plain circuit board to fully featured systems with multicore 32-bit processors with embedded connectivity modules, video and audio output and multiple buses. Some of the most popular development platforms are Atmel AVR, 8051, Freescale, Texas Instruments MSP430, PIC, Xilinx and various ARM platforms. Recently, open source devices like Arduino and Raspberry Pi are becoming a mainstream for 'Do It Yourself' types of the projects.

B. Mobile device implementation

Mobile devices can be categorized by the software, hardware, manufacturer, brand, price etc. Since this project requires a mobile application 'focus' will be on the mobile operating systems and SDKs.

There are just a few major mobile operating systems. One of the oldest were Palm OS and Windows CE but these are now obsolete. Soon came more popular Symbian and Blackberry OS but the real boom started with the introduction of the Apple iOS in 2007. Google introduced Android in 2008 and Microsoft released its Windows Phone operating system in 2010. These three operating systems hold 88.7% of the total market share in third quarter of 2012.

Apple iOS has restricted application ecosystem and closed source code. Also it must be programmed on an Apple personal computer in Xcode IDE. Programs are written in Objective-C or Objective-C++. All programs have to be licensed and tested by the Apple Inc. before they are allowed on the 'App Store'. The App Store has over 800 thousand applications and some of them interact with embedded system based devices [8]. Drawback of these applications is a high price and the fact that they are bound to equally expensive and unavailable hardware.

Android has an open ecosystem and source code. It is programmed in Java and XML in various development environments. Most popular are Eclipse, IDEA and NetBeans. IDE called Xamarin also allows developers to use of C# and .NET tools. 'Google Play' marketplace has over 700 thousand applications and 25 billion downloads (June 2012) [8]. But only a few applications are designed to be used with embedded systems and external hardware.

C. Types of communication

There are just a few types of communication between a smart mobile device and an embedded system. They can be categorized in two main groups. Wired transport usually relies on a serial transport using common bus in a form of a cable. Wireless transport relies on using public frequencies or GSM broadband connection. Most popular wireless implementations are Wi-Fi, Bluetooth, Near Field Communication (NFC), Radio Frequency Identification (RFID) and ZigBee. Table I. displays properties of these different technologies.

TABLE I. COMPARISON BETWEEN WIRELESS TECHNOLOGIES

	NFC	***RFID***	***Bluetooth***	***Wi-Fi***
Maximum Operating Range	10 cm	3 m	100 m	100 m
Operating Frequency	13.56 MHz	Varies	2.4 GHz	2.4/5 GHz
Directional Communication	Two way	One way	Two way	Two way
Bit Rate	106/212/ 424 Kbps	Varies	22 Mbps	144 Mbps

Wi-Fi is a very popular choice when connecting embedded systems and smart mobile devices since both can easily be modified to work together. Devices can be connected in three different ways: Point-to-Multipoint (Access Point), Point-to-Point (Ad hoc) and Multipoint-to-Multipoint (Mesh network). Most commonly used is Access Point (AP) method where all of the

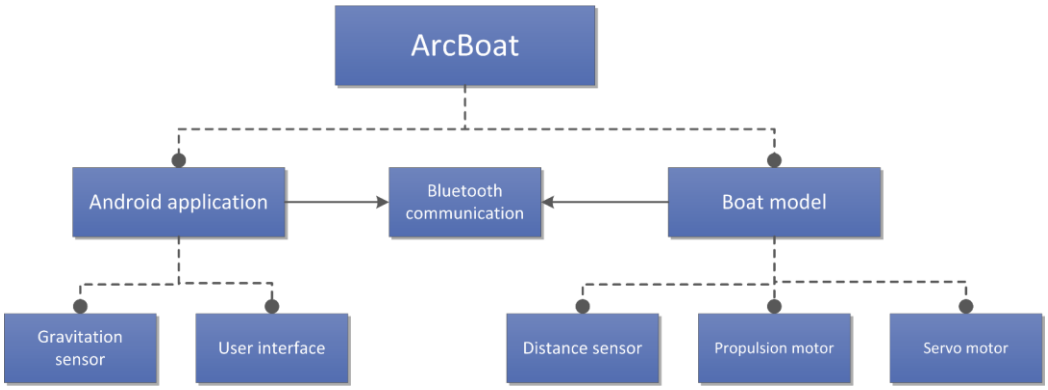


Figure I. ArcBoat distribution scheme

devices are connected to a single network node which manages all of the communication. A wireless ad hoc network is a decentralized type of wireless network. The network is ad hoc because it does not rely on a preexisting infrastructure, such as routers in wired networks or access points in managed (infrastructure) wireless networks. Instead, each node participates in routing by forwarding data for other nodes, and so the determination of which nodes forward data is made dynamically based on the network connectivity.

Bluetooth is a wireless technology standard for exchanging data over short distances. Bluetooth uses ISM frequency band (2.4-2.485GHz) and has 79 channels [10]. Devices are linked together into a Wireless Personal Area Network (WPAN) and can securely exchange the data. Most of the modern mobile devices have Bluetooth modules and it is easy to implement the communication. Bluetooth communication has a shorter range than Wi-Fi but also has lower power usage which is the main reason why Bluetooth communication is used in this project.

NFC and RFID are short range technologies based on lower frequency bands. Both have Point-to-Point architecture so only two devices can be connected at the same time. One device is an initiator and the other one is target. Data is transmitted half-duplex. The target device can be unpowered ('tag'). Main benefit of using NFC and RFID connection is that it operates only on short range which is convenient in mobile payment and identification systems.

IV. ARCBOAT IMPLEMENTATION

A. *Short description*

ArcBoat is an abbreviation of "Android remotely controlled boat" and is used as a symbolic name for the project. The goal is to design and build model of a boat remotely controlled with a mobile phone application. The boat gathers data from the sensors, receives heading and velocity parameters and controls the peripheral devices. Also it has automated collision avoidance system. Mobile application allows user to take control of model boats movement, read out the sensor data and adjust parameters of collision avoidance system. Used platforms and methods are available to both industry and general public. General scheme of implementation can be found in Figure I.

More information about the project, source code, circuit board design and wiring scheme can be found on the public GIT repository [3]. Project has open source and is licensed under Creative Commons license [4].

B. Used technology

Android has the biggest market share and there are not many applications that take advantage of embedded systems. Therefore it is selected as a development platform for this project. Application is programmed in Eclipse IDE with official Android SDK. Smartphone used for testing is HTC Desire running Android 4.2 Jelly Bean.

Arduino is one of the most popular cheap embedded systems. One of main reasons why it is selected for this project is availability of different 'shields' and modules. Program is written in C++ using Arduino framework and some external libraries. The specific Arduino variant used in this project is Arduino Uno. It has ATmega328 microprocessor, 32KB of flash memory and 2KB of RAM. The device is powered by a single 9V battery.

Selected communication method is Bluetooth 2.0. It provides excellent compatibility and low power consumption but has limited reach and bandwidth. 3G and 4G mobile connection are suggested as an improvement and possible upgrade to the project. Security and pairing are achieved using PIN authentication. Bluetooth shield is mounted on the Arduino board and interfaces through standard serial port using just two I/O pins (RX and TX) at baud rate of 38400.

C. Mobile application

Android application is the main interface between the system and the user. It has only one activity (screen) and it is laid out to be simple and practical. See Figure II. for the reference. Connection state can be found at the top of the screen. Beneath are readings received from the embedded system. At the bottom of the screen are two bars used for graphical interpretation of data from gravitational sensor and the calibration button. Connection can be established or terminated using

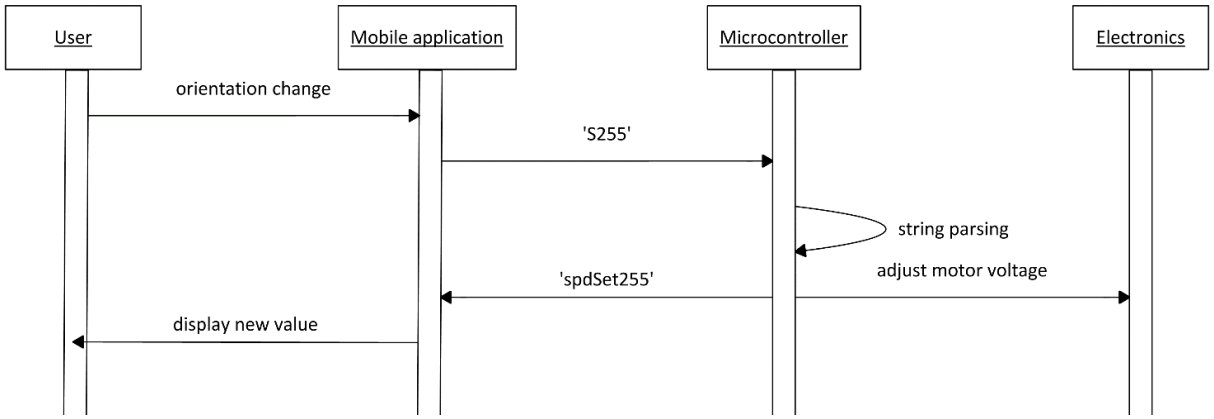


Figure III. Sequential diagram for speed adjustment

menu button or action bar. Collision avoidance parameters can also be accessed and changed from the menu or the action bar.

The application requires Bluetooth and Bluetooth_admin system permissions. When 'Connect' option is selected, application looks for the name of the device in the list of paired devices. If the designated device is found, two streams are created and initialized. The output stream manages the output queue and controls the transmission. Input stream manages the incoming data and is used inside of the listener thread that accepts the incoming data. Data from gravitational sensor is used for controlling the velocity and the heading of the model boat. 'Calibrate' option stores the current orientation data. Future readings are compared to those values and adjusted to fit the interface.

D. Embedded system

Embedded system is based on Arduino Uno and has Bluetooth HC05 shield attached to it. Shield communicates with Arduino through the serial COM port. The whole device is peered with the mobile application and receives control data and transmits the readings from the model. Besides, it has control over distance sensor, servo and propulsion motor.

Ultrasonic sensor HC-SR04 is connected using two digital pins. If the trigger pin has high input level for more than 10μs the sensor transmits 8 cycle ultrasonic burst and listens to the echo TTL signal. Built-in timer is used to control the time sequences and measure the round trip time. Distance is calculated as in

$$d = \frac{T_{high} * V_{sound}}{2} \quad (1)$$

Servo is controlled using an internal library. It is interfaced by transmitting 8 bit value representing the direction. Propulsion motor speed is regulated with the PWM signal.

Software interrupt is called each time a character arrives on the serial port of the Arduino. If character is the carriage return, the whole string is passed to the parser which determinates which command has been received. Data containing information on velocity and heading is transmitted to the mobile device periodically. The same timer is used for timing period for the data output and period at which ultrasonic sensor reads the distance.

E. Electronics

Inside of the boat hides a printed circuit board (PCB) used for voltage regulation of the propulsion motor. Since the highest TTL voltage of 5V is not high enough to power the motor, additional high voltage circuit is added. PWM signal is connected to the base of the TIP 122 Darlington transistor which translates 5V signal into the 18V signal. This voltage can successfully power the propulsion motor.

F. Communication

Communication is based on Bluetooth protocol. Link is managed by the HC05 Bluetooth integrated circuit. Properties such as baud rate, visibility, name of the device and access PIN can be changed using AT commands with reversed RX and TX pins on the shield.

Parameters exchanged through the Bluetooth link are speed, direction and distance to the obstacle. Each of the parameters is sent as ASCII string. Bluetooth shield receives this data and passes it to the microcontroller through the standard COM port using 'Universal asynchronous receiver/transmitter' (UART). Baud rate is set to 38400 bytes per second. Bluetooth control word has 8 data bits, and one stop bit. There is no parity control.

When received, the string is parsed to determine which parameter is received and what its value is. For example, parameter for changing the speed of the boat looks like "s<???>" where "s" symbolizes the speed and value is inserted instead of "<???>". Example of the successful communication is displayed in Figure III.

V. CONCLUSION

A lot of the current and future embedded systems can be customized and adapted to be controlled or monitored from the smartphones and tablets. This allows general users an easy and well-structured interaction with the embedded devices in home and industry. System consists of embedded device program and mobile application. Connection can be realized by direct wired connection or wirelessly. The latter is preferred and recommended.

ArcBoat is currently only a prototype of the system and a successful proof of the concept. With added camera and other sensors it can be used to explore inaccessible sea caverns and for marina surveillance. After adding GSM and GPS module and with improved construction it can be used as an automated tool for GSM signal measuring at open seas. This data can be used by the telecommunications companies for improvement of the signal coverage and discovering possible gaps or anomalies.

Some future improvements can be made to the functions and performance of the ArcBoat. GSM or Wi-Fi can be used if long distance control is required. Also, it is preferred to communicate using the stream of bits and bytes instead of streaming characters. With a few upgrades, the whole system can become modular so it can be used with different set of sensors, motors or communication methods. Small camera and multiple sensors can be added to the model for remote management. Application can be modified to display more data and information such as connection strength or number of lost packages. Application can easily be adapted to work with other mobile platforms and even personal computers.

VI. REFERENCES

[1] O. C. Maquelin, H. H. J. Hum and G. R. Gao, "Costs and benefits of multithreading with off-the-shelf RISC processors," in *EURO-PAR '95 Parallel Processing*, Springer Berlin Heidelberg, 1995, pp. 117-128.

[2] S. Ribarić, Građa računala, Zagreb: Algebra d.o.o., 2011.

[3] B. Vunderl, "Bitbucket repository," 2013. [Online]. Available: https://bitbucket.org/Xenosb/arcboat.

[4] Creative Commons, "Attribution-NonCommercial-ShareAlike 3.0," [Online]. Available: http://creativecommons.org/licenses/by-nc-sa/3.0/. [Accessed 2013].

[5] ADT, "Infographic List," 2011. [Online]. Available: http://infographiclist.com/2011/10/04/insights-and-stats-about-home-automation-systems-infographic/.

[6] R. Budihal, "EE Times," 2010. [Online]. Available: http://www.eetimes.com/discussion/other/4204667/Emerging-trends-in-embedded-systems-and-applications.

[7] "Mobile Statistics," 2012. [Online]. Available: http://www.mobilestatistics.com/.

[8] R. Triggs, "Apple App Store vs Google Play: year-end trends and statistics," 21 December 2012. [Online]. Available: http://www.androidauthority.com/apple-app-store-vs-google-play-distimo-141851/.

[9] Bluetooth SIG, "A Look at the Basics of Bluetooth Wireless Technology," 2013. [Online]. Available: http://www.bluetooth.com/Pages/Basics.aspx. [Accessed 16 February 2013].

[10] S. Faccin, C. Wijting, J. Kenckt and A. Damle, "Mesh WLAN networks: concept and system design," vol. 13, no. 2, pp. 10-17, 2006.

[11] A. Charlesworth, "The ascent of smartphone," *Engineering & Technology*, vol. 4, no. 3, 2009.

Optimal Control of a Family House Heating System

Mario Vašak *, Anita Martinčević**
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, Zagreb, Croatia
Email: mario.vasak@fer.hr*, anita.martincevic@fer.hr**

***Abstract*—Approximately 42% of total energy usage today is spent on the heating of the human environment. Energy consumption can be significantly reduced by proper design of the heating control system - performed studies suggest that in average about 12% reduction of energy usage can be achieved based on control system improvement. This paper describes an efficient control of a family house heating system in its very common configuration. We use model predictive control to fully take into account the prediction of disturbances (room temperatures), reference values of power per rooms and physical limitations of actuators. This will result in an optimal flow rate and temperature of medium required to meet all the demands. In this paper we will first present the model of a heating system, and from it derive model predictive control and verify it through simulation.**

I. Introduction

Each day we feel a lack of energy, therefore energy saving is a general social task. Correct automatic control of temperature and mass flow of the heating medium is the most effective way to save energy in heating systems. The heating system is designed for single-story house with two rooms and the attic [1]. For such a model in [2] model predictive control strategy has been derived. As the external conditions (temperature, insolation, etc.) can for the future be determined only in certain limits of uncertainty, the stochastic approach to the problem is chosen. The stochastic approach guarantees the meeting of all limitations only inside an interval of certainty. The stochastic MPC (SMPC) controller commands the heating power input for the rooms in the house so that the outdoor conditions are fully exploited in keeping the room temperatures within the comfort limits, and therefore minimum additional energy for heating is spent. In the mentioned work the nature and performance of the heating system was neglected and it was assumed that the heating input can change instantaneously. However, this assumption is not valid for real systems. The obtained required power input for heating per rooms is constant within one hour and will be a reference that must be achieved by optimal control of the heating system.

The goal of optimal control of the heating system is the minimization of energy consumption required to heat-up the heating medium and operate the pump, and also minimization of the deviation of heating powers brought into the room from the required values necessary to satisfy the comfort limits. Model predictive control fully takes into account the prediction of disturbances (room temperatures), reference values of powers per rooms and physical limitations of actuators. The model of the central heating system is highly nonlinear, and it must be iteratively linearized at each optimization step in order to implement the model predictive control. The result will be the optimum flow rate and temperature of the heating medium which will satisfy all demands and minimize the power consumption.

Section II describes the model of the heating system, section III deals with mathematical formulation of the optimization problem, and in section IV we verify the control strategy through simulations.

II. Modeling of the Heating System

The heating system consists of two radiators, one for each room, piping, a variable speed circulation pump and a boiler with a thermostat. The heating system is designed for a single-story house with two rooms and the attic, but only the ground floor is heated. A diagram of the heating system is shown in Figure 1.

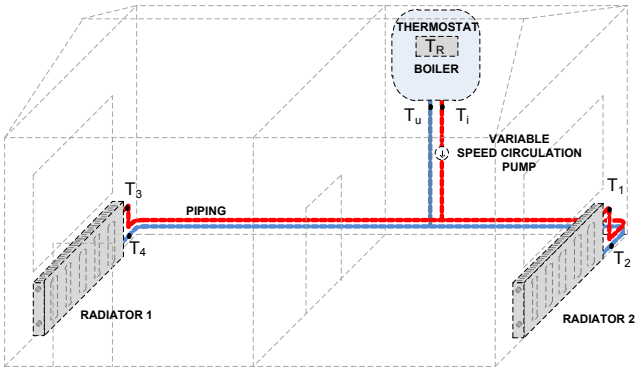


Fig. 1. Diagram of the heating system

A. Radiator

A radiator is part of a heating system which is used to transfer heat, mostly from hot water to the air. The working principle is based on the passage of a heating medium heated in a boiler through the radiator where the heat transmission medium heats up the radiator and loses heat. Radiators used to heat the ground floor are aluminum radiators composed of 25 segments. The number of segments was determined according to the recommendations of the radiator manufacturers. By installing a larger number of segments the initial costs are higher, but lower water temperature can be used for the same

heat effect, thus the energy is saved and the regulation of room temperatures is faster.

The mathematical model is based on the assumption that the radiation flux of the aluminum radiator is negligible in relation to the convection flux. A mathematical description of the heat transfer inside radiator is obtained assuming that the water temperature in the radiator does not depend on spatial coordinates and is equal to the average temperature of the water:

$$m_v c_v \frac{dT_{iz}}{dt} = c_v \frac{Q}{2}(T_{ul} - T_{iz}) - \alpha_r A_r \left(\frac{T_{ul} + T_{iz}}{2} - T_z \right), \quad (1)$$

where T_{ul} is the temperature of the heating medium at the entrance of the radiator, T_{iz} is the temperature of the heating medium at the exit from the radiator, Q is the mass flow of the heating medium from the boiler, c_v is the specific heat capacity of the heating medium, α_r is the overall heat transfer coefficient form heating medium to air, A_r is the active surface area from which heat is transferred to the air, m_v is the mass of the heating medium in the radiator and T_z is the temperature of air in the room.

From the Figure 1 it is evident that the modeled heating system is a closed system so mass flow Q is the same throughout the entire system, whereas it is assumed that both branches of pipes have the same resistance, such that each radiator receives half of the heating medium mass flow.

B. Piping

Piping is a part of the heating system used to transfer the heat from the heat source to the heat exchangers using a suitable heating medium. The pipeline is modeled using four segments, a hot and a cold pipe for each room separately. The heat conservation equation for one segment is:

$$Q c_v T_{ulc} = Q c_v T_{izc} + 2\alpha_c A_c \left(\frac{T_{ulc} + T_{izc}}{2} - T_z \right), \quad (2)$$

where T_{ulc} and T_{ilc} are temperatures of the heating medium at the entrance and exit of the segment, α_c is the overall heat transfer coefficient form heating medium to air through a copper transmission surface and A_c is the active surface area from which heat is transferred to air.

C. Boiler

Hot water, i.e. the heating medium is produced in the boiler. Assuming that the temperature disposition in the boiler is homogeneous, the power of the boiler is:

$$\eta P_b = c_v Q (T_i - T_u), \quad (3)$$

where η is the usability factor, T_u is the temperature of the water returning from the system to the boiler, and T_i is the temperature of water at the exit from the boiler. To avoid nonlinearity which occurs in real systems when modeling a thermostat, it's assumed that the temperature at the exit from the boiler has PT_1 behavior in relation to reference temperature T_R,

$$T_b \frac{dT_i}{dt} + T_i = T_R, \quad (4)$$

where T_b is the time constant of the boiler. If we assume that the boiler is gas powered, the consumption of gas can be minimized directly by minimizing the power of the boiler.

D. Variable speed circulation pump

A circulation pump is a device which is used to transfer fluid (i.e. heating medium) from a lower to a higher level, i.e. from lower to higher pressure. The advantage of using the pump in heating systems is a faster circulation of the hot medium through the system and thus heating the space faster and achieving the desired comfort faster. The power of the pump is proportional to the flow to the third power:

$$P_p \sim Q^3. \quad (5)$$

In order to use quadratic solvers, at each step of the optimization the cubic power characteristic of the pump is approximated by a square characteristic around the predicted optimal mass flow along prediction horizon calculated in the previous step. As the mass flow of the heating medium is regulated through the optimal control of the variable speed circulation pump, the thermostat radiator valves are not necessary.

E. Overall model of the heating system

All the described elements of the heating system together form a closed system of a central heating. Inputs into the system are defined by the flow pump Q and a reference temperature of the water in the boiler T_R. Air temperatures of the east T_{ze} and west T_{zw} room are disturbances which act on the heating process. Temperatures at the exit from the radiators (T_2, T_4) of eastern and western room and output boiler temperature T_i are system states. Outputs of the system are the returning water temperature T_u and the power transmitted into the eastern P_e and the western P_w room. The described model of the central heating system is a highly nonlinear system and for the implementation of model predictive controller it must be iteratively linearized around the predicted optimal values calculated in the previous optimization step and the familiar amounts of disturbance along the prediction horizon. Nonlinear terms which describe the heating system can be found in [3]. Linearized system can be expressed in state space:

$$\begin{bmatrix} \Delta\dot{T}_i \\ \Delta\dot{T}_2 \\ \Delta\dot{T}_4 \end{bmatrix} = \begin{bmatrix} -1/T_b & 0 & 0 \\ K_{Ti} & K_T & 0 \\ K_{Ti} & 0 & K_T \end{bmatrix} \begin{bmatrix} \Delta T_i \\ \Delta T_2 \\ \Delta T_4 \end{bmatrix} + \begin{bmatrix} 1/T_b & 0 \\ 0 & K_{Qe} \\ 0 & K_{Qw} \end{bmatrix} \begin{bmatrix} \Delta T_R \\ \Delta Q \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ K_{Tz} & 0 \\ 0 & K_{Tz} \end{bmatrix} \begin{bmatrix} \Delta T_{ze} \\ \Delta T_{zw} \end{bmatrix}. \quad (6)$$

Linearized output equations are:

$$\begin{bmatrix} \Delta T_u \\ \Delta P_e \\ \Delta P_w \end{bmatrix} = \begin{bmatrix} 0 & \Theta_T & \Theta_T \\ \Theta_{P,Ti} & \Theta_{P,T} & 0 \\ \Theta_{P,Ti} & 0 & \Theta_{P,T} \end{bmatrix} \begin{bmatrix} \Delta T_i \\ \Delta T_2 \\ \Delta T_4 \end{bmatrix} + \begin{bmatrix} 0 & \Theta_Q \\ 0 & \Theta_{P,Qe} \\ 0 & \Theta_{P,Qw} \end{bmatrix} \begin{bmatrix} \Delta T_R \\ \Delta Q \end{bmatrix} + \begin{bmatrix} \Theta_{Tz} & \Theta_{Tz} \\ \Theta_{P,Tz} & 0 \\ 0 & \Theta_{P,Tz} \end{bmatrix} \begin{bmatrix} \Delta T_{ze} \\ \Delta T_{zw} \end{bmatrix}. \quad (7)$$

The dependence of the system matrix coefficients on the stationary system state values can be found in [3]. Equations (6) and (7) fully describe the linearized heating system. The linearized system describes the given nonlinear system well only around a working point. Due to the high nonlinearity of the system in relation to the input mass flow, mass flow changes within the sampling time are limited to 20% of the stationary value.

III. Optimization

Use of automatic control in thermal processes can lead to significant energy savings. The heating system contains two actuators, which are managed with the purpose of reducing energy consumption. In [2] stochastic model predictive control is implemented on the same house model. The controller commands one hour constant heating power for rooms in the house so that the outdoor conditions are fully exploited in keeping the room temperatures within the comfort limits. The required heating power was obtained by neglecting the heating system performance and assuming that power of the actuator can be changed instantaneously. This assumption is not valid for real systems. So obtained required power will be a reference that must be achieved by optimal control of the heating system. The goal of optimal control of the heating system is the minimization of energy consumption required to heat up the medium and operate the pump, and the minimization of the deviation of power brought into the room from the required reference values necessary to satisfy the comfort limit.

$$\mathcal{J} = \min_{\Delta T_R, \Delta Q} \Big[\int P_b + \int P_p + R \left(\int P_{ref,e} - P_e \right)^2 + R \left(\int P_{ref,w} - P_w \right)^2 \Big], \quad (8)$$

where R is the weighting matrix, through which the weight of minimization of integral square deviation is specified. The weighting matrix determines how important it is to satisfy the requirements for minimal quadratic deviation from the referent value of power in relation to the requirement for minimizing the power of the boiler and the pump. Model predictive control fully takes into account the prediction of disturbances (room temperatures), reference values of power per rooms and physical limitations of actuators. Stochastic model predictive control from [2], that produces references for heating power inputs for the rooms, takes into account the outside temperature influence as well as the influence of solar irradiance on the house outer envelope along the prediction horizon. Additionally on this level also the heat load from occupancy and other activities in the rooms could be taken into account.

Integrals of boiler and pump power and the exerted power error in each room are obtained by expressing the above powers dependence as a function of inputs, disturbances and initial state. Dependence of mentioned powers of disturbances and system inputs can be described using PT_1 and PT_2 transfer functions. For PT_1 and PT_2 transfer functions there are analytical expressions for the step response, with which the given power integrals can easily be solved. The exact expressions for integral of the boiler and the pump power, and the power submitted in each room can be found in [2].

By solving the integrals mentioned in the optimization criterion, the quadratic optimization criterion problem is obtained. The optimization criterion in form of a quadratic problem is:

$$\mathcal{J} = \min_{\mathbf{X}_k} \frac{1}{2} \mathbf{X}_k' \mathbf{H}_k \mathbf{X}_k + \mathbf{f}_k' \mathbf{X}_k, \quad (9)$$

where vector $\mathbf{X}_k$ contains the optimal input values along the prediction horizon:

$$\mathbf{X}_k = \begin{bmatrix} \Delta T_{R,k|k} \\ \Delta Q_{k|k} \\ \Delta T_{R,k+1|k} \\ \Delta Q_{k+1|k} \\ \vdots \\ \Delta T_{R,k+N-1|k} \\ \Delta Q_{k+N-1|k} \end{bmatrix}. \quad (10)$$

$T_{R,k+1|k}$ presents the predicted optimal value of the reference boiler temperature at time $(k+1)$ obtained on the basis of the data available at the time k (disturbance profile and reference power values along the prediction horizon). Matrices $\mathbf{H}_k$ and $\mathbf{f}_k$ are matrices which define the quadratic problem, where the $\mathbf{H}_k \in R^{2N \times 2N}$ and $\mathbf{f}_k \in R^{2N}$. Elements of these matrices were obtained by the linearization of the heating system around the predicted optimal input values along prediction horizon from the previous step and the familiar amounts of disturbance along the prediction horizon. Linearization is therefore performed iteratively at each step. Algorithm of iterative linearization can be described using the following steps:

- System initialization : $k = 1$, $\mathbf{U}_0 \in R^{2N}$;

1. Linearization of nonlinear system around the predicted optimal input values along the prediction horizon from the previous step $\mathbf{U}_{k-1}$ and the familiar amounts of disturbance along the prediction horizon $T_{D,k}$,

$$\mathbf{U}_{k-1} = \begin{bmatrix} T_{R,k-1|k-1} \\ Q_{k-1|k-1} \\ \vdots \\ T_{R,k+N-2|k-1} \\ Q_{k+N-2|k-1} \end{bmatrix}, \mathbf{T}_{D,k} = \begin{bmatrix} T_{d,k} \\ T_{d,k+1} \\ \vdots \\ T_{d,k+N-1} \end{bmatrix}, \quad (11)$$

   where $T_{d,k} = \begin{bmatrix} T_{ze,k} & T_{zw,k} \end{bmatrix}$;
2. Definition of the quadratic optimization problem in form (9) - determination of matrices $\mathbf{H}_k$ and $\mathbf{f}_k$;

3. Obtaining the vector $\mathbf{X}_k$ by solving a finite horizon quadratic optimal control problem using a developed code in Matlab;
4. Obtaining the optimal input values $\mathbf{U}_k$=$\mathbf{U}_{k-1} + \mathbf{X}_k$;
5. Applying the $T_{R,k|k}$ and $Q_{k|k}$ to the system;
6. Time update $k = k + 1$.
7. Back to step 1;

The exact expressions are given in [3].

Mass flow given by the circulation pump is restricted to the interval $0.01 - 1\ kg/s$, referent temperature of the boiler is restricted to the interval $0 - 100\ ^oC$ and boiler is 7 kW sized. Additional restrictions are unnecessary because they follow directly from the given physical constraints of the system.

This optimization problem is solved by using Matlab Multi-parametric Toolbox [4] routines and appropriate quadratic programming solver like CPLEX [5].

IV. Results

The accuracy of the realized MPC regulator is tested by realizing a hierarchical heating process control system on two levels. On the higher level, using a SMPC regulator realized in [2], the optimal power input values per room necessary to maintain the desired temperature profile in the rooms are determined every hour through optimization. Optimal control of the heating system on a lower level gives the optimal amount of the mass flow and the necessary temperature of the heating medium, four times an hour, in order to spend as less energy as possible to get as little as possible deviation from the required reference power values generated by the higher control level. The central heating system generates the required power into the rooms by generating an optimal mass flow of the heating medium using a pump and heating the water in the boiler to a required optimal value.

The simulation was conducted during the first half of the year when the outside temperature values require the heating of interiors. The prediction horizon of both control levels is $N = 8$. The room temperatures must be kept within a limits of $20 - 25$ degrees with a probability of $95\ \%$. In line with the expectations and the given limitations, MPC on lower level calculates the optimal mass flow (Figure 5) and the referent boiler temperature (Figure 4) to satisfy the minimum deviation of the power transferred to the rooms from the demanded value generated by higher level, and at the same time satisfy the request for minimum boiler and pump power.

In Figures 4 and 5 the optimal input values into the heating system are shown. The heating system response to the calculated optimal inputs is shown in Figure 2 and the temperature response of western room to the input power generated by heating system is shown in Figure 3. Figure 6 shows an enlarged view of following the reference power during one day. It is obvious from Figure 6 that the ideal following the reference in the case of real actuators cannot be achieved. Optimal values of inputs to the heating system along the prediction horizon are found through optimization which minimize integrated squared difference between the actual and the reference value, taking into account that the user wants to

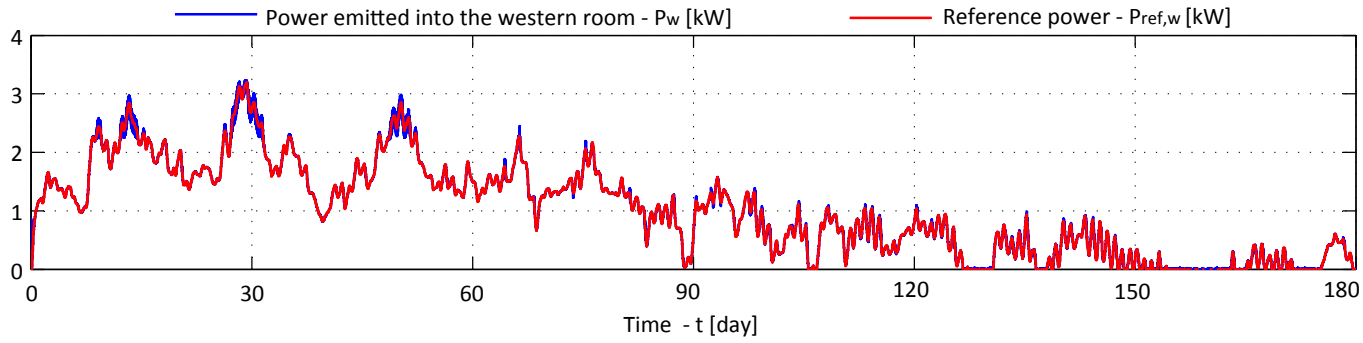


Fig. 2. Response of the heating system to the optimal inputs.

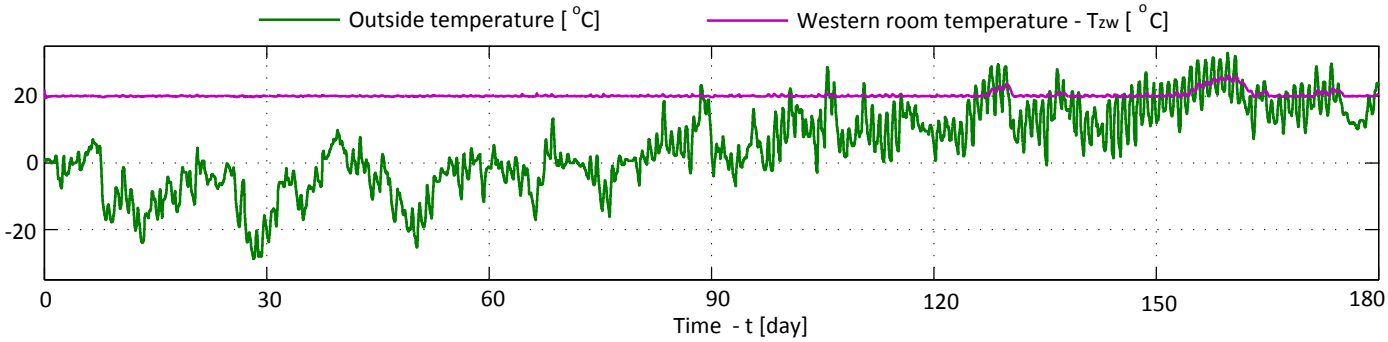


Fig. 3. Temperature response of the western room.

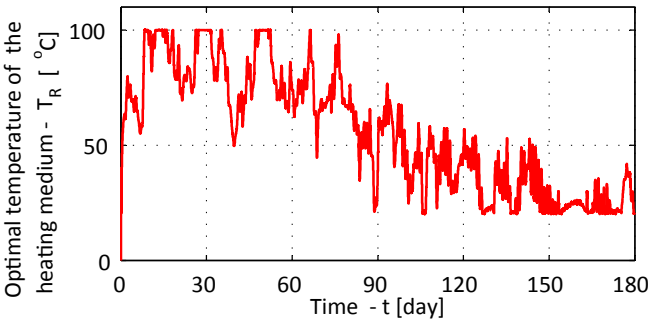


Fig. 4. Optimal temperature of the heating medium.

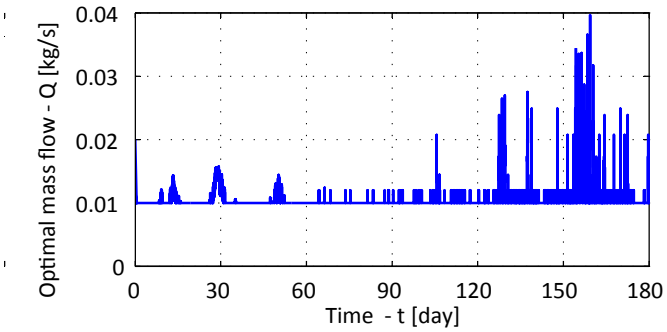


Fig. 5. Optimal mass flow.

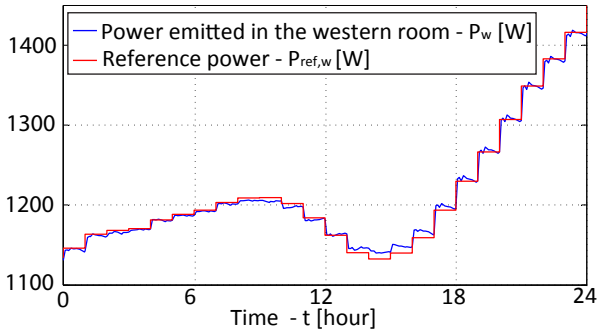


Fig. 6. Optimal temperature of the heating medium.

spend as little as possible for the required comfort (constant power within one hour ensures room temperatures maintained in the comfort interval), which means that the actuator power is as small as possible. The power submitted to the room can be increased by increasing the mass flow or by increasing the referent boiler temperature. The same applies to the demand for decreasing power submitted to the room.

From Figure 5 it can be seen that the temperature increase of the heating medium is an energetically cheaper solution to achieve the required power. For the previously stated reasons, with the temperatures of the heating medium inside the interval of $0 - 100\ ^oC$ the regulator keeps the mass flow on its minimal value (if there are no large leaps in the referent power requirements). With higher referent power values and the maximum temperature of the heating medium, the required referent power is gained by increasing the mass flow. With large leaps of the referent power value when it's necessary to achieve a large increase in the power submitted to the room in the minimum period of time (minimum deviation between the referent and the achieved power value), the necessary power value cannot be achieved fast enough only by heating up the heating medium, but it is necessary to increase the mass flow.

Low referent power values $(0 - 0.5kW)$ can be accomplished with lower temperatures of the heating medium. A higher level, aimed at minimizing the consumption of the actuators, generates low referent power values in case when the outdoor temperature is close to the user's defined comfort interval. The most efficient way of generating such power values is with the temperature of the heating medium close to the given comfort interval and the increase in mass flow (Figure 5 150th-180th day). There is no great decrease in the heating medium temperature in passing through the central heating system due to indoor temperatures within the given comfort interval, outdoor temperature close to the given interval and a high mass flow, so the power used by the boiler is minimal.

In spite of the stochastic behavior of the disturbance that affects the process [2] and the realization of the heating system which cannot follow the referent values ideally, we have managed to achieve a satisfactory behavior of the system due to the fact that the indoor temperature is within the interval set by the user (Figure 3).

Depending on the desired optimization effect, different reactions of the systems are accomplished by assigning different weights (unit energy costs) to the pump power and boiler power and changing the weighting coefficient R. In the conducted simulations where R=10, the emphasis was on minimizing the integral square deviations of powers.

V. Conclusion

In this paper we have modeled a heating system which consists of a variable speed circulation pump, a pipeline, a boiler and two radiators. The heating system is modeled for a house that has already been used to construct SMPC. The SMPC provides optimal input power demands which maintain indoor temperatures within the comfort interval set by the user. SMPC takes into account system dynamics, initial conditions, limitations and prediction of the disturbances along the prediction horizon and their uncertainty leaving enough space for preventing the disturbances in the following step from putting the system out of its limitations. Demanded optimal input power ensures maximal exploitation of natural resources (insolation, outside temperature, etc.) that we have at a certain time of the day. The problem is that the demanded optimal input power is constant within one hour, and obtained under the assumption of instantaneous changes of input power. Constant committed power and sudden steps are not achievable in a real system. In order to bring the optimal control of the temperature process as close as possible to the real system, hierarchical control has been realized. MPC of the heating system is used in order

to achieve minimum possible deviation form the required amount of exerted heating power, and at the same time to minimize the energy consumed by boiler and circulation pump. Hierarchical control ensures a minimal deviation from the required referent power values with minimal energy expenditure and maintaining the temperature within the interval set by the user. Despite the high sensitivity of the central heating model to changes in input flow, because of high nonlinearity, the algorithm synthesized over the heating system ensures minimal deviation between the actual and the reference power at minimum cost of the consumed energy and room temperatures in the desired range.

ACKNOWLEDGMENT

This work has been financially supported by the European Union through project ENHEMS-Buildings – Enhancement of Research, Development and Technology Transfer Capacities in Energy Management Systems for Buildings, grant No. IPA2007/HR/16IPO/001-040510.

REFERENCES

[1] A. Starčić, *Modeliranje i upravljanje toplinskim procesom u zgradi*, Faculty of Electrical Engineering and Computing - Zagreb, 2010.

[2] A. Martinčević, *Modelsko prediktivno upravljanje toplinskim procesom u zgradi uz uzimanje u obzir nesigurnosti podataka iz vremenske prognoze* seminar - diplomski studij, Zagreb, 2012, Fakultet elektrotehnike i računarstva, 31 str. Voditelj: Vašak, Mario

[3] A. Martinčević, *Optimalno upravljanje sustavom za grijanje*, seminar - diplomski studij, Zagreb, 2013, Fakultet elektrotehnike i računarstva, Voditelj: Vašak, Mario

[4] M. Kvasnica, P. Grieder, M. Baotić, and M.Morari, *Multi-Parametric Toolbox (MPT).* 2003. http://control.ee.ethz.ch/ hybrid/mpt/.

[5] ILOG, Inc., *CPLEX User Manual.* http: www.ilog.com.

[6] A. Martinčević, *Modelsko prediktivno upravljanje toplinskim procesom u zgradi uz uzimanje u obzir nesigurnosti podataka iz vremenske prognoze* završni rad - preddiplomski studij, Zagreb, 2011, Fakultet elektrotehnike i računarstva, 33 str. Voditelj: Vašak, Mario

[7] Grundfos research and tehnology, *The Centrifugal Pump*, http://dk.grundfos.com, 21. siječnja 2013.

[8] F. Oldewurtel, A. Parisio, C. N. Jones, M. Morari et al., *Energy Efficient Building Climate Control using Stochastic Model Predictive Control and Weather Predictions*, 2010 American Control Conference, pp. 5100-5105.

[9] L. Wang, *Control System Design and Implementation Using MATLAB*, Springer-Verlag London Limited, 2009

[10] P. Kaluderčić, *Analiza procesa grijanja zgrade elektrotehničkog fakulteta*,INFOTEH-JAHORINA, Vol.4, Ref. A-11, str.46-52, 2005

Model Predictive Control of a Heating, Ventilation and Air Conditioning System

Mario Vašak*, Antonio Starčić **
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, Zagreb, Croatia
Email: mario.vasak@fer.hr*, antonio.starcic@fer.hr**

Abstract - The largest energy consuming subsystem in a building is heating, ventilation and air conditioning (HVAC) system. It is thus essential to optimize the behavior of that subsystem in order to achieve energy savings. In this work focus is on efficient energy usage in a sample zone system consisting of two coupled zones. Ventilation, heating and cooling are done by air handling unit (AHU) and variable air volume (VAV) boxes. Optimization of the system means finding optimal control sequences for AHU and VAV boxes. The model of the coupled zones is linear, while the models of the AHU and VAV boxes are nonlinear. In order to find a suboptimal solution, which will keep zone temperature in given boundaries, while handling ventilation requirements, linear programming and linear zone system model are used to find appropriate cooling/heating powers, and then those powers are used to obtain control sequence for AHU and VAV boxes.

I. INTRODUCTION

There is a trend of reducing energy consumption in buildings today. One reason for that is economic, reducing energy expenses, and the other is environmental, since reducing energy consumption reduces CO_2 emissions. Building energy dissipation can be reduced in various ways which can be more or less invasive. Example of invasive solution is additional insulation of the house, while noninvasive is, for example, better control of heating and cooling. The cooling and heating system in this example consists of centralized AHU and localized VAV box. By implementing model predictive control (MPC) for this system significant energy savings can be made. MPC has significant advantages over traditional control like PID, since it utilizes information about future environment temperature and solar irradiation to optimize energy consumption.

In this paper, MPC control is implemented on a sample configuration of two coupled zones in which cooling and heating of zones is done by AHU and VAV boxes. The idea is to use state space model of the coupled zone system for synthesis of MPC that resides on linear programming. The MPC controller outputs the optimized heating/cooling powers for different zones. Next step is determining control sequence for AHU and VAV boxes which will accomplish these heating/cooling powers in the optimal way. Another task of the AHU is to supply outside air into the zones, so that the quality of zone air is preserved. In order to accomplish that ASHRAE Standard 62.1 [2] is implemented.

The idea behind is to make the control system development extremely scalable, since optimized heating/cooling powers for zones can be computed for a large number of zones if simple model of zones with lumped heating/cooling power is used. Accomplishing these heating/cooling powers via AHU / VAV setup is then posed as a static optimization problem.

Brief description of the considered zone setup is given in Section II. Synthesis of MPC by linear programming is described in Section III. Ventilation goals are given in Section IV. Section V. deals with the problem of obtaining control sequence for AHU and VAV boxes. Results are analyzed in Section VI.

II. THERMAL MODEL OF THE CONSIDERED ZONE SETUP

The zone setup for which control scheme is done is shown in Figure 1. There are two zones, with floor area of 25 meters squared and height of 3 meters. The complete list of materials used in the model of the zones can be found in [1]. The model is reduced to only 8 states in order to solve linear program faster.

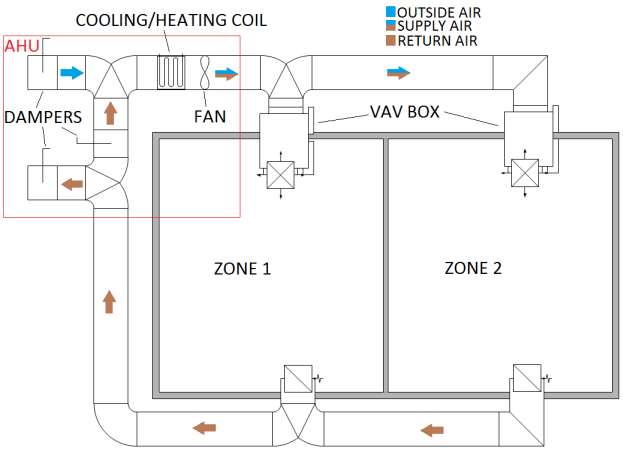


Figure 1 -Principle scheme of the zone setup.

III. SYNTHESIS OF MPC BASED ON LINEAR PROGRAMMING

Idea behind MPC is to find sequence of heating/cooling powers for zones which will keep zones temperatures in given boundaries on the prediction horizon, by taking into account current state of the model and all disturbances along the horizon. That way a sequence of input signals $U=(u_{t|t},u_{t+1|t},...,u_{t+N-1|t})$ is obtained. After that only the first input is applied, and the process is repeated in next step.

Model of the considered two-zone setup is continuous-time, so it must be discretized. Discretization is done by applying zero order hold with desired sampling time. The obtained discretized model can be described by:

$$x_{k+1}=A_d x_k+B_{du}u_k+B_{dd}d_k, \tag{1}$$

where A_d , B_{du} and B_{dd} are discrete state matrix, input matrix and disturbance matrix respectively. Vector x_k , is state vector and contains temperatures of walls and zones, u_k is input vector and contains heating/cooling powers for zones, d_k is disturbance vector and contains outside temperatures and solar irradiations. System response on the horizon can be described by:

$$X=\alpha\cdot x_{t|t}+\beta\cdot U+\gamma\cdot D, \tag{2}$$

where X is a stack of future states, U is a stack of future inputs and D is a stack of future disturbances:

$$\begin{aligned} X&=(x_{t+1|t},x_{t+2|t},...,x_{t+N|t});\\ U&=(u_{t|t},u_{t+1|t},...,u_{t+N-1|t});\\ D&=(d_{t|t},d_{t+1|t},...,d_{t+N-1|t}). \end{aligned} \tag{3}$$

Matrices α , β and γ are:

$$\alpha=\begin{bmatrix} A_d\\ A_d^2\\ \vdots\\ A_d^N \end{bmatrix}$$

$$\beta=\begin{bmatrix} B_{du} & 0 & 0 & ... & 0\\ A_dB_{du} & B_{du} & 0 & ... & 0\\ A_d^2B_{du} & AB_{du} & B_{du} & ... & 0\\ \vdots & \vdots & \vdots & \ddots & 0\\ A_d^{N-1}B_{du} & A_d^{N-2}B_{du} & ... & A_dB_{du} & B_{du} \end{bmatrix} \tag{4}$$

$$\gamma=\begin{bmatrix} B_{dd} & 0 & 0 & ... & 0\\ A_dB_{dd} & B_{dd} & 0 & ... & 0\\ A^2B_{dd} & A_dB_{dd} & B_{dd} & ... & 0\\ \vdots & \vdots & \vdots & \ddots & 0\\ A_d^{N-1}B_{dd} & A_d^{N-2}B_{dd} & ... & A_dB_{dd} & B_{dd} \end{bmatrix}$$

There are two limitations for this system. One is the requirement to keep temperatures of zones in given boundaries, and the other one limits heating and cooling power. These limitations can be described by:

$$\begin{aligned} \mathbb{X}^N&=\{X:\varepsilon_x X\le\xi_x\}\\ \mathbb{U}^N&=\{U:\varepsilon_u U\le\xi_u\}. \end{aligned} \tag{5}$$

Matrices ε_x , ε_u , ξ_x , ξ_u are defined according to limitations. Keeping zones temperatures in given boundaries means meeting these constraints:

$$\begin{aligned} T_{\min}\le T_1\le T_{\max}\\ T_{\min}\le T_2\le T_{\max}, \end{aligned} \tag{6}$$

where T_1 and T_2 are zone temperatures while $T_{\min}$ and $T_{\max}$ are lowest and highest zone temperature allowed. The heating and cooling power constraints are:

$$\begin{aligned} U_{\min}\le U_1\le U_{\max}\\ U_{\min}\le U_2\le U_{\max}, \end{aligned} \tag{7}$$

where U_1 and U_2 are power inputs into zone one and two, while $U_{\min}$ and $U_{\max}$ are minimum allowed power and maximum allowed power respectively.

The goal of MPC is to minimize input power while maintaining zone temperature:

$$\begin{aligned} &\min_u \|U\|_1=\min_u\sum_{k=0}^{N-1}|u_k|,\\ &\begin{pmatrix}\varepsilon_x\beta\\ \varepsilon_u\end{pmatrix}U\le\begin{pmatrix}\xi_x-\varepsilon_x\alpha x_{t|t}-\varepsilon_x\gamma D\\ \xi_u\end{pmatrix}. \end{aligned} \tag{8}$$

In order to define this as a linear program, absolute value nonlinearity must be resolved. That can be done by introducing slack variables Z , which meet the requirement $u_i\le z_i,-u_i\le z_i,i=1...N$, and minimize these variables. Now we can write our linear program like this:

$$\begin{aligned} &\min_{U,Z}\begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}U\\ Z\end{bmatrix}\\ &\begin{pmatrix}\varepsilon_x\beta\\ \varepsilon_u\end{pmatrix}U\le\begin{pmatrix}\xi_x-\varepsilon_x\alpha x_{t|t}-\varepsilon_x\gamma D\\ \xi_u\end{pmatrix}\\ &-Z\le U\le Z. \end{aligned} \tag{9}$$

This kind of problem can be solved using available linear programming solvers. Considered two-zone setup model response to optimal powers can be seen in Figure 2. The zone temperatures are kept between 20°C and 25°C. More on the topic of obtaining optimal cooling/heating powers can be found in [3].

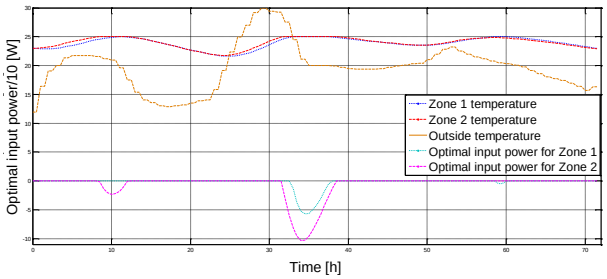


Figure 2 - Optimal input power response.

IV. RECIRCULATING VENTILATION SYSTEM AND ASHRAE STANDARD 62.1

With linear programming input cooling/heating powers which will be used in later control are obtained. Besides desired cooling/heating powers, ventilation criterions must be met. In order to preserve air quality in the zones ASHRAE Standard 62.1 [2] is applied.

The system that needs to be controlled consists of centralized AHU and localized VAV box. AHU's task is to mix return air and outside air, preheat or precool it. Mixing of the air is done by controlling air dampers. AHU also has supply fan, used to generate given airflow through supply duct. Every zone has its own VAV box, which is used to control air flow into the zone, and to heat up the air if necessary. The system control signal is $\{Q_1, Q_2, T_s, T_{s1}, T_{s2}, \rho\}$, where Q_1 is airflow to zone 1, Q_2 is airflow to zone 2, T_s is temperature of supply air from AHU to zones, T_{s1} and T_{s2} are temperatures of supply air for each zone from their VAV boxes and ρ is fraction of return air mixed into supply air. In order to simplify the model we assume that return airflow and supply airflow are equal, which means that input and output airflows for each zone are equal.

In order to implement ASHRAE Standard 62.1, minimal outside air intake flow, V_{ot}, must be determined. Firstly, breathing zone outdoor airflow V_{bzi} must be determined for each zone:

$$V_{bzi} = R_p \cdot P_{zi} + R_a \cdot A_z \tag{10}$$

- R_p - outside airflow rate required per person, in this paper it is set to $0.0025\left[\frac{m^3}{s}\right]$;
- P_{zi} - zone population, it is assumed that zone population is known through whole horizon;
- R_a - outdoor airflow rate required per unit area, in this work it is set to $0.0003\left[\frac{m^3}{m^2 s}\right]$;
- A_z - zone floor area, in this work it is $25\left[m^2\right]$.

Next step is determining zone outdoor airflow V_{ozi}:

$$V_{ozi} = \frac{V_{bzi}}{E_z}. \tag{11}$$

Coefficient E_z is called zone air distribution effectiveness, in this work it is set to 1, since it is assumed that there is ceiling supply air and floor return in each zone.

After that uncorrected outdoor air intake, V_{ou}, must be determined:

$$V_{ou} = D\sum_{i}^{N_z}(R_p \cdot P_z) + \sum_{i}^{N_z}(R_a \cdot A_z). \tag{12}$$

The occupancy diversity, D is set to 1, since it is assumed that population of the zone is known.

Last step is determining V_{ot}:

$$V_{ot} = \frac{V_{ou}}{E_v}. \tag{13}$$

Method for determining system ventilation efficiency, E_v, and a more detailed procedure for all other coefficients, can be found in [2].

Now all necessary requirements to proceed to final step are met, required heating/cooling powers to keep temperatures of the zones in given boundaries are obtained, and minimal outside air intake in order to comply with ASHRAE Standard 62.1 is also obtained, which means that air quality in zones is preserved.

V. FINDING CONTROL SEQUENCE WITH RESPECT TO COOLING/HEATING POWERS AND VENTILATION REQUIREMENTS

After required heating/cooling powers for zones and minimal outdoor air intake are determined, control sequence for AHU and VAV boxes can be calculated. The goal is to find control sequence such that heating/cooling powers for zones generated by AHU and VAV boxes correspond to heating/cooling powers obtained through MPC, for each zone, and airflow requirements are met, while energy spent for heating or cooling and recirculating air is minimized.

Power that AHU and VAV boxes are consuming can be put into this equation:

$$P_{total} = \left|P_{phc}\right| + P_{vav1} + P_{vav2} + P_{fan}. \tag{14}$$

Power P_{total} is sum of all powers, P_{phc} is preheating or precooling power in the AHU, P_{vav1} and P_{vav2} are powers of heating air in VAV boxes going into zone 1 and zone 2 respectively and P_{fan} is power used by fan to maintain required flow.

Preheating/precooling power is equal to:

$$\begin{aligned} P_{phc} &= C_{air} Q_s \left(T_s - T_{mix}\right) \\ P_{phc} &= C_{air}\left(Q_1+Q_2\right)\left(T_s - \frac{\rho\left(Q_1T_1+Q_2T_2\right)+(1-\rho)\left(Q_1+Q_2\right)T_o}{\left(Q_1+Q_2\right)}\right) \end{aligned} \quad (15)$$

Flow Q_s is supply flow and it is equal to return flow, T_o is outside temperature, T_{mix} is temperature of mixed air and C_{air} is thermal capacity of air set to $1206\left[\frac{J}{m^3K}\right]$.

Heating powers at VAV boxes are equal to:

$$\begin{aligned} P_{vav1} &= C_{air} Q_1 (T_{s1} - T_s); \\ P_{vav2} &= C_{air} Q_2 (T_{s2} - T_s). \end{aligned} \quad (16)$$

Temperatures T_{s1} and T_{s2} are temperatures of zone supply air for zone 1 and zone 2 respectively.

Power of the supply fan is approximated by:

$$P_{fan} = SFP \cdot \left(Q_1 + Q_2\right). \quad (17)$$

Coefficient SFP is specific fan power. In this paper it is set to $2\left[\frac{kW}{m^3 s}\right]$. More about SFP can be found in [4].

The goal is to minimize these powers with respect to constraints. Problem can be written like this:

$$\min_x f\left(x\right) such \quad that \begin{cases} c\left(x\right) \le 0 \\ ceq\left(x\right) = 0 \\ A \cdot x \le b \\ lb \le x \le ub. \end{cases} \quad (18)$$

Function $f\left(x\right)$ is the function that needs to be minimized, in this case:

$$f\left(x\right) = \left|P_{phc}\right| + P_{vav1} + P_{vav2} + P_{flow}, \quad (19)$$

because total power must be minimized. Argument of the function, vector x, is:

$$x = \left[Q_1 \quad Q_2 \quad T_s \quad T_{s1} \quad T_{s2} \quad \rho\right]^T. \quad (20)$$

Function $c(x)$ is a nonlinear constraint which sets minimal outdoor air intake required:

$$c(x) = \begin{cases} \frac{V_{oz1}}{Ev} - Q_1\left(1-\rho\right) \\ \frac{V_{oz2}}{Ev} - Q_2\left(1-\rho\right). \end{cases} \quad (21)$$

Airflow V_{oz1} and V_{oz2} are minimal required air intake flows for each zone. This way ASHRAE standard is applied.

In order to keep temperature in given boundaries, heating/cooling powers input for each zone must be equal to heating/cooling power input obtained by MPC. That is ensured with nonlinear function $ceq(x)$:

$$ceq(x) = \begin{cases} C_{air} Q_1 \left(T_{s1} - T_1\right) - P_1 \\ C_{air} Q_2 \left(T_{s2} - T_2\right) - P_2. \end{cases} \quad (22)$$

Matrix A, and vector b are linear constraints used to ensure that VAV box can only heat the air going into the zone. That can be done by setting difference between AHU supply air temperature, T_s and zone air supply temperatures, T_{s1} and T_{s2} less than or equal to zero. Matrix A, and vector b are:

$$\begin{aligned} A &= \begin{bmatrix} 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \end{bmatrix}; \\ b &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{aligned} \quad (23)$$

Vectors lb and ub are lower and upper bounds for all our variables:

$$\begin{aligned} lb &= \begin{bmatrix} 0 & 0 & 283 & 283 & 283 & 0 \end{bmatrix}; \\ ub &= \begin{bmatrix} 0.5 & 0.5 & 318 & 318 & 318 & 1 \end{bmatrix}. \end{aligned} \quad (24)$$

This kind of problem can be solved using *fmnicon* function in Matlab. The results are given in next section.

VI. RESULTS

In this chapter system behavior in couple of scenarios is analyzed.

A. Heating of the zones

If temperature outside is low heating/cooling powers will be applied to zones which will keep zones temperatures on the lower zone temperature boundary, in other words, zones will be only heated. Response of temperatures in that case can be seen in Figure 3. Temperatures which are result of applying optimal heating powers obtained by MPC, and response of temperatures of the system are the same. That means that heating/cooling powers for zones are well approximated using *fmincon* function.

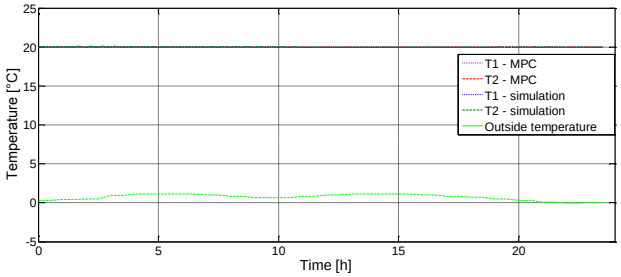


Figure 3 - Temperature response.

In Figure 4 minimal outside air intake for AHU, and zone occupancies are given. In a situation where outside air temperature is outside of the zone temperature boundaries, outdoor air intake should be minimal. It can be seen that this is true in this situation, since minimal air intake required by standard, and air intake that was obtained by MPC are the same.

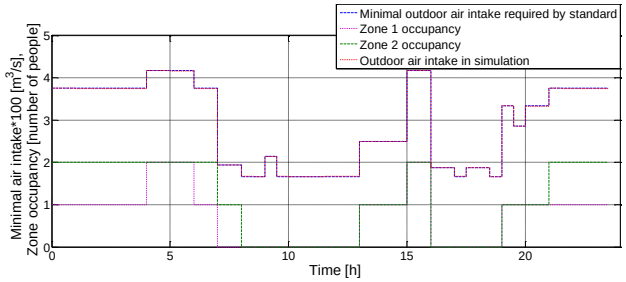


Figure 4 - Outdoor air intake and zone occupancy.

In Figure 5 total power used in system and obtained by MPC are given. Return air ratio $\rho \cdot 1000$ is also shown. It is obvious that AHU utilizes return air to minimize energy consumption. It can be seen that AHU consumes more power than obtained by MPC. That is because additional power is needed to heat up outside air coming into the system and to power the fan.

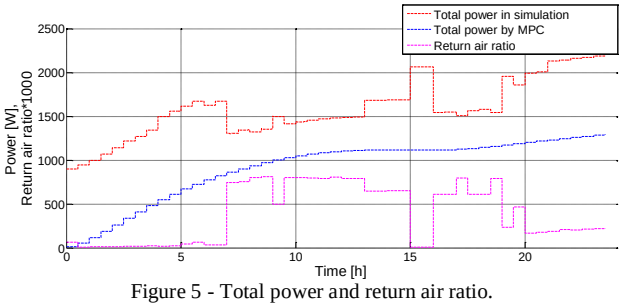


Figure 5 - Total power and return air ratio.

B. *Cooling the zones*

Cooling of the zones is the same as heating, if the outside air is above upper zone temperature boundary. But when the outside air temperature is inside given zone temperature boundaries, AHU should be able to utilize that to minimize energy consumption.

In Figure 6 temperature response is given. As before, zone temperatures are inside boundaries. All the times when outside temperature is above the required temperature interval for the zones and zone temperatures are on the upper boundaries, it can be concluded that exertion of cooling power to zones is needed to keep the zone temperature inside temperature boundaries.

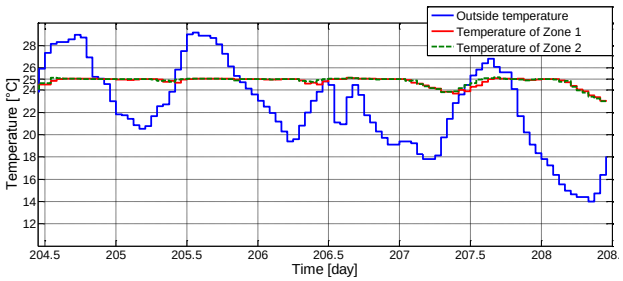


Figure 6 - Temperature response.

In Figure 7 outdoor air intake and zone occupancies are given. It can be seen that AHU takes in more outside air than required by standard, which means that AHU is utilizing outside air in order to minimize energy consumption.

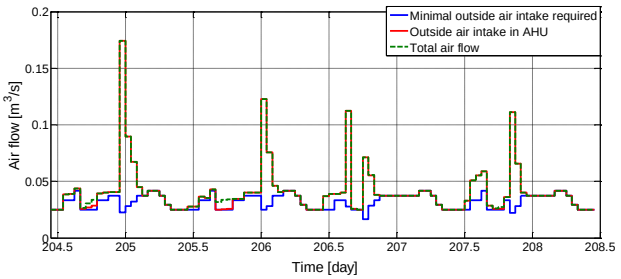


Figure 7 - Outdoor air intake.

In Figure 8 total power and return air ratio are given. It can be seen that AHU sometimes uses less power than demanded to be injected into zones by MPC. This confirms the assumption that AHU utilizes outside air in order to minimize energy consumption while delivering the right amount of cooling energy into the zones.

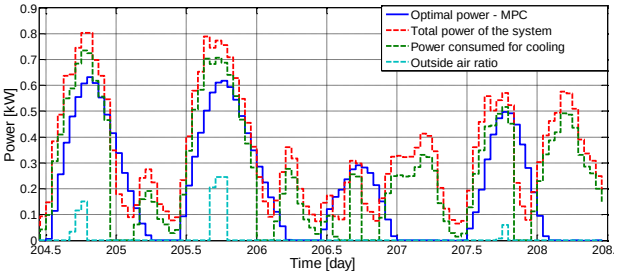


Figure 8 - Power response.

The case when the designed AHU/VAV control performs sub optimally is when cooling/heating powers obtained by MPC are zero and temperatures of the zones are somewhere inside zone temperature boundaries. Since AHU an VAV boxes produce equal cooling/heating powers as the ones obtained by MPC, and since AHU must take some air from the outside, it is possible that outside air will be heated (or cooled) to match that cooling/heating power. But since the temperature is not

on the lower or upper boundary it is possible that more energy than needed will be consumed.

VII. Conclusion

It is clear that obtaining control sequence for AHU and VAV boxes in a manner described in this work could be used in practice. The nonlinearity and scalability problems are addressed by decomposing the control problem in two stages, computation of heating/cooling power exertions for zones and static optimization of HVAC system operation to adhere to the required power exertions to zones while ventilation constraint is also taken into account.

Acknowledgments

This work has been financially supported by the European Union through project ENHEMS-Buildings -- Enhancement of Research, Development and Technology Transfer Capacities in Energy Management Systems for Buildings, grant No. IPA2007/HR/16IPO/001-040510.

References

[1] A. Starcic, Modeliranje i upravljanje toplinskim procesom u zgradi, Zagreb: Faculty of Electrical Engineering and Computing, 2010.

[2] ANSI/ASHRAE Standard 62.1-2007, Atlanta: ASHRAE, Inc., 2007.

[3] A. Martincevic, Modelsko prediktivno upravljanje toplinskim procesom u zgradi uz uzimanje u obzir nesigurnosti podataka iz vremenske prognoze, Zagreb: Faculty of Electrical Engineering and Computing, 2011.

[4] J. Railio i P. Mäkinen, »SPECIFIC FAN POWER – a tool for better performance of air handling,« , *Clima 2007 WellBeing Indoor*, Helsinki, 2007.

Tuning of model predictive controller for wind turbine load reduction

Goran Banjac, Vedrana Spudić and Mato Baotić
Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia
E-mail: {goran.banjac, vedrana.spudic, mato.baotic}@fer.hr

Abstract – **Classical objective of wind turbine control is to maximize the power extracted from wind. However, with the increase in wind turbine size the structural loads become a limiting factor for safe wind turbine operation and wind turbine operating life. In this paper we investigate the design of a model predictive controller that reduces wind turbine loads at the expense of wind turbine power variability. By appropriate tuning of the model predictive controller we show that the trade-off between load reduction and power variations is favorable.**

I. INTRODUCTION

The use of wind energy for electrical energy production has experienced an increasing growth in last two decades, which was caused by the growth in electrical energy consumption and by the requirement to reduce the emissions of CO_2 and other harmful gases [1]. Wind turbine development led to increase in its rated power and, consequently, in its dimensions and mass. As a result of the increase of wind turbine construction, the wind turbine is exposed to increased loading which can cause an excessive wear and tear of the wind turbine components. The reduction of wind turbine structural loads can be achieved by a redesign and/or improvements of wind turbine controller. One typical solution is an individual pitch control algorithm described in [2].

In this paper we investigate the trade-off between wind turbine power variability and structural loads in high winds. Namely, a typical objective in wind turbine control during high winds is to keep the power output constant, which inevitably causes variations of structural loads. The idea investigated here is to reduce structural loads on behalf of power tracking relaxation, i.e. allowing certain variability of the power output. Initial considerations showed that the relaxation of the power tracking requirements on behalf of structural load penalization can lead to reduction of damage equivalent loads, which can be achieved by model predictive controller design [3]. In this paper we further explore this trade-off and look for an optimal ratio between those objectives.

In this paper we consider a control scheme in which the wind turbine controller remains the same (we use a conventional wind turbine controller design), while we design an add-on on the supervisory level – a load reduction controller (LCR). The load reduction controller sets the power reference P_{ref} to the wind turbine controller, as depicted in Fig. 1. The load reduction

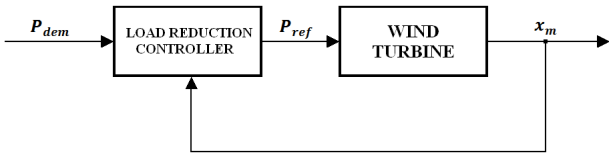


Figure 1. The supervisory controller for load reduction

controller receives an external power demand, P_{dem}, which the wind turbine needs to produce on average. Typically the power demand is equal to the nominal power. This controller configuration was used in [4]. The load reduction controller and the wind turbine controller are separated by time scales – wind turbine controller operates at a sampling rate of 12.5 ms, while the load reduction controller operates at 1 second sampling time [5]. The load reduction controller adapts the operating point of the plant to the disturbance from wind, while the wind turbine controller ensures tracking of this operating point.

The paper is structured as follows:
Section II describes the wind turbine operation, including the wind turbine control and the arising wind turbine structural loads. Section III describes the model predictive control formulation which comes to a quadratic optimization. Section IV shows the wind turbine model used in the control algorithm performance analysis. In section V the principle for load reduction is presented. Section VI shows the results. In section VII we give some concluding remarks.

II. WIND TURBINE OPERATION

Wind turbine is an electromechanical system that converts the kinetic energy of wind to electrical energy [1]. In the last 30 years, a great progress has been achieved in wind turbine development – from the initial power of about 30 kW and rotor diameter of about 15 m to the power of few MW and rotor diameter of more than 120 m. The increase in power of two orders of magnitude in just 30 years shows the achievement which has not been recorded for any other electrical plant in the history.

The primary aim of a wind turbine control system is to enable the delivery of the required amount of power. The maximal available power is a function of wind speed, wind turbine rotational speed and turbine's geometry [6]. For wind turbine control, two control variables are available – the pitch angle of the blades and the torque reference to the electrical generator. Changing the pitch

angle of the blades, wind turbine rotational speed can be controlled. Two basic wind turbine control configurations are tracking the demanded power and maximizing power production.

A. Wind turbine operation principle

The interaction between the wind flow and wind turbine blades is referred to as wind turbine aerodynamics. This interaction is governed by principles of fluid dynamics – the wind turbine blades rotate in a wind field from which they extract momentum. The actuator disc model describes the physical principles of the kinetic wind energy conversion into the kinetic energy of wind turbine rotation and potential energy which is manifested by rotor thrust force [7]. According to the actuator disc model, the power captured by the wind turbine, P, is modeled by:

$$P = \frac{1}{2} R^2 \pi \rho C_P(\lambda, \beta) v_{rot}^3, \quad (1)$$

where R is the rotor radius, ρ is air density, C_P is the power coefficient, β is the collective pitch angle, $\lambda = \frac{\omega_r R}{v_{rot} - \dot{x}_t}$ is the tip speed ratio (ω_r is rotor speed, $\dot{x}_t$ is the tower top velocity) and v_{rot} is the rotor effective wind speed.

The rotor effective wind speed is an artificial signal that depicts the effect of the turbulent wind over the entire rotor in such manner that the rotor torque determined from it by using static aerodynamic characteristic is equal to the rotor torque obtained by the uneven wind field over the entire rotor. For details see [8].

Static expressions for wind turbine rotor torque T_r and thrust force F_t are given as:

$$T_r = \frac{1}{2} R^3 \pi \rho C_Q(\lambda, \beta) v_{rot}^2, \quad (2)$$

$$F_t = \frac{1}{2} R^2 \pi \rho C_T(\lambda, \beta) v_{rot}^2, \quad (3)$$

where $C_Q = \frac{C_P}{\lambda}$ is the torque coefficient and C_T is the thrust coefficient.

Rotor thrust force that acts on rotor in the fore-aft plane causes the tower top motion. The tower bending moment M_{tower} is modeled as the elastic force obtained by tower top deflection acting on the tower arm [8]:

$$M_{tower} = -F_t \cdot H, \quad (4)$$

where H is the tower height.

The wind turbine transmission system consists of a low speed shaft, a high speed shaft and a gearbox that connects them. The low speed shaft is relatively long and therefore torsionally flexible. The low speed shaft torsional moment is modeled as the elastic torque accumulated at shaft:

$$T_{shaft} = -k\phi, \quad (5)$$

where k is equivalent drive train torsional spring constant and ϕ is torsional angle of the low speed shaft.

B. Wind turbine control principle

Wind turbine control is typically implemented as a hybrid control system that comprises two different ranges of operation: below and above rated wind speed, [1].

For wind speeds below rated, wind turbine extracts the power which is less than the rated generator power. In order to maximize the plant efficiency, the objective of wind turbine control for wind speeds bellow rated is to maximize the power extracted from wind. This is achieved by rotor speed control – it can be shown that for every wind speed below rated there exists the speed of wind turbine rotor for which the power extraction efficiency is optimal [7].

For wind speeds above rated, the maximal power that can be extracted from wind is larger than the rated generator power. The control objective in this case is to limit the power extracted from wind which is performed by adjusting pitch angle of the blades. Classical PI controller is often used for performing this task. For more details about wind turbine internal controller implementation, see e.g. [7].

The classical control algorithm is primarily aimed at nominal speed and nominal power tracking, and does not explicitly take structural loads into account. With the increase of wind turbine dimensions, structural loads are considerable and the controller design that reduces the structural loads on account of tracking can be beneficial for the overall wind turbine operation.

C. Wind turbine loads

Wind turbine aerodynamics, gravitation and inertia are some of the causes of wind turbine loads. Wind turbine is not a static structure – rotor rotates and, due to large and slender structure, the structural vibrations are introduced, which significantly contribute to dynamic loading of the structure. For large wind turbines it is typical that the dynamical loads, which cause material fatigue, are the dominant factor for component dimensioning since it determines the lifetime of the installation [7].

For controller design that aims at fatigue reduction, one requires an appropriate measure of fatigue, i.e. the measure of damage that is caused during given operating conditions. To that purpose the concept of damage equivalent loads is used [9]. Damage equivalent loads can be obtained from a given load in two steps: (1) extracting load cycles from the load history by the rainflow counting algorithm, and (2) computing damage equivalent loads by utilizing the Miner's rule. The details of this procedure are described in [8]. Since the damage equivalent loads cannot be described in an explicit mathematical form, for controller design we use an approximation first introduced in [5] (described in Section IV), while for controller validation we use appropriate *Matlab* toolboxes developed for damage equivalent loads computation along with wind turbine models developed for structural load simulations (described in Section V).

III. MODEL PREDICTIVE CONTROL

Model predictive control, MPC, is an advanced control technique which had a significant and widespread impact on industrial process control [10]. MPC partly gained confidence in the industrial environment even before the

theoretic analysis was completed. Reasons for that are: (1) MPC is a generic control technique which can routinely and inherently deal with safety and technical constraints because they are *a priori* taken into account while calculating the control signal, (2) basic MPC formulation is naturally extended to the complex Multiple-Input/Multiple-Output systems with almost no modifications and (3) it allows simple parameter tuning.

In the beginning, MPC was limited only for the slow processes, e.g. in the chemical industry, because of the optimization problem which required quite large computer and time resources. Optimization techniques development and computer power growth made wider MPC applications possible. Meanwhile, the theoretic MPC analysis has been completed and the strict mathematical conditions which have to be satisfied for MPC stability guaranty have been defined [11].

Standard linear MPC uses discrete linear time-invariant model of the process which is given as follows:

$$\begin{aligned} x(t+1) &= A \cdot x(t) + B \cdot u(t) + F, \\ y(t) &= C \cdot x(t) + D \cdot u(t) + G, \end{aligned} \tag{6}$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^m$ is the input, $y \in \mathbb{R}^p$ is the output, $A \in \mathbb{R}^{n \times n}$ is the state matrix, $B \in \mathbb{R}^{n \times m}$ is the input matrix, $C \in \mathbb{R}^{p \times n}$ is the output matrix, $D \in \mathbb{R}^{p \times m}$ is the feedforward matrix, $F \in \mathbb{R}^n$ is the state shift and $G \in \mathbb{R}^p$ is the output shift. Parameter t indicates discrete-time step. The constant matrices F and G (which are not typically included in linear plant model) are used in order to use the absolute values of the states, instead of deviation from the operating point. Thus, a more natural tracking problem can be defined. The addition of constant matrices does not influence the structure of the optimization problem at hand.

The optimization problem related to the model predictive control for optimal tracking problem is defined as:

$$\begin{aligned} \min_{u(0),\cdots,u(N-1)} & \sum_{k=0}^{N-1} \big(y(k) - y_{sp}\big)^T Q \big(y(k) - y_{sp}\big) + \\ & + \big(u(k) - u_{sp}\big)^T R \big(u(k) - u_{sp}\big) \end{aligned} \tag{7}$$

$$\text{subj. to} \begin{cases} x(k+1) = A \cdot x(k) + B \cdot u(k) + F, \\ y(k) = C \cdot x(k) + D \cdot u(k) + G, \\ u_{min} \le u(k) \le u_{max} \end{cases}$$

where $Q \in \mathbb{R}^{p \times p}$, $Q = Q^T \succcurlyeq 0$ and $R \in \mathbb{R}^{m \times m}$, $R = R^T \succ 0$ are weight matrices, $u_{sp} \in \mathbb{R}^m$ is the input set point, $y_{sp} \in \mathbb{R}^p$ is the output set point and $N \in \mathbb{N}$ is the prediction horizon. The optimization problem (7) is typically referred to as Constrained Finite-Time Optimal Control (CFTOC) problem.

In model predictive control the control signal is calculated using predictions of the system states obtained from model of the process (1). Model predictive control methodology comprises the following steps:

1. Measure or estimate process state $x_m(t)$ at the time step t,
2. Formulate and solve the optimization problem (7) for $x(0) = x_m(t)$,
3. Apply $u(t) = u(0)$ to the process,
4. Increment discrete-time step: $t \leftarrow t + 1$. Return to step 1.

This algorithm is repeated at every time step. This control strategy is called *Receding Horizon* and it introduces feedback into the control system.

IV. Model Predictive Control for Wind Turbine Load Reduction

Damage equivalent load is a convenient measure for qualitative comparison between load histories. Since the damage equivalent loads cannot be described in an explicit mathematical form, it is not suitable for direct use in the cost function of the optimization problem. Therefore, the aim is to find the simpler objective formulation, which approximates DEL. DEL approximation concept is introduced in [3] and it shows that by penalization of (the low-frequent part of) the deviations of thrust force and shaft torque from its set points the reduction of the damage equivalent loads of structural loads mostly influenced by aerodynamic forces can be achieved. The set points of the structural loads are easily computed from the mean wind speed and the power demand, see [3]. As in [3], we focus here on the shaft and tower loads, which can be easily interpreted as a direct consequence of the thrust force and rotor torque.

In this paper we use the wind turbine model developed by the National Renewable Energy Laboratory (NREL), which is widely considered as a representative of a typical utility-scale land- and sea-based multimegawatt turbine [12]. It is a conventional three-bladed wind turbine with adjustable rotor speed and blade pitch angle. This model is used by the research teams from all over the world as a standard in offshore wind turbine specifications [12], and it is also used in this work for testing the control algorithm behavior. The NREL wind turbine model consists of the closed loop nonlinear system containing typical controllers mentioned in Section II.

To design the load reduction controller we use a linear discrete-time state space model derived in [6]:

$$\begin{aligned} x(k+1) &= a \cdot x(k) + d \cdot u(k) + \\ &\quad + b_d \cdot d_v(k) + f, \\ y(k) &= c \cdot x(k) + d \cdot u(k) + \\ &\quad + d_d \cdot d_v(k) + g, \end{aligned} \tag{8}$$

where $x = \begin{bmatrix} \beta & \omega_r & \omega_g^{filt} \end{bmatrix}^T$ is the state vector (ω_g^{filt} is the filtered generator speed), $u = \begin{bmatrix} P_{ref} \end{bmatrix}$ is the input, $d_v = [v_{rot}]$ is the disturbance (v_{rot} is the rotor effective wind speed), $y = [F_t \quad T_{shaft}]^T$ is the output, $f \in \mathbb{R}^3$ and $g \in \mathbb{R}^3$ are state shift and output shift which came as the result of linearization of the nonlinear model [6]. f and g are operating point dependent constant vectors and a, b, b_d, c, d and d_d are operating point dependent constant matrices of the appropriate dimensions.

The rotor effective wind speed is a stochastic variable, however it can be well estimated from the typical wind

turbine measurements. For the use of model predictive controller, prediction of all the variables should be possible. Here, we use a persistence model to predict the effective wind speed. Since the effective wind speed is averaged over the large rotor area, it has slower dynamics and therefore it can be considered constant throughout the short prediction horizon we use here. In order to pose the optimization problem (7), the state space model (8) should be transformed in the form of model (6). Thus, using an assumption that wind speed is constant over the prediction horizon, the state space model (8) can be transformed as follows:

$$\begin{bmatrix} x(k+1) \\ d_v(k+1) \end{bmatrix} = \begin{bmatrix} a & b_d \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x(k) \\ d_v(k) \end{bmatrix} + \begin{bmatrix} b \\ 0 \end{bmatrix} u(k) + \begin{bmatrix} f \\ 0 \end{bmatrix},$$
$$y(k) = \begin{bmatrix} c & d_d \end{bmatrix} \begin{bmatrix} x(k) \\ d_v(k) \end{bmatrix} + d \cdot u(k) + g. \quad (9)$$

This model can be readily used to formulate the optimization problem (7) and thereby design the model predictive controller. The set points u_{sp} and y_{sp} can be determined from the static model of the NREL wind turbine derived in [6].

The optimization problem (7) is a quadratic program, which we solve by using the *Matlab* function *quadprog* [13].

V. TUNING OF MODEL PREDICTIVE CONTROLLER FOR WIND TURBINE LOAD REDUCTION

The objective of load reduction controller is to reduce wind turbine structural loads. However, in the controller design we use an indirect approach – we reduce the variability of the thrust force and shaft torque, which should lead to reduction of aerodynamically introduced DEL. For the controller tuning we evaluate the final objective, i.e. the minimization of the tower bending moment DEL and the low speed shaft torque DEL. The criterion for the controller tuning is therefore defined as:

$$f(Q,R) = a\left(\frac{DEL(T_{shaft})}{DEL_0(T_{shaft})}\right)^2 + b\left(\frac{DEL(M_{tower})}{DEL_0(M_{tower})}\right)^2, \quad (10)$$

where $DEL(\cdot)$ denotes DEL of the component for the current controller parameters Q and R, $DEL_0(\cdot)$ denotes DEL of the component at constant power requirement, i.e. DEL calculated for the same operating conditions but without using load reduction controller, while a and b are parameters which determine the criterion function, e.g. the greater ratio between a and b, the greater DEL reduction of low speed shaft with respect to DEL reduction of tower.

For determining DEL of the component, *MCrunch* is used. *MCrunch*, published by NREL, is a set of *Matlab* scripts that allows determining damage equivalent loads of wind turbine components using the wind turbine measurements after simulating the wind turbine operation [14]. It is a postprocessor utility, thus it cannot be used for online determining DEL as a result of current control actions, but it can be used for tuning of MPC parameters.

For fixed parameters a and b, the optimal weight matrices Q and R are determined using *Matlab* *fminsearch* function which adjusts weight matrices in order to minimize the function f defined in (10). *Fminsearch* is the unconstrained multivariable nonlinear optimization function which uses derivative-free method for finding the function minimum [15].

VI. RESULTS

For the operating point, defined by the power demand P_{dem} of 3 MW and rotor effective wind speed v_{rot} of 18 m/s, and a set of values of the parameters a and b defined in (10), optimal values of the weight matrices Q and R are obtained. Table I shows the relative changes of damage equivalent loads for optimal weight matrices. Notation $\Delta\,\mathrm{DEL}(\cdot)$ stands for $\mathrm{DEL}(\cdot) - \mathrm{DEL}_0(\cdot)$.

It can be seen that larger value of the parameter b results in larger DEL reduction for the tower. With same values of the parameters a and b, much larger DEL reduction is obtained for low speed shaft in relation to tower DEL reduction. Approximately same extent of DEL reduction for low speed shaft and tower are obtained for parameter values $a = 0.15$ and $b = 0.85$.

Fig. 1 shows the output from load reduction controller, P_{ref}, for optimal values of matrices Q and R obtained for the equal values of parameters a and b, power demand P_{dem} of 3 MW and effective rotor wind speed v_{rot} of 18 m/s with turbulent component included. Constraints on the control input value P_{ref} are set as follows:

$$P_{ref} \in [1\ MW, 5\ MW] \cap [0.75\ P_{dem}, 1.25\ P_{dem}]. \quad (11)$$

The first part represents power constraint for the NREL wind turbine and the second part defines maximal deviation of MPC output P_{ref} from the power demand P_{dem}. The second part of the constraint (11) is defined by a user. Fig. 1 shows that the output signal is quite far from defined constraints. Maximal relative deviations of the LRC control input P_{ref} from the power demand P_{dem} for a set of parameters a and b are shown in the Table II. It is shown that, for larger value of parameter b, P_{ref} has larger maximal deviation from the P_{dem}. For parameter values $a = 0.15$ and $b = 0.85$ for which almost the same relative changes of DEL are obtained, P_{ref} oscillates around P_{dem} with maximal variation of about 7.5 %, thus the cost for load reduction is variation in wind turbine produced power.

Fig. 3 shows low speed shaft torque and tower bending moment for equal values of parameters a and b without load reduction control and with LRC included. It can be seen that the LRC managed to find such the power reference sequence which reduces deviations of the low speed shaft torque from its set point $T_{shaft,sp}$. Dynamics of signals for each component is very similar, but low speed shaft torque signal with LRC is much closer to the set point value of the low speed shaft torque. Differences between tower bending moment without and with LRC are not obvious from Fig. 2, what corresponds to load reduction of 2.42 % from Table I. Different time scales for low speed shaft torque and tower bending moment are chosen because of much greater time constant of tower

TABLE I. RELATIVE CHANGES OF DAMAGE EQUIVALENT LOAD

a	b	$\frac{\Delta \mathrm{DEL}(T_{shaft})}{\mathrm{DEL}_0(T_{shaft})}$	$\frac{\Delta \mathrm{DEL}(M_{tower})}{\mathrm{DEL}_0(M_{tower})}$
0.5	0.5	-10.89 %	-2.42 %
0.2	0.8	-7.73 %	-3.54 %
0.15	0.85	-5.74 %	-3.96 %
0.1	0.9	+1.78 %	-5.08 %
0	1	+43.04%	-6.69 %

TABLE II. MAXIMAL VARIATIONS OF LRC CONTROL INPUT FROM POWER DEMAND

a	b	$\frac{\max(P_{ref}) - P_{dem}}{P_{dem}}$	$\frac{\min(P_{ref}) - P_{dem}}{P_{dem}}$
0.5	0.5	+6.20%	-4.69 %
0.2	0.8	+6.73 %	-4.94 %
0.15	0.85	+7.52 %	-5.48 %
0.1	0.9	+9.90 %	-7.26 %
0	1	+16.98%	-13.24 %

swinging in relation to the time constant of low speed shaft oscillations.

In order to validate the parameters of load reduction controller which are obtained using *Matlab* unconstrained nonlinear optimization function *fminsearch*, the controller is tested for the other operating conditions, keeping the same calculated values of the weight matrices Q and R. New operating point has to satisfy the same conditions as the first one: rated power has to be below rated and the wind speed has to be above rated one. New operating point is defined by power demand P_{dem} of 2.5 MW and effective rotor wind speed v_{rot} of 15 m/s containing the new turbulent component. Tables III and IV show the variables presented in Tables I and II, but simulated in new operating conditions. DEL reduction for this operating point is a bit smaller, but the same trend in a relative change of DEL for different parameter values can be seen. Table IV shows that the change of operating point led to the increase in maximal variation of power reference P_{ref} from power demand P_{dem} of about 30 %.

VII. CONCLUSION

The paper analyses the design of model predictive controller that reduces wind turbine structural loads at the expense of wind turbine power variability. Load reduction controller is designed on the supervisory level while the internal wind turbine controller remains the same.

Results of the wind turbine simulation with the load reduction controller show that significant reduction of wind turbine structural loads can be achieved through relatively small variations in produced power. Depending on critical components of wind turbine in view of structural loads, the LRC parameters can be adjusted in order to put the larger emphasis on the critical component load reduction. Simulations show that the parameters optimized for one operating point have a favorable behavior in the new operating point too.

ACKNOWLEDGEMENT

This research has been supported by the European Communitys Seventh Framework Programme under grant No. 285939 (ACROSS), by the Croatian Science Foundation under grant No. I-4463-2011 (MICROGRID), and by the Ministry of Science, Education and Sports of the Republic of Croatia under grant No. 036-0361621-3012.

REFERENCES

[1] Burton, T.; Sharpe, D.; Jenkins, N.; Bossanyi, E. (2001). Wind Energy Handbook. John Wiley & Sons, Ltd.

[2] Bossanyi, E. A. (2002). Individual Blade Pitch Control for Load Reduction. Wind Energy.

[3] Spudić, V.; Jelavić, M.; Baotić, M. (2011). Wind Turbine Power References in Coordinated Control Wind Farms. Automatika.

[4] Spudić, V.; Jelavić, M.; Baotić, M. (2012). Explicit model predictive control for reduction of wind turbine structural loads.

[5] Spudić, V.; Jelavić, M.; Baotić, M; Perić, N. (2010). Hierarchical wind farm control for power/load optimization. Torque, Crete, Greece

[6] Spudić, V.; Jelavić, M.; Baotić, M.; Vašak, M.; Perić, N. Reconfigurable control extension: Wind turbine model. AEOLUS, University of Zagreb, Faculty of Electrical Engineering and Computing.

[7] Jelavić, M. (2009). Wind turbine control for structural dynamics loads reduction. PhD thesis, University of Zagreb, Faculty of Electrical Engineering and Computing.

[8] Spudić, V. (2012). Coordinated optimal control of wind farm active power. PhD thesis, University of Zagreb, Faculty of Electrical Engineering and Computing.

[9] Sutherland, H. J. (1999). On the Analysis of Wind Turbines. Sandia National Laboratories.

[10] Maciejowski, J. M. (2000). Predictive Control with Constraints. Prentice-Hall, Inc.

[11] Borelli, F.; Bemporad, A.; Morari, M. (2011). Predictive control for linear and hybrid systems. Cambridge.

[12] Jonkman, J.; Butterfield, S.; Musial, W.; Scott, G. (2009). Definition of a 5-MW Reference Wind Turbine for Offshore System Development. Technical Report, Nation Renewable Energy Laboratory.

[13] Matlab Help R2011b Documentation, Optimization Toolbox. http://www.mathworks.com/help/optim/ug/quadprog.html

[14] Buhl Jr., M.L. (2012). MCrunch User's Guide for Version 1.00. Nation Renewable Energy Laboratory.

[15] Matlab Help R2011b Documentation, Optimization Toolbox. http://www.mathworks.com/help/optim/ug/fminsearch.html

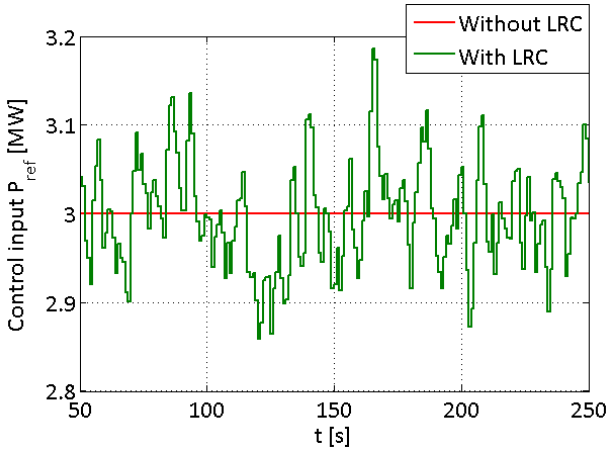


Figure 2. Control input which represents power reference for wind turbine

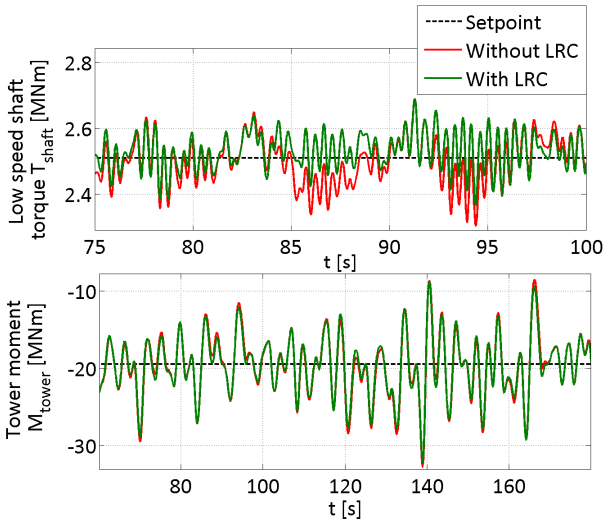


Figure 3. Low speed shaft torque and tower bending moment with applied P_{ref} depicted in Fig. 2

TABLE III. RELATIVE CHANGES OF DAMAGE EQUIVALENT LOAD – VALIDATION

a	b	$\frac{\Delta\,\mathrm{DEL}(T_{shaft})}{\mathrm{DEL}_0(T_{shaft})}$	$\frac{\Delta\,\mathrm{DEL}(M_{tower})}{\mathrm{DEL}_0(M_{tower})}$
0.5	0.5	-13.77 %	-1.49 %
0.2	0.8	-9.86 %	-2.61 %
0.15	0.85	-6.81 %	-2.88 %
0.1	0.9	+3.28 %	-3.43 %
0	1	+45.89%	-2.91 %

TABLE IV. MAXIMAL VARIATIONS OF LRC CONTROL INPUT FROM POWER DEMAND – VALIDATION

a	b	$\frac{\max(P_{ref}) - P_{dem}}{P_{dem}}$	$\frac{\min(P_{ref}) - P_{dem}}{P_{dem}}$
0.5	0.5	+8.03 %	-6.79 %
0.2	0.8	+8.91 %	-6.94 %
0.15	0.85	+9.97 %	-7.66 %
0.1	0.9	+13.14 %	-9.94 %
0	1	+22.35%	-17.22 %

Control of the Grid Side Converter in a Wind Turbine

Tamara Hadjina, Mato Baotić, Nedjeljko Perić
Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia
Email: tamara.radosevic@fer.hr, mato.baotic@fer.hr, nedjeljko.peric@fer.hr

***Abstract*—Due to constantly increasing share of wind power in total installed power, transmission system operators must adhere to ever stricter grid codes in order to maintain stable, secure and reliable power system operation. Grid codes require from wind power plants to participate in active and reactive power control and frequency control. In particular, wind power plants need to have fault-ride through capability and must respect power quality demands of the grid codes. To satisfy those requirements wind turbines utilize power electronic devices - converters - as an interface between the wind generator and the grid. A lot of research effort is focused on improvement in the design and control of the grid side converter, which connects the turbine to the grid. This paper presents some of the established grid side converter control strategies. The challenges regarding grid side converter control are also discussed with a brief overview of recent research trends.**

I. Introduction

In recent decades we are witnessing rapid development of grid-connected wind systems and constantly increasing penetration of wind power [1]–[3]. In order to be able to increase the wind energy penetration, new grid interconnection requirements called grid codes have been developed by countries with a high wind power share.

Power converters are the interface that enables an efficient and flexible interconnection of wind power plants (WPPs) and other renewable energy sources to the electric power system [4]. An effective use of power electronics, especially on the grid side, is allowing WPPs to behave similarly to conventional power plants [5].

Variable speed wind turbine (WT) with synchronous generator and full-power frequency converter is, currently, one of the most promising wind turbine configurations [5]. This type of WT provides great flexibility by enabling full active and reactive power control and fault ride-through capability during voltage dips. The variable speed wind turbine with synchronous generator is connected through a frequency converter to the grid as shown in Fig.1. The frequency converter, in literature often referred to as back-to-back converter, consists of two three-phase voltage converters and a DC link. Since the DC link effectively decouples the generator from the grid, the grid side converter must handle all grid connection issues.

In this paper grid code requirements are discussed with special emphasis put on the requirements which can be fulfilled by a proper control of the grid side converter. The grid side converter control strategy is described and two most popular current control strategies are discussed. Finally, an example is given that illustrates good and bad aspects of the current control strategy in stationary reference frame.

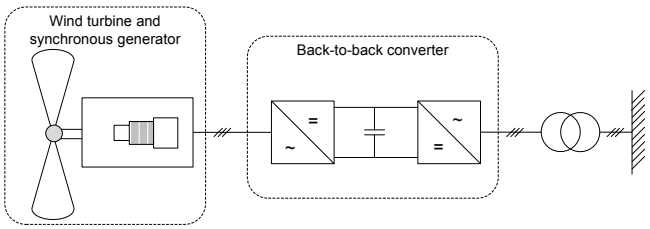


Fig. 1. Variable speed wind turbine with synchronous generator and full scale frequncy converter

II. Grid code requirements

Grid codes (GCs) give technical regulations regarding the connection of wind power plants to the transmission system. Some of the GCs criteria vary significantly among the countries with high penetration level of wind energy (e.g. Germany, Denmark, Spain, Belgium). However, the European Wind Energy Association (EWEA) recommends that regulations for the European grid connection should be developed in a more consistent manner [6]. The general requirement, which is common to all of the GCs, is that the wind power plants should behave in the same way as conventional power plants based on large synchronous generators [3].

The following requirements are common in most grid codes [2]:

- Normal operation behaviour:
  - frequency and voltage ranges,
  - active power control,
  - reactive power control,
  - high power quality.
- Behavior under grid faults:
  - fault ride through (FRT),
  - reactive current injection (RCI).

These requirements are widely discussed in literature [5], [7]–[9].

In normal operating conditions WPP must operate within given voltage and frequency variation limits. WPP is also

expected to take part in grid operation and control. This is achieved by regulation of WPP's output active and reactive power. Depending on the WT type, different strategies are used to follow the active/reactive power reference which is forwarded to the WPP by the transmission system operator. In the case of the WT type given in Fig. 1 active power is controlled directly by means of generator torque and pitch (position ot WT blades) control, and indirectly through DC-link voltage control. Keeping the DC-link voltage at a constant level assures that the total available active power is transfered to the grid. Since the DC-link completely decouples the reactive power of the generator from the reactive power of the grid, the reactive power injected to the grid can be controlled by the grid side converter.

Power quality requirements are expressed through permitted levels of grid current harmonic distortion. In particular, the injected current should not have the so-called total harmonic distortion (THD) larger than 5%:

$$THD = \frac{\sqrt{\sum_{h=3}^{50} I^2(h)}}{I(1)} \quad (1)$$

where h denotes the h^{th} harmonic and $I(h)$ is the RMS value of the h^{th} harmonic of the grid current [10].

Fault ride through requirements dictate the behaviour of the WPP under different types of voltage sags and are therefore also called low voltage ride through (LVRT) requirements. Voltage sags can be roughly divided into balanced and unbalanced faults. Balanced faults imply an equal voltage drop in all three phases while unbalanced faults involve different voltage drops in different phases and a possible phase shift. A widely accepted classification of voltage sags is presented in [11]. Wind turbines are expected to withstand voltage dips down to a certain percentage of the nominal voltage and for a specific duration. Note that some grid codes demand from the WT to stay connected to the grid after the voltage drops to zero [3].

One of the most demanding GC requirements is the reactive current injection during grid faults. The aim of the increased RCI is to support the system voltage during voltage sags. Some grid codes require WPPs to follow the reactive current reference with response time of 20 ms.

In the case of the WT connected to the grid through a full scale power converter one can dominantly affect the power quality, reactive power control, and FRT and RCI abilities of WPPs by the proper control of the grid-side converter.

III. Grid-side converter control

A grid-side converter, which converts DC link voltage to AC voltage in order to inject AC current to the grid, is a three-phase two-leg voltage source inverter with six switches, usually IGBT with a diode in parallel [12]. Voltage source converter is a force comuted converter whose switching states are obtained using some of the well established modulation techniques such as pulsewidth modulation (PWM) or space vector modulation (SVM) [13], [14]. PWM converters are widely used as they provide bidirectional power transfer with low harmonic current distortion. A grid-connected PWM converter with LCL filter is shown in Fig.2.

The grid-side converter control is usually based on cascaded controllers. The outer loops, with slower dynamics, are controlling the reactive power delivered to the grid and the dc-link voltage, U_{dc}. The outer loops generate references ($i_{g,ref}$) for the faster inner loops which control the current injected to the grid [7], [15].

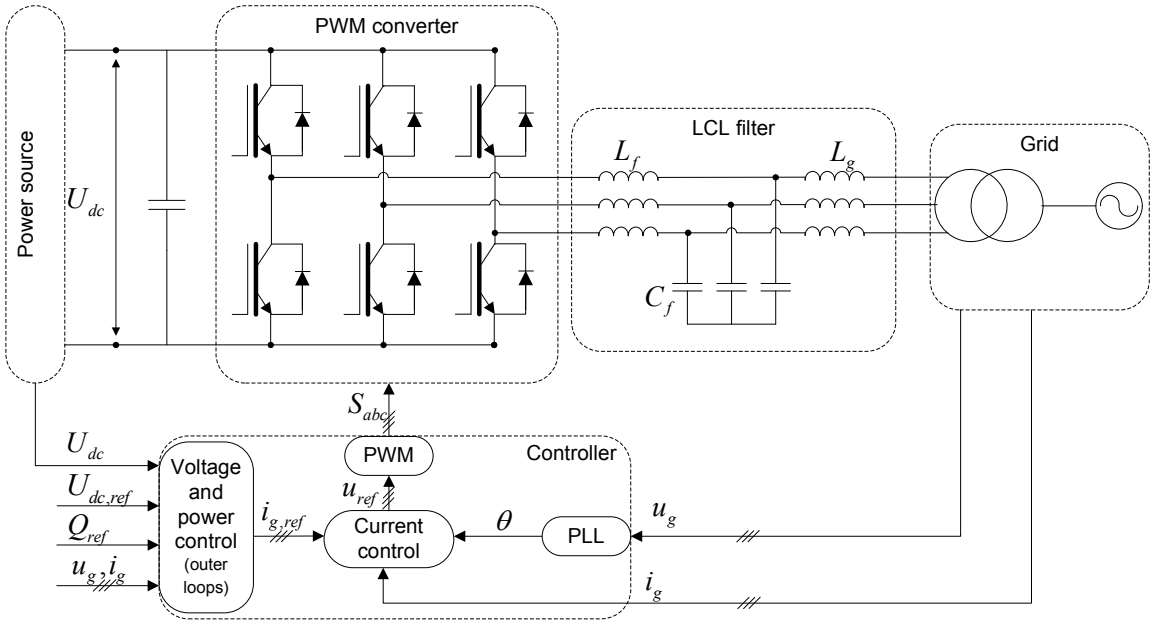


Fig. 2. Grid-side converter control system

IV. CURRENT CONTROL

The current control loops are critical parts of the converter control system since they are sensitive to grid voltage variations while, by the grid code requirements, they must ensure fast dynamic response [5], [7]. One of the major responsibilities of the current controller is the power quality. Consequently, one can find variaty of methods in literature describing compensation of grid harmonics (in order to obtain an improved power quality) [15]–[17].

The grid current harmonic distortion comes from two different origins. The low frequency harmonics are caused by the grid voltage distortion. Dealing with low frequency harmonics is a control issue connected to the robustness of the controller against grid disturbances. The high frequency grid current harmonic distortion is related to harmonics at the PWM frequency of the converter. Mitigation of the PWM harmonics is accomplished by using LCL filters because of their good filtering characteristic above resonance frequency. However, resonance may trigger undesired oscillations or even instability. To avoid filter resonance passive damping resistor can be added to the filter or some of the active damping methods can be used [18]–[20].

Assuming that the impendances of the three-phase system are balanced the differential equations describing each phase of the PWM converter with LCL filter are:

$$L_{\mathrm{f}}\frac{\mathrm{d}\,i_i}{\mathrm{d}\,t} = u_i - i_i R_{\mathrm{f}} - u_{\mathrm{c}i} \quad (2)$$

$$C_{\mathrm{f}}\frac{\mathrm{d}\,u_{\mathrm{c}i}}{\mathrm{d}\,t} = i_i - i_{\mathrm{g}i} \quad (3)$$

$$L_{\mathrm{g}}\frac{\mathrm{d}\,i_{\mathrm{g}i}}{\mathrm{d}\,t} = u_{\mathrm{c}i} - i_{\mathrm{g}i} R_{\mathrm{g}} - u_{\mathrm{g}i} \quad (4)$$

$$u_i = s_i U_{dc} \quad (5)$$

where $i \in \{a, b, c\}$, i_i and u_i are the output current and voltage of the converter, $u_{\mathrm{c}i}$ is the voltage at the filter capacitor, $i_{\mathrm{g}i}$ and $u_{\mathrm{g}i}$ are the grid current and voltage, U_{dc} is the dc-link voltage, $s_i \in \{0, 1\}$ is the switching state of the phase leg of the converter ($S_{abc} = [s_a\ s_b\ s_c]$).

For the purpose of the control system design the influence of the capacitor C_f can be neglected since it only deals with the PWM switching frequencies. At frequencies lower than half of the resonance frequency of the LCL filter, where the control actions occur, the LCL-filter model and the L-filter models are practically the same [5]. Therefore, all the resistances of the filter can be replaced with $R = R_f + R_g$ and the inductance of the L-filter becomes $L = L_f + L_g$. A simplified model of the PWM converter with L filter, used for the control system design, is shown in Fig. 3.

A. The dq frame control

The dq control is using the transformation of the control variables from their natural abc frame to dq frame that rotates synchronously with the grid voltage. As a result of the transformation the control variables are becoming DC signals which enables the use of well known control methods and the use of PI controllers [21].

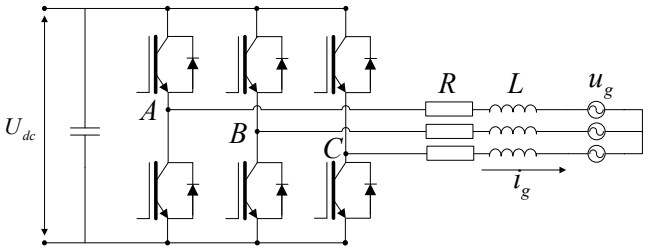


Fig. 3. PWM converter with L filter

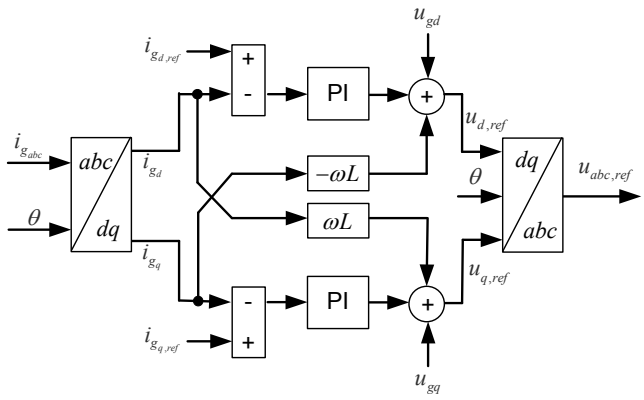


Fig. 4. Current control in the dq frame

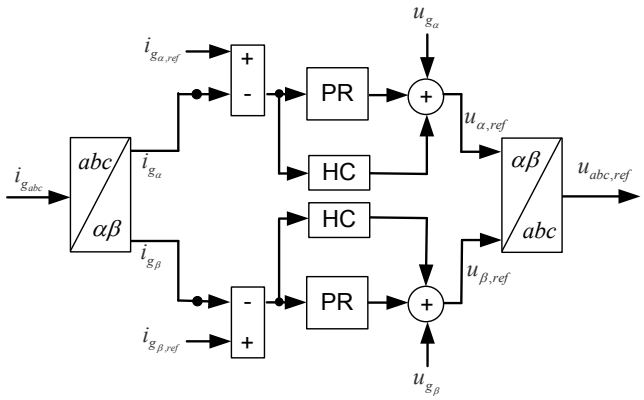


Fig. 5. Current control in the $\alpha\beta$ frame

A simplified model of the PWM converter with LCL filter in dq frame is given with the following equations:

$$L\frac{\mathrm{d}\,i_{gd}}{\mathrm{d}\,t} = u_d - u_{gd} - i_{gd}R + \omega L i_{gq} \quad (6)$$

$$L\frac{\mathrm{d}\,i_{gq}}{\mathrm{d}\,t} = u_q - u_{gq} - i_{gq}R - \omega L i_{gd} \quad (7)$$

Control strategy derived from this model is given in Fig. 4. The control structure involves cross coupling betveen d and q axes as well as a feedforward of a grid voltage. Another distinct feature of this control strategy is the necessity of the information about the phase angle of the grid voltage, θ. The phase angle of the grid voltage is obtained using phase-locked loop (PLL) [22]–[24].

B. The $\alpha\beta$ frame control

The $\alpha\beta$ control is using the $abc \rightarrow \alpha\beta$ transformation to transform the control variables from their natural abc frame to stationary reference frame. A simplified model of the PWM converter with L filter in $\alpha\beta$ frame is given with the following equations:

$$L\frac{\mathrm{d}\, i_{g\alpha}}{\mathrm{d}\, t} = u_\alpha - u_{g\alpha} - i_{g\alpha}R \quad (8)$$

$$L\frac{\mathrm{d}\, i_{g\beta}}{\mathrm{d}\, t} = u_\beta - u_{g\beta} - i_{g\beta}R. \quad (9)$$

Control structure in $\alpha\beta$ frame is shown in Fig. 5. Control variables in $\alpha\beta$ frame are time varying and therefore PI controllers can not be used to track the reference. Proportional-resonant (PR) controllers have become a standard solution for current control in grid-connected converters because of their ability to track sinusoidal reference [17], [25], [26]. Transfer function of the PR controller is given with:

$$G_{PR}(s) = K_p + K_i\frac{s}{s^2 + (\omega)^2}. \quad (10)$$

PR controller acts on a narrow band arround its resonant frequency and therefore the implementation of harmonic compensator for low-order harminics is possible without influencing the dynamics of the current controller. The transfer function of the harmonic compensator is given with:

$$G_{HC}(s) = \sum_{h=3,5,7} K_{ih}\frac{s}{s^2 + (\omega h)^2} \quad (11)$$

Advantages of the $\alpha\beta$ control over the dq control are the easy implementation of the harmonics compensation, no need for grid synchronization and no axes coupling. However, this control strategy is sensitive to grid frequency variations and the tuning of the PR controller is more difficult than the tuning of the PI controller. Often, tuning the parameters of the PR controller is done by the root locus method [7], [15].

V. Example

As an example a PR controller with harmonic compensation is tested. Model of the converter is made using PLECS [27] and the control algorithm is implemented in MATLAB/Simulink. The model consists of six IGBT switches with parallel diode, DC-link which is modelled as a constant DC voltage source, PWM modulator, LCL filter, grid transformer and the grid which is modelled as a three phase voltage source as shown in Fig. 6. Grid voltage is slightly distorted with 0.5% of 5^{th} and 7^{th} harmonics. Grid voltage and grid current are measured on the low voltage side of the grid transformer. Control structure consists of two PR controllers with harmonic compensation of the 5^{th} and 7^{th} harmonics in both axes. PR controllers and harmonic compensators are implemeted as discrete time controllers. Discrete time realization is obtained by using the impulse invariant discretization method as suggested in [28]. Grid current references calculated by the outer control loops are usually given as dq current components, therefore to transform the references to $\alpha\beta$ frame an information about the grid voltage phase is obtained using a simple synchronous reference frame PLL (SRF-PLL) [5], [29]. Reactive power control loop and DC-link voltage control loop are not simulated in this example, the reactive current reference is set to zero and the active current reference is changed once during the simulation (at $t_1 = 0.2$ s).

Simulation is conducted for real time period of 0.7s. Response to grid current reference change is observed as well as the system behaviour during two types of voltage sags. The simulation results are presented in Fig.7. As expected, PR controller with harmonic compensation has very good performance regarding low harmonic rejection since the THD in normal operating condition stays below 5%. Grid current reference change, which ocurrs after 0.2s in simulation, is accomplished in short period and the THD is quickly diminished to a satisfactory level. Fast transient response to the reference change is accomplished by good tuning of the PR controller parameters. It can be easily concluded that this type of control together with the proper control of the WT can comply to GC requirements in normal operating conditions.

During grid faults, which occur in 0.3s and 0.5s and last for 0.1s, this type of control approach has significant drawbacks. The SRF-PLL does not guarantee synchronisation with the grid voltage during unbalanced grid faults. Therefore, the $\alpha\beta$ grid current references are not properly calculated and the entire control system is not synchronized with the grid. Instead of providing voltage support to the grid during faults, additional disturbance is passed on to the grid by injecting distorted current. In such conditions WT would be forced to disconnect from the grid.

VI. Conclusion

The ever-increasing integration of wind power plants, which should fullfil the requirements imposed by the grid operator regarding fault ride-through and grid support during grid faults (RCI), has encouraged engineers and researchers to improve the existing control solutions for grid-connected converters.

The ongoing research indicates that the current control is a critical part of the converter control system. Two main tasks within current control are the robustness against grid disturbances and the dynamic behaviour of the controlled system. Control strategies that were previously developed for controlling voltage source converters in applications like UPS or motor drives, can not be directly used for controlling the grid-side converter of the wind turbine or other renewable energy sources. The drawbacks of such control strategies are illustrated in a simulation example that was presented in the paper.

Research studies are now focused on improving the synchronization algorithms using different types of phase-locked loops. Many researches are done on optimization of the proportional-resonant controllers. However, optimal control algorithms have also shown to have great potential in controlling grid-connected converters.

Acknowledgement

This research has been supported by the Croatian Science Foundation under grant No. I-4463-2011 (MICROGRID) and

by the Ministry of Science, Education and Sports of the Republic of Croatia under grant No. 036-0361621-3012.

REFERENCES

[1] *Global Wind Report 2011*, http://www.gwec.net, GWEC - Global Wind Energy Conucil, 2011.

[2] M. Altin, O. Go¨ksu, R. Teodorescu, P. Rodriguez, B.-B. Jensen, and L. Helle, "Overview of recent grid codes for wind power integration," in *Optimization of Electrical and Electronic Equipment (OPTIM), 2010 12th International Conference on*, may 2010, pp. 1152 –1160.

[3] M. Tsili and S. Papathanassiou, "A review of the grid code technical requirements for wind farms," *IET Renewable Power Generation*, vol. 3, pp. 308–332, 2009.

[4] F. Blaabjerg and Z. Chen, *Power Electronics for Modern Wind Turbines*. Morgan & Claypool, 2006.

[5] R. Teodorescu, M. Liserre, and P. Rodríguez, *Grid Converters for Photovoltaic and Wind Power Systems*. John Wiley & Sons, 2011.

[6] *European grid code requirements for wind power generation*, EWEA Working Group on Grid Code Requirements - Position Paper, Brussels, 2008.

[7] A. Timbus, M. Liserre, R. Teodorescu, P. Rodriguez, and F. Blaabjerg, "Evaluation of current controllers for distributed power generation systems," *IEEE Transactions on Power Electronics*, vol. 24, pp. 654–664, 2009.

[8] L. Villate and H. Camblong, "Connection requirements for wind farms: A survey on technical requierements and regulation," *Renewable and Sustainable Energy Reviews*, vol. 11, no. 8, pp. 1858 – 1872, 2007.

[9] P. Rodriguez, A. Timbus, R. Teodorescu, M. Liserre, and F. Blaabjerg, "Reactive power control for improving wind turbine system behavior under grid faults," *IEEE Transactions on Power Electronics*, vol. 24, pp. 1798–1801, 2009.

[10] *IEEE Standard for Interconnecting Distributed Resources with Electric Power Systems*, IEEE Standard 1547, 2003.

[11] M. Bollen, "Characterisation of voltage sags experienced by three-phase adjustable-speed drives," *Power Delivery, IEEE Transactions on*, vol. 12, no. 4, pp. 1666 –1671, oct 1997.

[12] N. Mohan, T. Undeland, and W.P.Robbins, *Power Electronics - Converters, Applications and Design*. John Wiley & Sons, 1995.

[13] H. W. van der Broeck, H. C. Skudelny, and G. Stanke, "Analysis and realization of a pulse width modulator based on voltage source space vectors," *IEEE Transactions on Industrial Applications*, vol. 24, pp. 142–150, 1988.

[14] J.Holtz, "Pulsewidth modulation - a survey," *IEEE Transactions on Industrial Electronics*, vol. 39, pp. 410–420, 1992.

[15] A. Vidal, F. Freijedo, A. Yepes, P. Fernandez-Comesana, J. Malvar, O. Lopez, and J. Doval-Gandoy, "Assessment and optimization of the transient response of proportional-resonant current controllers for distributed power generation systems," *Industrial Electronics, IEEE Transactions on*, vol. 60, no. 4, pp. 1367 –1383, april 2013.

[16] M. Liserre, R. Teodorescu, and F. Blaabjerg, "Multiple harmonics control for three-phase grid converter systems with the use of pi-res current controller in a rotating frame," *Power Electronics, IEEE Transactions on*, vol. 21, no. 3, pp. 836 –841, may 2006.

[17] X. Yuan, W. Merk, H. Stemmler, and J. Allmeling, "Stationary-frame generalized integrators for current harmonics of concern under unbalanced and distorted operating conditions," *IEEE Transactions on Industrial Applications*, vol. 38, pp. 523–532, 2002.

[18] E. Wu and P. Lehn, "Digital current control of a voltage source converter with active damping of lcl resonance," *IEEE Transactions on Power Electronics*, vol. 21, pp. 1364–1373, 2006.

[19] M. Liserre, A. Dell'Aquila, and F. Blaabjerg, "Genetic algorithm-based design of the active damping for an lcl-filter three-phase active rectifier," *Power Electronics, IEEE Transactions on*, vol. 19, no. 1, pp. 76 – 86, jan. 2004.

[20] J. Dannehl, F. Fuchs, and P. Thogersen, "Pi state space current control of grid-connected pwm converters with lcl filters," *Power Electronics, IEEE Transactions on*, vol. 25, no. 9, pp. 2320 –2330, sept. 2010.

[21] V. Blasko, J. C. Moreira, and T. A. Lipo, "A new field oriented controller utilizing spatial position measurement of rotor end ring current," in *PESC*, 1989, pp. 295–299.

[22] S. Chung, "A phase tracking system for three phase utility interface inverters," *IEEE Transactions on Power Electronics*, vol. 15, no. 3, pp. 431–438, May 2000.

[23] P. Rodriguez, J. Pou, J. Bergas, J. Candela, R. Burgos, and D. Boroyevich, "Decoupled double synchronous reference frame pll for power converters control," *IEEE Transactions on Power Electronics*, vol. 22, no. 2, pp. 584–592, March 2007.

[24] A. V. Timbus, M. Liserre, R. Teodorescu, and F. Blaabjerg, "Synchronization methods for three phase distributed power generation systems. an overview and evaluation," in *PESC*, 2005, p. 2474–2481.

[25] S. Fukunda and T. Yoda, "A novel current-tracking method for active filters based on sinusoidal internal model," *IEEE Transactions on Industrial Electronics*, vol. 37, pp. 888–895, 2001.

[26] Y. Sato, T. Ishizuka, K. Nezu, and T. Kataoka, "A new control strategy fot voltage-type pwm rectifiers to realize zero steady-state control error in input current," *IEEE Transactions on Industrial Applications*, vol. 34, pp. 480–486, 1998.

[27] *PLECS User Manual*, Plexim GmbH, 2011.

[28] A. Yepes, F. Freijedo, J. Doval-Gandoy, O. Lo´pez, J. Malvar, and P. Fernandez-Comesana, "Effects of discretization methods on the performance of resonant controllers," *Power Electronics, IEEE Transactions on*, vol. 25, no. 7, pp. 1692 –1712, july 2010.

[29] V. Kaura and V. Blasko, "Operation of a phase locked loop system under distorted utility conditions," *Industry Applications, IEEE Transactions on*, vol. 33, no. 1, pp. 58 –63, jan/feb 1997.

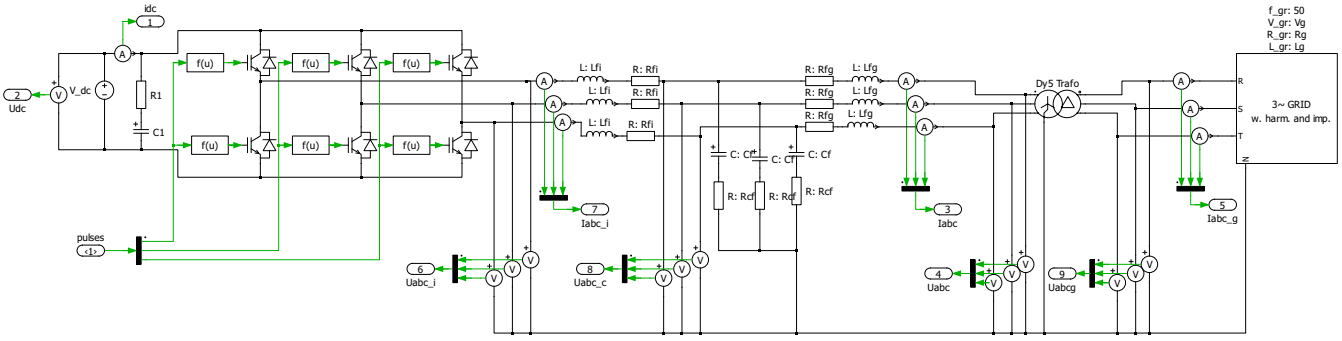


Fig. 6. Model of the grid-connected converter with LCL filter in PLECS

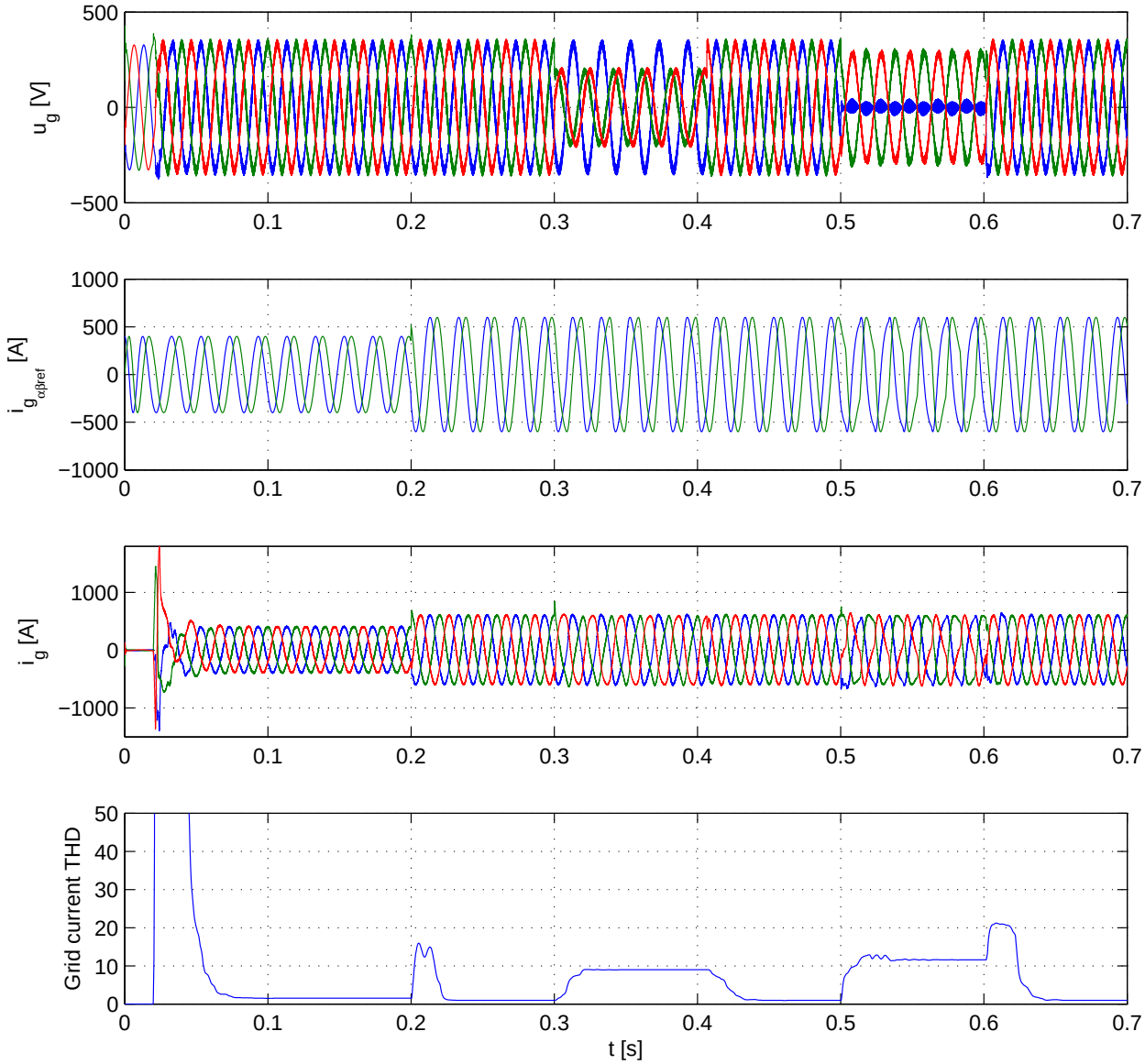


Fig. 7. Simulation example showing grid voltage, grid current reference in $\alpha\beta$ frane, grid current and total harmonic distortion of the grid current. Current reference has changed in $t_1 = 0.2$s. Two voltage faults occur at $t_2 = 0.3$ s and $t_3 = 0.5$ s.

A Leader-Follower Approach to Formation Control of Multiple Non-Holonomic Mobile Robots

Toni Petrinić* and Ivan Petrović**
* HEP-ODS d.o.o., Elektroprimorje Rijeka, Pogon Cres-Lošinj, Dražica 10, 51550 Mali Lošinj, Croatia
toni.petrinic@gmail.com
** University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Control and Computer Engineering, Unska 3, 10000 Zagreb, Croatia
ivan.petrovic@fer.hr

Abstract - Many cooperative tasks in real world environments need the robots to maintain some desired formations when moving. Formation control refers to the problem of controlling the relative position and orientation of robots in a group, while allowing the group to move as a whole. In this paper, a novel leader-follower formation control law for multiple non-holonomic mobile robots based on kinematic models and trajectory tracking techniques is proposed.

I. INTRODUCTION

Research interest in formation and cooperative control of robotic systems has recently grown enormously in the control community. The usage of groups of robots has several advantages compared to a single robot. For instance, the tasks that can be accomplished are inherently more complex than those that can be accomplished by a single robot. Also, the system becomes more flexible (since robots can have a variety of roles, the same group of robots can be employed for many different objectives) and robustness (system can be designed so that a robot can take over the tasks of another robot in case of failure). The range of applications includes: (i) dangerous tasks on locations where it is harmful for humans to work, e.g. areas of toxic contamination, nuclear contamination or forest fires; (ii) tasks that cannot be performed by a single robot, e.g. transporting or repositioning of large objects; (iii) tasks that scale up or down in time, e.g. certain tasks may require a decreasing or increasing amount of robots as the operation proceeds over time; (iv) exploration, searching and rescue tasks.

In the literature, three different control approaches for mobile robot formation are described: behaviour-based approach [1-6], virtual-structure approach [7-8] and leader–follower approach [9-16].

In the behaviour-based approach, the selection of robots behaviour (e.g. obstacle avoidance, target seeking) is based on some rules and each behaviour has its specific purpose or task. The advantage of this approach is that it is natural to derive control strategies when robots have multiple competing objectives and that the robots are controlled in a decentralized mode. Therefore, it is suitable for a large number of robots. A disadvantage is complex mathematical analysis and it is hard to guarantee exact formation control and stability of the system as a whole.

In the leader-follower approach, one of the robots is assigned as the leader and others are followers. The leader follows a predefined trajectory, while the followers are keeping the position and direction with a certain distance to the leader. The advantage of this approach is that it is easy to understand and implement. However, this approach has some disadvantages. It asks for a centralized control strategy (followers use the position of the leader robot as a control input), which makes it less suitable for a large number of robots. There is no explicit feedback from the followers to the leaders. For example, if the leader moves too fast, or the follower is blocked by an obstacle, the formation will be destroyed. Another disadvantage is that, if the leader has failed, the entire formation can not be kept and the consequence will be very serious. One method which is often used to maintain the formation is to define an internal shape variable of the follower robot according to the relative angle and relative distance from its leader or defining the relative distance from two of its leaders. However, the leader-follower approach will not maintain its formation if the followers are perturbed, which results in a limited robustness with respect to the formation.

In the virtual-structure approach the robot formation is considered as a virtual rigid structure (e.g. a circle, square, etc). The advantage of this approach is that the formation is easy to describe. The main difference between the leader-follower approach and the virtual structure approach is that for the virtual structure approach extra parameters that specify the formation, so called mutual parameters, are added to the controllers. If these parameters are added to the control feedback, the structure is more robust, even if some robots are perturbed. A disadvantage is that it is often more difficult to determine stability than for the leader-follower approach.

This research has been supported by the European Community's Seventh Framework Programme under grant No. 285939 (ACROSS).

In this paper, we investigate leader-following problem by assigning a virtual leader robot for each robot to guarantee the formation stability. Virtual formation constitutes of virtual leader and virtual followers. The positions of virtual followers are defined as offsets from the origin of the virtual leader coordinates in the directions of robot local coordinate system.

The rest of the paper is organized as follows. First, we describe the model of unicycle robot in section 2. Then, the main results with design procedures and stability analysis are presented in section 3. In section 4, simulation results are given. Finally, the paper ends with concluding remarks in section 5.

II. ROBOT MODEL

In this paper, we deal with a group of unicycle robots. Mathematically the kinematic model of a unicycle robot is described by the following equations:

$$\dot{x} = v\cos\theta, \quad \dot{y} = v\sin\theta, \quad \dot{\theta} = \omega, \tag{1}$$

where v, ω denote the translational and rotational velocities which are assumed to be the control input of the system. Mobile robots of this type are subject to non-holonomic constraints [18]:

$$\dot{x}\sin\theta - \dot{y}\cos\theta = 0. \tag{2}$$

It is important to mention that dynamics of a robot is not modelled by the kinematic model. Therefore, it is assumed that robot exactly realizes velocity commands v and ω. Of course, due to its body and actuator dynamics, and also non-idealities such as friction, gear backslash, wheel slippage, actuator dead zone and saturation, a robot cannot exactly realize velocity commands. Nevertheless, as we have assumed that the command trajectory is feasible for the robot, robot can approximately track the reference velocity commands.

In simulations we use the robot model for robot soccer [19]. The robot model is developed in MATLAB® Simulink and simulations are done in MATLAB® Simulink also.

Let $v_{\max}$ and $\omega_{\max}$ are the absolute maximum linear and angular velocities of the robot, respectively. As a safety measure, in case when the formation tracking controller temporarily generates command velocities higher than robot limitations, velocity saturation will occur. In order to preserve the curvature radius originated from v and ω, a velocity scaling is needed as follows. If the scaled down values of linear and angular velocities are v_s and ω_s respectively, we have [17]:

$$\begin{aligned} \delta &= \max\{|v|/v_{\max}, |\omega|/\omega_{\max}, 1\} \\ v_s &= sign(v)v_{\max}, \omega_s = \omega/\delta \quad if \quad \delta == |v|/v_{\max} \\ v_s &= v/\delta, \omega_s = sign(\omega)\omega_{\max} \quad if \quad \delta == |\omega|/\omega_{\max} \\ v_s &= v, \omega_s = \omega \quad if \quad \delta == 1 \end{aligned}. \tag{3}$$

III. VIRTUAL ROBOT TRACKING BASED FORMATION CONTROLLER

In the leader-follower based formation control, the desired poses of followers can be thought of as offsets to the origin of the leader robot coordinate system at any given time. These offsets mimic virtual robots which the actual followers must track. Note that these offsets may be time invariant or time variant. In this paper, for simplicity, we assume that offset is time invariant. Tracking paths and velocities of the virtual robot require a combination of a nominal feed-forward command, which calculates the pose of the virtual robot and it's linear and angular velocities to which the actual designated follower must reach to, with a feedback action on the error.

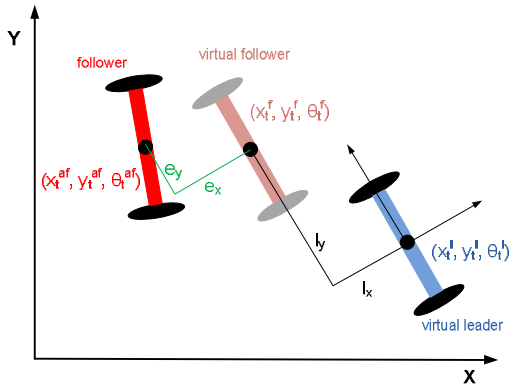


Figure 1. Leader-follower based formation control

Assuming that the virtual leader robot's pose at time t is $(x_t^l, y_t^l, \theta_t^l)$ and linear and angular velocities are (v_t^l, ω_t^l). The desired position for a virtual follower (x_t^f, y_t^f) can be described as offsets l_x and l_y from the origin of the virtual leader robot coordinates in the directions of robot X and robot Y coordinates, respectively:

$$\begin{bmatrix} x_t^f \\ y_t^f \end{bmatrix} = \begin{bmatrix} \cos\theta_t^l & -\sin\theta_t^l \\ \sin\theta_t^l & \cos\theta_t^l \end{bmatrix} \begin{bmatrix} l_x \\ l_y \end{bmatrix} + \begin{bmatrix} x_t^l \\ y_t^l \end{bmatrix}. \tag{4}$$

The derivatives of desired virtual follower pose variables are:

$$\begin{bmatrix} \dot{x}_t^f \\ \dot{y}_t^f \\ \dot{\theta}_t^f \end{bmatrix} = \begin{bmatrix} \cos\theta_t^l & -l_x\sin\theta_t^l - l_y\cos\theta_t^l \\ \sin\theta_t^l & l_x\sin\theta_t^l - l_y\cos\theta_t^l \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_t^l \\ \omega_t^l \end{bmatrix}. \tag{5}$$

The desired linear and angular velocities of virtual follower are given by:

$$v_t^f = \pm\sqrt{(\dot{x}_t^f)^2 + (\dot{y}_t^f)^2}, \tag{6}$$

$$\omega_t^f = \frac{\ddot{y}_t^f \dot{x}_t^f - \ddot{x}_t^f \dot{y}_t^f}{(\dot{x}_t^f)^2 + (\dot{y}_t^f)^2}. \tag{7}$$

The ω_t^f is derived through defining θ_t^f as:

$$\theta_t^f = a\tan 2(\dot{y}_t^f, \dot{x}_t^f) + k\pi, \quad k = 0,1, \tag{8}$$

where $k=0$ is for forward motion and $k=1$ is for backward motion. Note if l_x and l_y are constants and $(\dot{v}_t^l, \dot{\omega}_t^l) = (0,0)$, than angular velocities of the virtual leader and the virtual follower have the same value $\omega_t^f = \omega_t^l$.

The problem of minimizing difference between virtual robot, i.e. reference configuration, and actual designated follower robot configuration is equivalent to the problem of stabilization of tracking error dynamics. The tracking error $e = \left[e_x, e_y, e_\theta\right]^T$ can be defined in the configuration space of the robot, however it is convenient to transform it to the robot local coordinate system. This transformation is performed as follows

$$\begin{bmatrix} e_x \\ e_y \\ e_\theta \end{bmatrix} = \begin{bmatrix} \cos\theta_t^{af} & \sin\theta_t^{af} & 0 \\ -\sin\theta_t^{af} & \cos\theta_t^{af} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_t^f - x_t^{af} \\ y_t^f - y_t^{af} \\ \theta_t^f - \theta_t^{af} \end{bmatrix}, \tag{9}$$

where error component e_x is the longitudinal error, e_y is the orthogonal error, while e_θ is the orientation error as illustrated in Figure 1 and $(x_t^{af}, y_t^{af}, \theta_t^{af})$ is the actual follower pose. By taking a derivative of the above equation and taking into account robot kinematics (1) and (2), the error dynamics become [20]

$$\begin{aligned} \dot{e}_x &= \omega_t^{af} e_y - v_t^{af} + v_t^f \cos e_\theta \\ \dot{e}_y &= -\omega_t^{af} e_x + v_t^f \sin e_\theta \\ \dot{e}_\theta &= \omega_t^f - \omega_t^{af} \end{aligned} . \tag{10}$$

To make the formation error (10) converge to zero, we first stabilize e_x with v_t^{af}. Because e_y cannot be directly controlled, we choose appropriate ω_t^{af} to decrease the effect of $\omega_t^{af} e_y$ and $\omega_t^{af} e_x$.

If we choose

$$v_t^{af} = k_x e_x + v_t^f \cos e_\theta , \tag{11}$$

the close-loop dynamics of $\dot{e}_x$ become

$$\dot{e}_x = -k_x e_x + \omega_t^{af} e_y , \tag{12}$$

where k_x is a positive constant. From (12) we can conclude that, if e_y is zero, then e_x can exponentially converge to zero.

To determine the control input ω_t^{af}, we chose a Lyaponuv function as

$$V(t) = \frac{e_x^2}{2} + \frac{e_y^2}{2} + \frac{1-\cos e_\theta}{k_y} , \tag{13}$$

with k_y as a positive constant.

The time derivative of Lyaponuv function (13) with (10) and (11) is

$$\dot{V}(t) = -k_x e_x^2 + \sin e_\theta \left[v_t^f e_y + \frac{\omega_t^f - \omega_t^{af}}{k_y} \right]. \tag{14}$$

To make the derivative of the Lyapunov function (14) negative, we choose

$$\omega_t^{af} = \omega_t^f + v_t^f k_y e_y + k_\theta k_y \sin e_\theta . \tag{15}$$

where k_θ is a positive constant.

Finally, from (14) and (15) we obtain

$$\dot{V}(t) = -k_x e_x^2 - k_\theta \sin^2 e_\theta \le 0 . \tag{16}$$

Now we analyse the stability of the formation tracking control law.

Theorem 1: Under the control law (11) and (15), the equilibrium of system (9) is locally stable.

Proof: $V(t)$ is positive definite and $\dot{V}(t)$ with (11) and (15) is negative semi-definite around the origin. Therefore, $V(t)$ is a Lyaponuv function, which indicates the local stability [21]

Remark 1: In [20], the local stability is solved under condition $v_t^f > 0$. Note that in our case (16) does not depend on v_t^f.

Theorem 2: Assume that (i) v_t^f and ω_t^f are continuous, (ii) v_t^f, ω_t^f, k_x, k_y, k_θ are bounded, (iii) k_x, k_y, k_θ are positive constants and (iv) $\dot{v}_t^f$ and $\dot{\omega}_t^f$ are sufficiently small. Under these condition, $e=0$ is uniformly asymptotically stable over $[0,\infty)$.

Proof: We first linearize the nonlinear system (10) around the origin:

$$\dot{e} = Ae = \begin{bmatrix} k_x & \omega_t^f & 0 \\ -\omega_t^f & 0 & v_t^f \\ 0 & -k_y v_t^f & -k_\theta k_y \end{bmatrix} e . \tag{17}$$

The characteristic equation for A is:

$$a_3 s^3 + a_2 s^2 + a_1 s + a_0 = 0 , \tag{18}$$

where

$$\begin{aligned} a_0 &= 1 \\ a_1 &= k_\theta k_y + k_x \\ a_2 &= (\omega_t^f)^2 + k_y (v_t^f)^2 + k_x k_y k_\theta \\ a_3 &= k_x k_y (v_t^f)^2 + k_\theta k_y (\omega_t^f)^2 \end{aligned} . \tag{19}$$

Since all coefficients of (18) are positive and $a_1 a_2 > a_0 a_3$, the real parts of all roots are negative according the Routh criterion [22]. Therefore, the theorem was proved.

Let the actual pose of leader robot be $(x_t^{al}, y_t^{al}, \theta_t^{al})$ and its velocities $(v_t^{al}, \omega_t^{al})$. The leader robot tracks

virtual leader robot with zero offsets $(l_x, l_y) = (0,0)$. The design of control law for the leader is the same as for the follower robot. In equations (9)-(19) parameters $x_t^f, y_t^f, \theta_t^f, x_t^{af}, y_t^{af}, \theta_t^{af}, v_t^{af}, \omega_t^{af}$ should be replaced with $x_t^l, y_t^l, \theta_t^l, x_t^{al}, y_t^{al}, \theta_t^{al}, v_t^{al}, \omega_t^{al}$.

IV. Simulation Results

For simulation purposes an 8-shaped trajectory was used as is illustrated in Figure 2. The group consists of three robots, one leader robot and two followers. Follower 1 is in parallel formation with the leader while follower 2 is in serial formation with the leader. Since we describe formation in Cartesian coordinates with respect to the leader robot (4), we will show in simulations that our control law is capable of dealing with two mentioned formation. In [10] formation is described in polar coordinates with respect to the leader robot and it is proposed two control laws, one for parallel formation and one for serial formation.

In simulation, we choose following values of constants for our control law: $k_x = 1$, $k_y = 1$ and $k_\theta = 10$. Leader robot and virtual leader robot initial poses were $[x_t(0), y_t(0), \theta_t(0)] = [0,\ 0,\ \pi/2]$. Initial poses of follower 1 and 2 were set to the values $[x_t^f(0), y_t^f(0), \theta_t^f(0)] = [-0.1,\ -0.05,\ \pi/2]$ and $[x_t^f(0), y_t^f(0), \theta_t^f(0)] = [0,\ 0.1,\ \pi/3]$, respectively. Desired formation for the follower 1 was $(l_x, l_y) = (0,\ 0.05)$ while for the follower 2 was $(l_x, l_y) = (-0.02,\ -0.05)$. Initial velocities for all robots were zero. Reference linear and angular velocities of virtual leader for left and right part of 8-shaped trajectory were set to the values $(v_t^l, \omega_t^l) = (0.1,\ 0.4)$ and $(v_t^l, \omega_t^l) = (0.1,\ -0.4)$, respectively. Note that linear velocities of virtual followers v_t^f (see Figure 6 and 8) depend on desired formation, i.e. offset.

Figure 3 presents how the leader robot tracks the virtual robot. Since all initial tracking errors were set to zero, leader robot quickly reaches reference values of linear and angular velocities (Figure 4). Initial position error of follower 1 which can be seen on Figure 5, causes slow reaching of reference velocities value (Figure 6). An initial pose (position and orientation) error of follower 2 causes a big value of robots angular velocity on the beginning of fetching desired formation (Figure 8).

V. Conclusion

This paper proposes a stable leader-follower formation control law based on kinematic model and trajectory tracking technique. Offset that define formation can be time invariant or time variant. The proposed control law may be used for pure trajectory tracking (with and without initial tracking error). Stability of the control law is proved through the use of a Lyaponuv function. Simulation results confirm theoretical results.

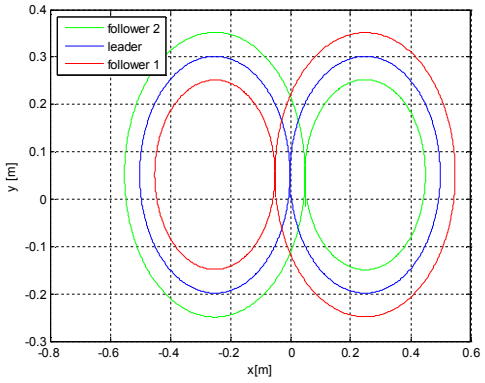


Figure 2. The paths of virtual leader and virtual followers

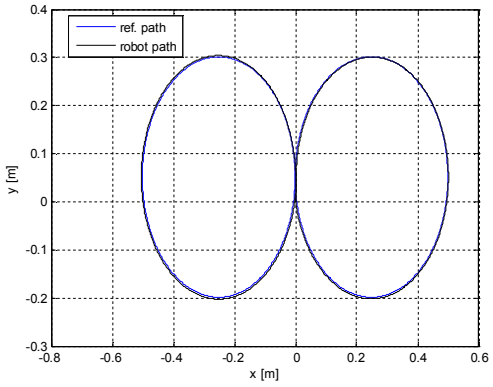


Figure 3. Formation control of the leader robot

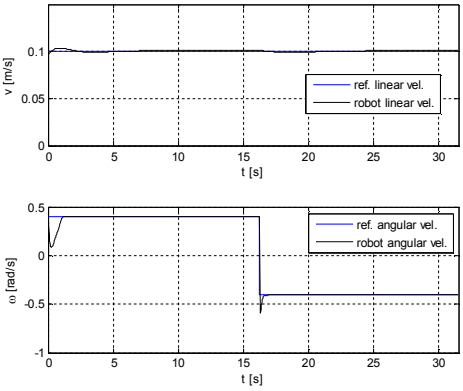


Figure 4. Linear and angular velocities of the leader robot

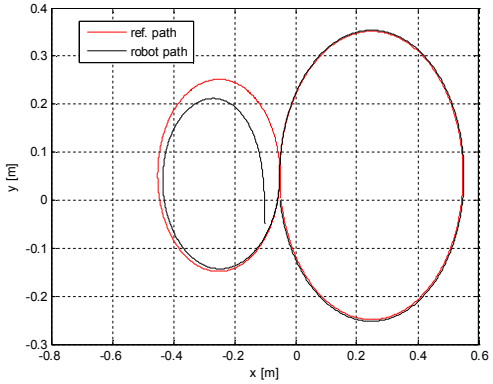


Figure 5. Formation control of the follower 1

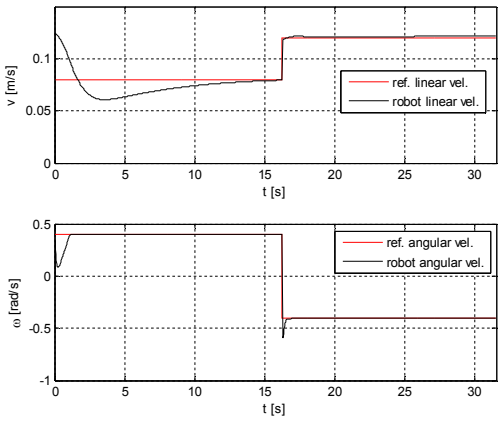


Figure 6. Linear and angular velocities of the follower 1

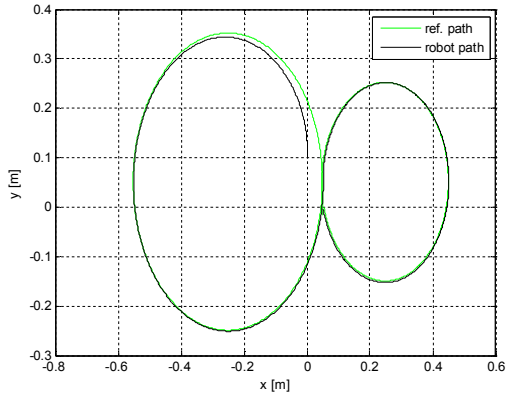


Figure 7. Formation control of the follower 2

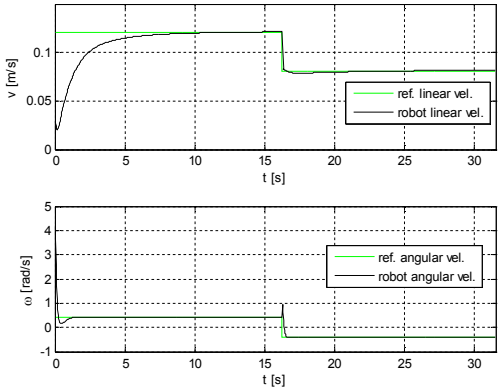


Figure 8. Linear and angular velocities of the follower 2

REFERENCES

[1] D. W. Casbeer, D. B. Kingston, R. W. Beard, T. W. McLain, S.-M. Li, and R. Mehra, "Cooperative forest fire surveillance using a team of small unmanned air vehicles," International Journal of Systems Sciences, pp. 351–360, May 2006.

[2] J. Cortes, S. Martinez, T. Karatas, and F. Bullo, "Coverage control for mobile sensing networks," IEEE Transactions on Robotics and Automation, 2002.

[3] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbour rules," IEEE Transactions on Automatic Control, pp. 988–1001, 2003.

[4] T. Balch and R. C. Arkin, "Behaviour-based formation control for multirobot teams," IEEE Transactions on Robotics and Automation, Vol. 14, pp. 926-939, 1998.

[5] M. J. Matarić, "Behaviour-based control: examples from navigation, learning, and group behaviour," Journal of Experimental & Theoretical Artificial Intelligence, Vol. 9, pp. 323–336, 1997.

[6] M. J. Matarić, "Learning in behaviour-based multi-robot systems: policies, models, and other agents," Journal of Cognitive Systems Research, Vol. 2, pp. 81-93, 2001.

[7] T. H. A. van den Broek, N. van de Wouw, and H. Nijmeijer, "Formation Control of Unicycle Mobile Robots: a Virtual Structure Approach," Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference, pp. 8328-8333, Dec. 2009.

[8] A. Sadowska, H. Huijberts, D. Kostić, N. van de Wouw, and H. Nijmeijer, "Formation Control of Unicycle Robots Using the Virtual Structure Approach," The 15th International Conference on Advanced Robotics (ICAR), pp. 365-370, June 2011.

[9] A. S. Brandao, M. Sarcinelli-Filho, R. Carelli, and T. F. Bastos-Filho, "Decentralized control of leader-follower formations of mobile robots with obstacle avoidance," IEEE International Conference on Mechatronics (ICM), pp. 1-6, April 2009.

[10] T. Gustavi, Control and Coordination of Mobile Multi-Agent Systems, doctoral thesis, Royal Institute of Technology, Sweden, 2009.

[11] J. P. Desai, J. Ostrowski, and V. Kumar, "Controlling formations of multiple mobile robots," Proceedings of IEEE International Conference on Robotics and Automation, Vol. 4, pp. 2864-2869, May 1998.

[12] J. Shao, G. Xie, and L. Wang, "A tracking controller for motion coordination of multiple mobile robots," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 783-788, Aug. 2005.

[13] V.-H. Tran and S.-G. Lee, "A Stable Formation Control Using Approximation of Translational and Angular Accelerations," International Journal of Advanced Robotic Systems, Vol. 8, No. 1, pp. 65-75, 2011.

[14] J. Wu and Z. Jiang, "Formation control of multiple mobile robots via switching strategy," International Journal of Information and Systems Sciences, Vol. 5, No. 2, pp. 210-218, 2009.

[15] S.-C. Liu, Da-L. Tan, and G.-J. Liu, "Robust Leader-follower Formation Control of Mobile Robots Based on a Second Order Kinematics Model," Acta Automatica Sinica, Vol. 33, No. 9, pp. 947-955, September 2007.

[16] D. Miklić, S. Bogdan, and R. Fierro, "AVLib: A Simulink® Library for Multi-Agent Systems Research," 20th Mediterranean Conference on Control & Automation, 2012.

[17] M. Brezak, I. Petrović, and N. Perić, "Experimental Comparison of Trajectory Tracking Algorithms for Nonholonomic Mobile Robots," Proceedings of 35th Annual Conference of the IEEE Industrial Electronics Society, pp. 473-478, 2009.

[18] T. Petrinić, I. Petrović. Control of nonholonomic differential drive mobile robots (in Croatian). Proceedings of the 33th international convention MIPRO, Vol. III. - CTS & CI, pages108-113., 2010.

[19] G. Klančar, M. Brezak, I. Petrović, D. Matko. Two approaches to mobile robots simulator design. Proceedings of the 6th EUROSIM Congress on Modelling and Simulation, pages 1-9, 2007.

[20] Y. Kanayama, Y. Kimura, F. Miyazaki, and T. Noguchi, "A stable tracking control method for an autonomous mobile robot," Proceedings of IEEE International Conference Robotics and Automation, pages 384-389, 1990.

[21] J. Kasać. Stability analysis of nonlinear systems controlled by analytic fuzzy controller (in Croatian), doctoral thesis, FSB, Zagreb, 2005.

[22] K. Ogata, Modern Control Engineering, 4th ed., Prentice Hall, 2002.

One Variant of Self-tuning DC Motor Adaptive Controller

Ž. Jurić *, H. Šiljak **
* Faculty of Electrical Engineering/Department of Automatic Control & Electronics, Sarajevo, Bosnia & Herzegovina
** Burch International University/Department of Electrical & Electronics Engineering, Sarajevo, Bosnia & Herzegovina
zjuric@utic.net.ba; hsiljak@ibu.edu.ba

Abstract – This paper proposes a self-tuning adaptive controller in order to produce response of a second order block from the typically nonlinear motor. Periodic reset of controller's estimator block is employed as an improvement of the original design. Classical model of DC motor is well-controlled via this improved controller, while for the system containing a motor and thyristor-fired phase-controlled rectifier a new improvement is introduced, the arcsine compensation block. Controller is tested in different simulation conditions and it shows high ability to compensate any effects, produced either by the reference value change or the external error induced by load or thyristor rectifier.

I. INTRODUCTION

The principles of adaptive control through self-tuning regulators are well known and widely applied for a long time already [1]. Since the object of control very often is a DC motor, this topic deserves special attention, paying attention to different modeling variants and different issues arising in the process of control.

These issues often require improvements of the original adaptive control algorithm in order to meet the performance requirements. One of such improvements addressed in this paper is controller reset, procedure already known in theory and practice [2]. Another one is related with the wind-up phenomenon, which may cause serious problems in achieving the desired response [3], so it has been studied thoroughly, both in case of wind-up caused by the object of control [4] and the controller itself [5].

If we focus our attention at the model of object being controlled – the DC drive, one may notice that the original dynamical model of DC machine is usually augmented with additional equations (i. e. components), such as thyristor rectifiers of various sorts [6]. Since these components inherently bring in nonlinearity in the system, in order to achieve good performance nonlinearity has to be compensated through complementary nonlinear blocks within the controller [7].

This paper intends to give a systematic approach to DC drive adaptive control development based on the improvements listed in this section. While they are relatively known in literature, their synergy and details of application in this particular class of control objects represents the major contribution of this paper.

II. A BASIC DESIGN

The object of the control is a standard DC motor described with the following set of equations:

$$\begin{aligned} L_s\, di_s/dt + R_s\, i_s &= u_s \\ L_a\, di_a/dt + R_a\, i_a + L_s\, i_s\, \omega &= u_a \\ J\, d\omega/dt + B\, \omega + M_l &= L_s\, i_s\, i_a \end{aligned} \qquad (1)$$

This is a 3rd-order nonlinear model with stator current i_s, armature current i_a and angular speed ω as the state variables. The load torque M_l acts as a disturbance. The aim of the control is to keep the angular speed on the given value, and to allow good transient response when it is changed, but assuming that the controller has no a priori information about the parameters of the motor. The basic structure of the control loop is shown in Fig. 1.

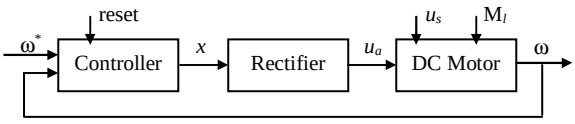


Figure 1. Basic structure of the control loop

The control variable is the armature voltage u_a, and it is generated as the output of the controlled rectifier (where x is the parameter which determines the output voltage). If the stator voltage u_s is constant (i. e. when $u_s = U_s = \text{const.}$), the current i_s becomes constant after the initial transients vanishes, so we get the linear model

$$\begin{aligned} J\, d\omega/dt + B\, \omega + M_l &= G\, i_a \\ L_a\, di_a/dt + R_a\, i_a + G\, \omega &= u_a \end{aligned} \qquad (2)$$

where $G = L_s\, U_s / R_s$. The design will first be based on such simplified model, and later the consequences of such simplification will be explored. Equations (2) may be solved for ω. In the operator p-domain, we get

$$\omega(t) = H(p)\, [u_a(t) - (L_a\, p + R_a)\, M_l(t) / G] \qquad (3)$$

Here, $p = d/dt$, and $H(p)$ is the transfer function

$$H(p) = K / (p^2 + a\, p + b) \qquad (4)$$

where parameters a, b and K are given by

$$a = \mathrm{R}_a/\mathrm{L}_a + \mathrm{B}/\mathrm{J},\ b = (\mathrm{B}\,\mathrm{R}_a + \mathrm{G}^2)/(\mathrm{J}\,\mathrm{L}_a),\ \mathrm{K} = \mathrm{G}/(\mathrm{J}\,\mathrm{L}_a) \quad (5)$$

The principal structure of the controller is given in Fig. 2.

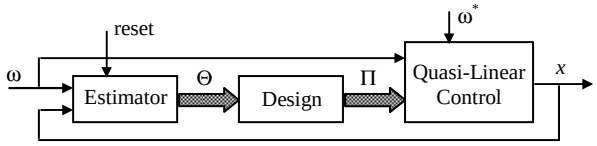


Figure 2. Principal structure of the controller

Here, $x(t)$ is the output of the controller. The estimator estimates the unknown parameter vector $\Theta = \{a, b, \mathrm{K}\}$. The estimation is based on the observation of the input and the output of the motor (together with the rectifier). The block "Design" uses the estimated Θ to calculate the vector Π of parameters that are necessary to obtain the wanted response. The last block is an ordinary linear controller with externally tunable parameters (that makes the controller nonlinear in nature) together with (also nonlinear) anti-windup compensation. The following text gives the more detailed explanation of all of these blocks.

The estimator is based on weighted Gaussian least squares approach. At first, we will assume that there is no load torque, i. e. $\mathrm{M}_l = 0$. Let us take $\omega_f(t) = \mathrm{H}_f(p)\,\omega(t)$ and $x_f(t) = \mathrm{H}_f(p)\,x(t)$ where $\mathrm{H}_f(p)$ is a transfer function of some low-pass filter (for our purpose, it is enough to use a second-order filter). Using (3) and (4), we have

$$p^2\omega_f(t) = -a\,p\,\omega_f(t) - b\,\omega_f(t) + \mathrm{K}\,x_f(t) \quad (6)$$

This may be regarded as a linear regression model $y(t) = \varphi^{\mathrm{T}}(t)\,\Theta$ where $y(t)$ and $\varphi^{\mathrm{T}}(t)$ are given by

$$\begin{aligned} y(t) &= p^2\,\omega_f(t) \\ \varphi^{\mathrm{T}}(t) &= \{-p\,\omega_f(t), -\omega_f(t), x_f(t)\} \end{aligned} \quad (7)$$

The purpose of the filter $\mathrm{H}_f(p)$ is to replace the derivative operators p and p^2, which can not be performed reliably in a presence of noise, with much less sensitive operators $p\,\mathrm{H}_f(p)$ and $p^2\,\mathrm{H}_f(p)$. Really, $p\,\omega_f(t)$ and $p^2\,\omega_f(t)$ are, in fact, $(p\,\mathrm{H}_f(p))\,\omega(t)$ and $(p^2\,\mathrm{H}_f(p))\,\omega(t)$. The weighted least squares approach assumes that $\Theta = \Theta(t)$ is a time-variant vector such that the weighted mean square error

$$\mathrm{E}[\Theta] = \int_0^t e^{-\alpha(t-u)}\,(y(u) - \varphi^{\mathrm{T}}(u)\,\Theta)^2\,\mathrm{d}u \quad (8)$$

is minimal, where α is "forgetting factor" (then, the "true" value of Θ is obtained for $t \to \infty$). It is known [1] that such Θ must satisfy the set of differential equations

$$\begin{aligned} \mathrm{d}\Theta(t)/\mathrm{d}t &= \mathrm{P}(t)\,\varphi(t)\,[\,y(t) - \varphi^{\mathrm{T}}(t)\,\Theta(t)] \\ \mathrm{dP}(t)/\mathrm{d}t &= \alpha\,\mathrm{P}(t) - \mathrm{P}(t)\,\varphi(t)\,\varphi^{\mathrm{T}}(t)\,\mathrm{P}(t) \end{aligned} \quad (9)$$

where $\mathrm{P}(t)$ is an auxiliary symmetric 3×3 matrix (the covariance matrix). In scalar form, (9) becomes

$$\begin{aligned} \mathrm{d}a/\mathrm{d}t = q_1\,e,\ \mathrm{d}b/\mathrm{d}t = q_2\,e,\ \mathrm{dK}/\mathrm{d}t = q_3\,e \\ e = \mathrm{d}^2\omega_f/\mathrm{d}t^2 + a\,\mathrm{d}\omega_f/\mathrm{d}t - \omega_f\,b - x_f\,\mathrm{K} \\ q_i = -p_{1i}\,\mathrm{d}\omega_f/\mathrm{d}t - p_{2i}\,\omega_f + p_{3i}\,x_f,\ i = 1..3 \\ \mathrm{d}p_{ij}/\mathrm{d}t = \alpha\,p_{ij} - q_i\,q_j,\ i = 1..3,\ j = i..3 \end{aligned} \quad (10)$$

Note that all quantities in (10) are time-dependent ones, although this is not shown explicitly, due to the shortage of the space. So, the estimator is a 9th-order nonlinear model, in which a, b, K and p_{ij}, $i = 1..3$, $j = i..3$ are state variables, and the other quantities are auxiliary ones. The integrators that implement p_{ij} state variables in the estimator have the possibility of resetting, so it is possible to reset these state variables (i. e. the covariance matrix) to their initial values using an external signal.

The first goal of the controller is to achieve a second-order dynamic response in respect to a set point change in the closed loop. We will first assume that the rectifier is an identity block, i.e. that $u_a = x$ (later we will remove this restriction). Also, let us assume for the moment that the main part of the controller is a general linear controller

$$x(t) = -\frac{\mathrm{S}(p)}{\mathrm{R}(p)}\,\omega(t) + \frac{\mathrm{T}(p)}{\mathrm{R}(p)}\,\omega^*(t) \quad (11)$$

Under such assumption, the closed loop transfer function in respect to a set point change is

$$\mathrm{W}(p) = \mathrm{K}\,\mathrm{T}(p)\,/\,[(p^2 + a\,p + b)\,\mathrm{R}(p) + \mathrm{K}\,\mathrm{S}(p)] \quad (12)$$

We want to find $\mathrm{R}(p)$, $\mathrm{S}(p)$ and $\mathrm{T}(p)$ such that

$$\mathrm{W}(p) = \Omega_0^2\,/\,(p^2 + 2\,\xi\,\Omega_0\,p + \Omega_0^2) \quad (13)$$

where Ω_0 and ξ are some predefined values (i. e. design parameters). In addition, the order of $\mathrm{S}(p)$ and $\mathrm{T}(p)$ must be smaller or equal to the order of $\mathrm{R}(p)$, due to the physical limitations of the controller. The simplest solution for $\mathrm{R}(p)$, $\mathrm{S}(p)$ and $\mathrm{T}(p)$ that satisfies all these requirements is $\mathrm{R}(p) = p + r_1$, $\mathrm{S}(p) = s_0\,p + s_1$ and $\mathrm{T}(p) = t_0\,(p + \lambda_0)$, where λ_0 is an arbitrary value, and

$$\begin{aligned} r_1 = \lambda_0 + 2\,\xi\,\Omega_0 - a,\ \ s_0 = (\Omega_0^2 + 2\,\xi\,\Omega_0\,\lambda_0 - b - a\,r_1)/\mathrm{K} \\ s_1 = (\Omega_0^2\,\lambda_0 - b\,r_1)/\mathrm{K},\ \ t_0 = \Omega_0^2/\mathrm{K} \end{aligned} \quad (14)$$

Therefore, the block "Design" in Fig. 2 should generate vector $\Pi = \{r_1, s_0, s_1, t_0, \lambda_0\}$ in according to (14). It is pure algebraic, non-inertial block.

Ideally, the main part of the controller may be a linear controller described by (11). However, in reality, the rectifier that acts as an actuator will always have saturation, which may produce wrong control due to the windup effect. Assuming that the saturation-type characteristic of the actuator is described as $u_a = \mathrm{sat}(x)$, it is shown in [4] that the windup effect may be avoided if the control law is changed in according to the following equation:

$$x(t) = \mathrm{sat}\,[\,\frac{\mathrm{T}(p)}{\Lambda(p)}\,\omega^*(t) - \frac{\mathrm{S}(p)}{\Lambda(p)}\,\omega(t) + (1 - \frac{\mathrm{R}(p)}{\Lambda(p)})\,x(t)] \quad (15)$$

Here, $\Lambda(p)$ is an arbitrary polynomial. It is easy to see that (15) reduces to (11) if $\mathrm{sat}(x) = x$. The simplest design is obtained when $\Lambda(p) = p + \lambda_0$, due to some cancellations that appears in (15) with such setting. Therefore, the main part of the controller is a quasi-linear anti-windup controller whose work is described by the following equation:

$$x(t) = \mathrm{sat}\left[t_0\,\omega^*(t) - \left(s_0 + \frac{s_1 - s_0\lambda_0}{p+\lambda_0}\right)\omega(t) + \frac{\lambda_0 - r_1}{p+\lambda_0}\,x(t)\right] \quad (16)$$

III. Using the Thyristor-fired Rectifier as an Actuator

In the previous section, we used the simplified approach where it is assumed that the controller directly produces the armature voltage, i. e. that $u_a = x$, eventually including the saturation effect. Now, we will consider more realistic situation, where a three-phased thyristor-fired phase-controlled rectifier is used as an actuator. The rectifier has 6 thyristors (two per each phase) and it produces on the output non-constant pulsating voltage with the pulsation frequency $6f$, where f is a frequency of AC voltage supply of the controller (assuming that $f = 50$ Hz, the pulsation frequency is 300 Hz). The mean value of the output voltage depends on the firing angle x, which can vary between $-\pi/2$ and $\pi/2$. Fig. 3. shows the shape of the output value of such rectifier assuming that the firing angle x varies from $\pi/2$ to $-\pi/2$ with step $\pi/8$, and that the change in x occurs every 0.02 seconds:

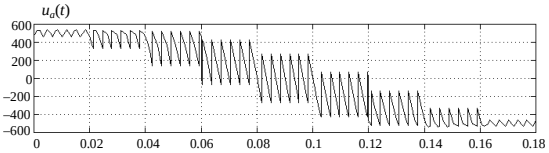


Figure 3. The output from the thyristor-fired rectifier

Such rectifier may be described using the following set of equations:

$$\begin{aligned} u_a(t) &= c_1(t)\,u_1(t) + c_2(t)\,u_2(t) + c_3(t)\,u_3(t) \\ u_i(t) &= \mathrm{U}_{max}\sin(2\pi f t - 2\pi i/3),\ \ i{=}1..3 \\ c_i(t) &= \sigma_i(t)\,(1-\sigma_{i+1}(t))\,(1-\sigma_{i+2}(t)) - \\ &\quad -(1-\sigma_i(t))\,\sigma_{i+1}(t)\,\sigma_{i+2}(t),\ \ i{=}1..3 \\ \sigma_i(t) &= 2\,\mathrm{sgn}\sin(2\pi f t - 2\pi i/3 + \pi/2 - x) - 1,\ \ i{=}1..5 \end{aligned} \quad (17)$$

The pulsations are fast enough in comparison to the mechanical bandwidth of the controlled motor that they will not have big influence to the controlled variable (i. e. the angular speed), although they will have significant influence to the currents in the motor. So, for the control purpose, we will pretend that u_a is constant when x is constant. We will see in Section 4 that although this is not really true, this works very good for the control purpose.

The real problem with the thyristor-fired rectifier in a control loop is a highly nonlinear dependency between x and the mean value of the output voltage. Namely, straightforward but tedious calculation based on (17) shows that the average value of $u_a(t)$ is proportional to $\sin x$. Therefore, the static gain of the cascade of the rectifier and the motor is highly dependent of the operating point (i. e. of the set point value of the angular speed), and tends to zero when x approaches $\pm\pi/2$. Consequently, we can expect big problems when the set point is changed, because this causes a change in the operating point and the overall parameters of the whole system.

It will be shown later in the paper that the designed regulator can deal with such changes in parameters, but only if the covariant matrix is reset after each change of the set point (this effectively starts a new estimation). Much better results are obtained by inserting an extra arc sine compensation block between the controller and the control input of the rectifier, so the control input of the rectifier is equal to the arc sine of the output of the controller. Such compensation block may be regarded as a part of the controller, so such augmented controller may be described by equation similar to (16), but with extra arc sine in front of the right side of the equation. It will be shown that such improved controller is very robust to any set point changes.

IV. An Improved Design

So far, the presentation given in the paper ignores completely the load torque M_l. However, the designed controller is quite robust to various disturbances caused by the existence of a load torque. This fact is quite obvious if the load torque acts as a viscose friction. Really, in such case we have $\mathrm{M}_l = \beta\,\omega$, where β is some coefficient. But, from (1) we can conclude that such load has the same effect as the incrementing the parameter B of the machine by β. As B is continuously estimated, the controller will easily adapt to such change in the parameter B.

The surprising fact is that the controller is also robust to step changes in the load torque. This looks surprising because it follows from (3) that a step change in the load torque acts similarly as a step disturbance superposed on the control voltage. It is known that compensation of a step disturbance on the input of the controlled object requires a controller with a pure integral action (i. e. with a pole at $p = 0$). But, the designed controller has no pure integral action (i. e. $\mathrm{R}(p)$ has no pole at $p = 0$). This apparently strange fact may be explained as follows. From (12), it follows that the closed loop gain $\mathrm{K_C}$ is

$$\mathrm{K_C} = \mathrm{W}(0) = \frac{\mathrm{K\,T}(0)}{b\,\mathrm{R}(0) + \mathrm{K\,S}(0)} = \frac{\mathrm{K}\,t_0}{b\,r_1 + \mathrm{K}\,s_1} \quad (18)$$

where we used derived expressions for $\mathrm{R}(p)$, $\mathrm{S}(p)$ and $\mathrm{T}(p)$. But, if we insert values for t_0, r_1 and s_1 from (14), we get $\mathrm{K_C} = 1$.This means that the closed-loop static gain with the designed controller is always 1, no matter of values of a, b and K. In other words, in the steady state the value of ω is always equal to the given set point, regardless of the estimated values of a, b and K (assuming that the controller can produce signal that is necessary to establish wanted set point value, without going into a saturation). The estimated values of a, b and K only have influence on the dynamic behavior. So, although the disturbance effectively changes the parameters of the controlled object, the controller will adapt to such changes.

It is possible to improve further the robustness of the controller to the step changes in the load torque by introducing a pure integral action in the controller. Namely, it is possible to find another solution for $\mathrm{R}(p)$, $\mathrm{S}(p)$ and

T(p) such that R(p) has a pole at p = 0, and that (13) still holds. The simplest such solution is $R(p) = p(p + r_1)$, $S(p) = s_0 p^2 + s_1 p + s_2$ and $T(p) = t_0 (p + \lambda_0)(p + \lambda_1)$ where λ_0 and λ_1 are arbitrary values, and

$$\begin{aligned} r_1 &= \lambda_0 + 2\xi\Omega_0 - a + \lambda_1, \quad t_0 = \Omega_0^2/\mathrm{K}, \quad s_2 = \Omega_0^2 \lambda_0 \lambda_1/\mathrm{K} \\ s_0 &= [\Omega_0^2 + 2\xi\Omega_0\lambda_0 - b + (\lambda_1 - a)(r_1 - \lambda_1)]/\mathrm{K} \\ s_1 &= [\Omega_0^2(\lambda_0 + \lambda_1) - b\, r_1 + 2\xi\Omega_0\lambda_0\lambda_1]/\mathrm{K}, \end{aligned} \tag{19}$$

This means that the designing block should now generate vector $\Pi = \{r_1, s_0, s_1, s_2, t_0, \lambda_0, \lambda_1\}$ in according to (19). The main part of the controller may be again expressed using (15), but now $\Lambda(p)$ must be at least a 2nd-order polynomial, because R(p), S(p) and T(p) are 2nd-order polynomials too. The simplest design is obtained when $\Lambda(p) = (p + \lambda_0)(p + \lambda_1)$. Further simplification is possible by taking $\lambda_1 = \lambda_0$. Therefore, the improved controller is described by the following equation:

$$\begin{aligned} x(t) = \mathrm{sat}\,[t_0\,\omega^*(t) + \frac{(2\lambda_0 - r_1)\,p + \lambda_0^2}{p^2 + 2\lambda_0 p + \lambda_0^2}\,x(t) - \\ - (s_0 + \frac{(s_1 - 2\lambda_0 s_0)\,p + (s_2 - \lambda_0^2 s_0)}{p^2 + 2\lambda_0 p + \lambda_0^2})\,\omega(t)] \end{aligned} \tag{20}$$

V. SIMULATIONS

The proposed design is demonstrated using a set of simulation experiments. The controlled object is a model of a real motor with parameters L_s = 0.12 H, R_s = 20 Ω, L_a = 0.1 H, R_a = 5.96 Ω, J = 0.056 kg m^2 and B = 0.2 N m s. In the first set of experiments, we assume that u_s is constant and set to U_s = 380 V. This gives the following values of the parameters of the linear model (4): K = 407.142, a = 63.17 and b = 1141.14.

The experiments are based on obtaining responses to step changes in the set point value of the angular speed between 50 rad/s and 100 rad/s and vice versa. The transfer function of the filter used in the estimator is chosen to be $H_f(p) = 1/(p^2 + p + 1)$. The "forgetting factor" in the first experiment is set to zero, i. e. α = 0. The parameters of the wanted response are chosen to be Ω_0 = 50 and ξ = 0.7. The free parameter λ_0 in the designing block is chosen to be λ_0 = 5. Also, in the first experiment, we assumed that the rectifier is nearly "ideal", i. e. that the controller generates directly the necessary armature voltage. However, we include the saturation effect even in the first experiment, i. e. we simply take that $u_a(x) = \mathrm{sat}(x)$, where $\mathrm{sat}(x) = x$ for $|x| \le U_{max}$ and $\mathrm{sat}(x) = U_{max}\,\mathrm{sgn}\,x$ for $|x| > U_{max}$ (without the saturation, the controller sometimes requires armature voltage over 10000 V during some short time intervals, which is impossible to realize in practice). We choose U_{max} = 600 V. The initial values of the machine parameters are chosen to be $a(0)$ = 0, $b(0)$ = 0 and K(0) = 1 (it is not possible to set K(0) = 0, because the designing block requires division by K). The initial value of the covariance matrix is set to P(0) = 10000 I, where I is 3×3 identity matrix. Fig. 4 shows that regardless of the absence of the initial estimates of a, b and K, these parameters converges rapidly to their true values. Fig. 5. shows the obtained response for ω and the control voltage u_a. Obviously, the obtained response is practically identical to the wanted response, except on the beginning, when the controller still has no good information about the object parameters.

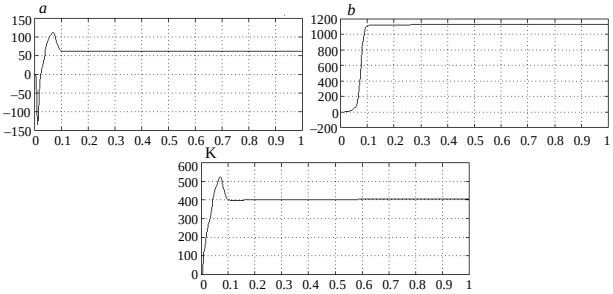


Figure 4. Convergence of parameters – simple experiment

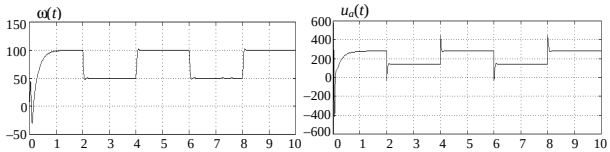


Figure 5. Obtained response – simple experiment

In the next experiment, the machine is modeled as a nonlinear object described with (1) instead of a linear model described with (2), i. e. without assumption that the stator current is constant. Under the constant stator voltage the current really becomes constant after initial transients vanishes, but during the time interval in which the current is non-constant, the estimator will accumulate completely wrong information about the process. So, it is not strange that parameters converge to wrong values, as shown in Fig 6. As a consequence, we get the response that differs considerably from the wanted response, as shown in Fig 7.

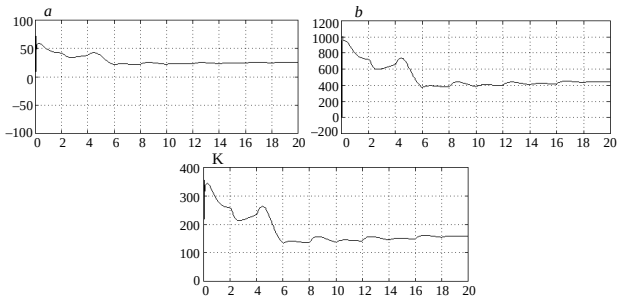


Figure 6. Wrong convergence – influence of transients

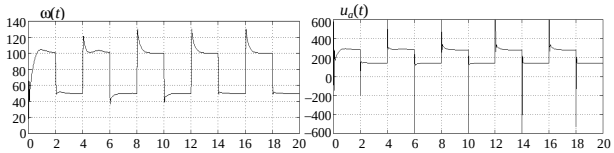


Figure 7. Bad response caused by wrong convergence

There are various approaches to solve this problem. For example, setting the "forgetting factor" α to some small value greater than zero can help, because in such case the estimator takes new measurements with greater weight than older ones, so it succeeds to "forget" values that are accumulated during the period of intensive current change. The next experiment is performed by taking α = 0.5. Fig. 8 shows that the correct convergence of parameters is really achieved with such approach, although relatively slowly. Consequently, the obtained response is much better, and becomes quite good when the values of parameters approach their correct values, as shown in Fig. 9.

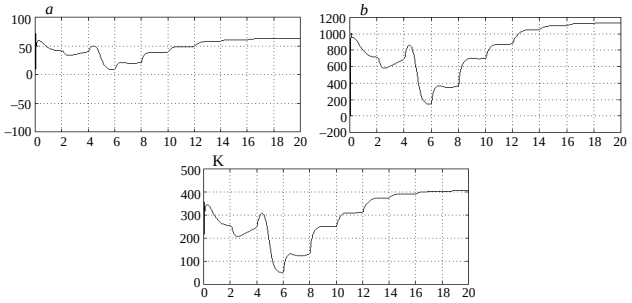


Figure 8. "Forgetting" can be helpful to achieve good convergence

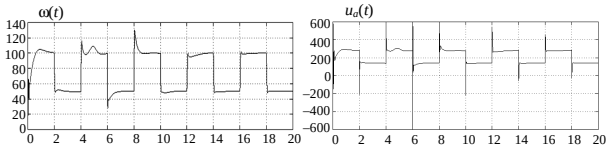


Figure 9. Improved response when "forgetting" is introduced

Note that greater values of α cause worse response at the begining, but the convergence of parameters is faster. From the other side, when α is too small, the convergence is too slow. Experiments show that value α = 0.5 is a good compromise.

Next experiment shows that even better results may be achieved by resetting the covariance matrix periodically. After each reset, currently achieved values of the object parameters become new initial values Then, the estimation starts from the beginning, with the initial value of the covariant matrix. In the experiment, the reset is performed after each change of the set point. The experiment shows that such strategy significantly speed up the convergence, as shown in Fig. 10. As a consequence, we get much better response, shown in Fig 11, not so far from the ideal one.

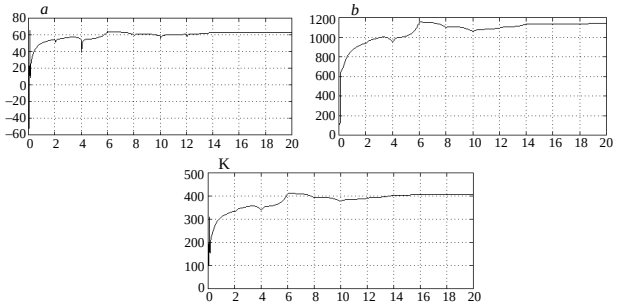


Figure 10. Resetting covariance matrix can speed up the convergence

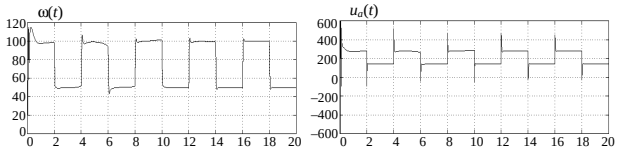


Figure 11. Nearly ideal response achieved by resetting the convariance matrix on each set point change

Now, the "ideal" rectifier that is used in the previous experiments is replaced with a realistic thyristor-fired rectifier described in Section 3. We first tested what will happen if no other changes in the control loop are performed. So, in the next experiment, the only principal difference is in the rectifier. However, the definition of sat(x) is changed. Now, we have sat(x) = x for $|x| \le \pi/2$ and sat(x) = $(\pi/2)$ sgn x for $|x| > \pi/2$, because x is now the firing angle, which must lay in the mentioned range. We can expect big problems after a set point change, as this changes the operating point, and consequently the parameters of the object, due to nonlinear nature of the rectifier. Really, Fig. 12 shows that the response is completely nonsense after a set point change (notice that the actuator goes into the saturation):

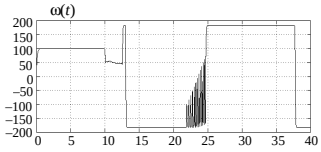


Figure 12. An useless response caused by untreated nonlinearity

The reason for such nonsensical response is the fact that the object parameters do not converge to sensible values. After a set point change, the parameters become wrong (as the gain now depends of the operating point), the controller goes into the saturation, the estimator stops acquiring information about the object, and finally, the parameters starts converging to zero, as shown in Fig. 13. Note that initial value of K was set to K(0) = 100000, as this order of magnitude is expected due to the order of magnitude of ratio $u_{a\,avg}/x$ in the rectifier.

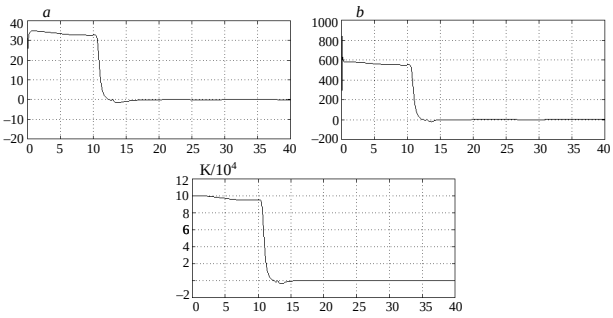


Figure 13. Loss of convergence after changing the operation point

A partial solution is resetting the covariance matrix after each set point change. This allows new estimation of object parameters after each change of the operating point. Fig. 14 shows that this strategy allows acceptable response, although quite far from ideal one. However, this shows that the unmodified controller (without compensation of the nonlinearity) is still capable to track big changes in the object parameters.

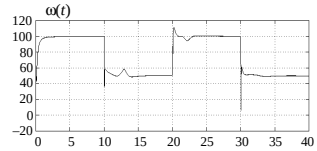


Figure 14. Acceptable response is possible even without compensation of the nonlinearity

Much better solution is compensation of the main nonlinearity in the rectifier, as described in Section 3. So, in the next experiment, we inserted an additional arc sine compensation block between the controller and the rectifier. We also changed again the definition of sat(x) to be sat(x) = x for $|x| \le 1$ and sat(x) = sgn x for $|x| > 1$, as the controller now generates sine of the firing angle instead of the angle itself. Fig. 15 shows that the response is now ideal. In addition to the angular speed, the figure shows the firing angle as the control variable, and the armature current. It is possible to see big ripple in the armature

current, caused by pulsations of the armature voltage. However, it is important that there is no any ripple in the controlled quantity (i. e. the angular speed). Fig. 16 also shows that the convergence of the object parameters to sensitive values is also achieved.

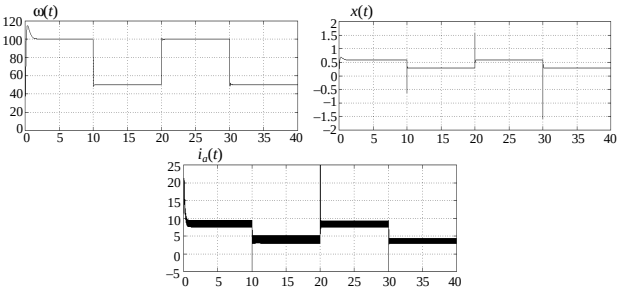


Figure 15. Ideal response achieved by compensation of the nonlinearity

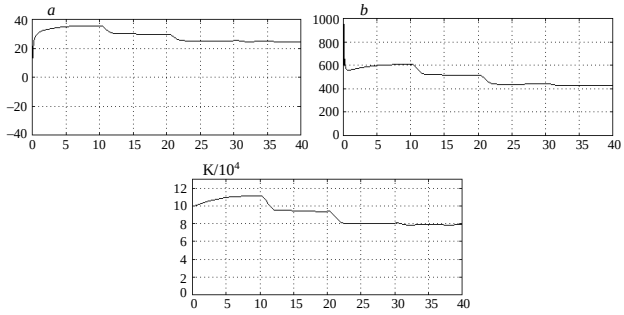


Figure 16. Convergence of parameters of compensated object

The initial value for K was again set to K(0) = 100000. However, except in the very beginning, the system works well even with the terribly wrong initial estimate K(0) = 1, as shown in Fig. 17. The bad response during first few seconds (that causes the saturation of the actuator) is caused by bad estimation of K in this period (the saturation will not cause serious problems due to applied anti-windup measures). The convergence of K with such initial estimate is also shown in the same figure.

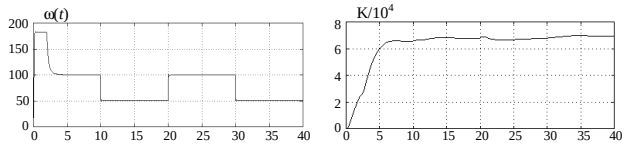


Figure 17. Response is good even with very bad initial estimates

Now, we will remove the assumption that u_s is constant. Instead, we will assume that u_s comes from the similar type of the rectifier too. This will induce pulsations in both the stator voltage and the stator current. However, experiments shows that their influence to the controlled angular speed is neglible, as shown in Fig. 18.

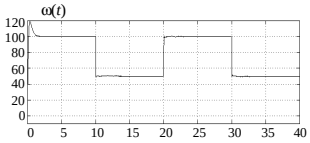


Figure 18. Response in presence of stator voltage pulsations

Finally, we tested the influence of load disturbances. Viscose load does not cause any change in the response, as explained in Section 4. Fig. 19 shows the influence of the step load of M_l = 20 N m that is applied 5 seconds after the beginning.

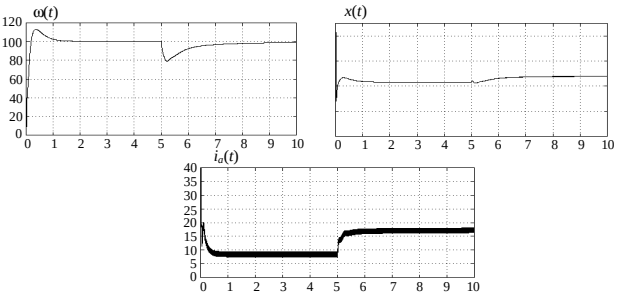


Figure 19. Influence of step load disturbance – basic design

We can see that even the basic design can remove quickly the influence of the disturbance. This is explained in Section 4. However, the advanced design developed in this section allows much stronger and faster removal of the disturbance as shown in Fig. 20. The experiments show that the dynamic of the disturbance removal depends of cutoff frequency of the filter used in the estimator. Lower cutoff causes faster response, but too low cutoff frequency can cause unstability.

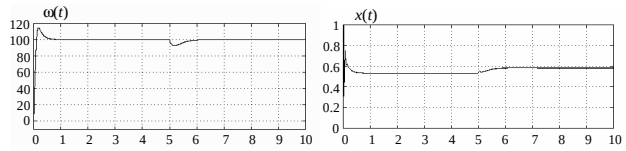


Figure 20. Influence of step load disturbance – advanced design

VI. CONCLUSION

This paper presents one advanced design of self-tuning controller for DC motor. Although the controller itself is somewhat complex, it is very robust and independent of the machine parameters. The response to the changes in the wanted angular speed is excellent, and the removal of the disturbances caused by the non-ideal rectifier and the presence of the load is also very good.

In applications, the working mode of the system dictates the need for covariance matrix reset. This flexibility makes the control algorithm suitable for a wide range of applications (including rare special cases of fast-changing parameters compared to the classical case of slow-changing parameter set). The controller design is deliberately shown in steps, since in some applications even the simple versions of the control scheme can do well.

REFERENCES

[1] K. J. Astrom, U. Borisson, L. Ljung, and B. Wittenmark, "Theory and applications of self-tuning regulators," Automatica, 13:457-476, 1977.

[2] T. Hägglund and K. J. Åström, "Supervision of adaptive control algorithms," Automatica, 36, pp. 1171–1180, 2000.

[3] P. J. Campo and M. Morari, "Robust Control of Processes Subject to Saturation Nonlinearities," Comput. & Chem. Eng., 14, 343. 1990.

[4] M. V. Kothare, P. J. Campo, M. Morari, and C. N. Nett, "A unified framework for the study of anti-windup designs," Automatica, 30, no. 12, pp. 1869–1883, 1994.

[5] K. J. Åström and L. Rundqwist, "Integrator windup and how to avoid it," Proceedings of the 1989 American control conference, Vol. 2, pp. 1693-1698, 1989.

[6] S. K. Pillai, Analysis of Thyristor Power-Conditioned Motors, Universities Press, 2003.

[7] Q. Yan and J. Sun, "A Model Reference Adaptive Control with Nonlinear Compensation for Robot Manipulators," Proceedings of the 1989 American control conference, Vol. 2, pp. 1023-1028, 1989.

An Experience on Manufacturing Execution System in the Field of Metals Industry

Jadranko F. Novak

Nomen d.o.o., Rijeka, Croatia

jadranko.novak@mynomen.hr

Abstract -- The paper presents an informal specification of a software product dedicated to manufacturing execution functions in long products hot rolling mill application field. The product is an implementation of so-called Level 3 functions of a CIM pyramid model. The main functional groups of the product are discussed, as well as principal features are described.

I. INTRODUCTION

Application functionalities covered by the product herein described are to be classified in a so-called Level 3 of the Computer Integrated Manufacturing (CIM) scheme. It is important to note that the product takes part of a project for a totally automated hot rolling mill, having a real-time and near-to-real-time integration between the three levels of automation: Level 1 Automation (Equipment Control), Level 2 Automation (Process Control) and Level 3 Automation (Production Control).

The classification has to be taken very informally. Its automated functions are divided in three levels. The goal of this paper is not to discuss the normative issues of the automation, but to get the attention of the reader to focus on application functions of the product. For the exposure to be correct, it has to be said that the upper classification is very simplified, and for that it does not follow strictly the actual normative and de-facto schemas for classification of automation functions and terminology.

II. PRODUCT DOMAIN

The product makes it possible to formalize (it is defined and described by the system) the following entities of interest for the Level 3 automation:

- production equipment and tools (steel plant and mill stand rolls);
- all materials involved in a production process (raw materials, finish products, auxiliary materials);
- finish product structure (bill of materials);
- list of operations and resources associated to a single production process (routings, or bill of procedures);
- clients and suppliers;
- work, sales, and purchase orders;
- materials and tools stocks;
- foreseeing and planning.

Also, it makes it possible:

- to handle events and to control the flow of activities related to production execution;
- to control co-operation with the lower level of the plant automation system (Level 2);
- to control integration with neighboring systems on the same level (continuous casting system, for example);
- to control integration with upper levels of automation/computerization (Enterprise Resource Planning System – ERP).

The main goal of the system is to have under control a complete production cycle relative to a given client's demand(s). The system acts as an intelligent advisor to operative department of the hot rolling mill plant.

A brief description of main functional groups of the system will be given in the text that follows.

III. MASTER DATA MANAGEMENT

A. Master Data

The following master data structures exist in the system:

- materials master (billets, finish products, auxiliary materials),
- stand rolls master,
- clients and suppliers master,
- stock master,

- hot rolling mill plant resources master.

The system provides all functions necessary for maintenance of the master data.

B. Quantity Issues

A quantity of a billet and finish product is kept as a <*weight, pieces, packs*> triplet, which are base (not derived) attributes of a material. The three values are memorized in all contexts dealing with material quantity. All quantity entry/modify functions provide a consistent and convergent algorithm for expressing a quantity as a quantity triplet. Exceptions: a stand roll has no quantity attribute (there may be one and only one roll of a given code), a quantity of an auxiliary material is expressed by a single value.

IV. CATALOGUES AND CONFIGURATION TABLES

There are number of catalogues and configuration tables used primarily to classify data into families, types, groups, etc., according to the logical data model of the application. For example:

- all materials are classified by so-called "material class" (billet, finish product, rolls, etc.),
- finish products are grouped into so-called "finish products families",
- orders are classified by order types,
- etc.

There are system-level and user-level catalogues and configuration tables. System-level catalogs are used to define system parameters and constants; this approach makes the system flexible for customization purposes. It is possible to establish one-to-many relation between a system- and a user-level parameter of the same type. For instance, to a "stand roll" (a system-level material class), various user-defined classes may be associated, as: *rolls for mono-block, rolls for intermediate mill*, etc.

V. BILL OF MATERIALS AND ROUTINGS

The bills, i.e. Bill of Materials (BOM) and Routings (Bill of Procedures - BOP), are data structures used to give a formalized answer to the following two questions:

- What is a row material (billet, auxiliary material) and how many of these is needed to produce a unit quantity of the given finish product?
- What are the operations to be performed upon the row material to obtain this finish product, and what are plant resources and tools used?

Several interesting features, related to the BOM and Routings, are supported by the system:

- Unlimited number of variants of the structure of the same finish product may be described by a single BOM. A variant is defined by following triplet: <*billet, routings, finish product*>. In other words, a single finish product may be obtained from the billet A applying the Routings R, or from the billet B applying the Routings Q, etc. One of variants is declared as a "primary variant".
- Routings is a list of operations (heating, rolling…). To each operation a set of resources may be associated. For instance, to the rolling operation a list of stands used by it is given, for each stand a stand roll and roll groove is defined. Even more, to each stand an unlimited number of <*roll, groove*> pairs may be associated, one of these is declared a "primary pair". A roll and a roll groove is not identified by a unique code, but by a type to which it belongs.
- The system provides for optimal selection of a physical roll pair on a basis of its physical attributes (roll type, groove type) and its wear, i.e. proposes one which best fits the requirements of the given work order lot.
- Also a concept of "time-based-validity" is handled for both bills. There may be more than one bill of the same code in the system, but only one valid in a given time interval. This enables a user to prepare future releases of products maintaining the same identification codes of bills; the system will consider only valid ones.

VI. STOCK MANAGEMENT

A. Stock Types

There may be unlimited number of stocks for billets, finished products, auxiliary materials and stand rolls. Each stock is seen as a number of locations.

B. Location Detail

A location may contain one or more so-called location details. A location detail is a set of a material of the same identifier. Location detail identifier consists of the following component identifiers: material id, lot id (heat number for billets), and id of order of origin. The order of origin may be a purchase order, in case of bought materials (billets, auxiliary materials), or work order in case of produced materials (finish products). A location detail may be free or allocated (billets booked by a work order, finish products booked by a sales order, etc.).

C. Dedicated Locations

A location may or may not be dedicated to a given material. For billets and finish products, a concept of dedicated location is typically used. A material-to-location association may be performed at location definition time (stock master), or upon a receipt of the first material into an empty location.

D. Quality Control Issues

A so-called material quality identification attribute (Q-ID) may be associated to a material. As far as stock management is concerned, a stock location may be dedicated to materials of given quality. The system compares a Q-ID of a material being moved and a Q-ID of a destination location; only movements having compatible Q-IDs are enabled.

E. Free Space Issues

For each location its physical dimensions are defined too, so a volume of the location is known. Each location detail is also associated a volume. This enables the system to calculate a remaining free space of a single location, in all transactions dealing with a material receiving. Also, for each location a so-called "cram factor" is defined. If dimensions of a material are multiples of location dimensions, and if there is no lost space (a very remote case), a cram factor is 1, otherwise it is less than one. The user may modify a default value (1) of the cram factor at location definition time. It is possible, in all transactions dealing with material receiving, that a user declares a location "full". Declaring "location is full" triggers a recalculation of the cram factor. This mechanism frees a user of location dimensions adjusting and helps a lot in obtaining realistic values for a remaining free space of the stock.

F. Stock Transactions

Three basic types of stock transactions are supported: issue, receipt and location transfer. Other specific transactions are also provided, as a transaction for inventory count. Stock transactions may be embedded in various context of order handling, like: picking billets and stocking finish products of a given work order, picking finish products for a given sales order, receiving billets ordered by a given purchase order. Also, various views and reports upon a stock data are available. Movements of materials by crane may also be supported by the system.

G. Stock Historian

Each exit and entry of a material from or to a stock location is memorized in an archive table called stock historian.

VII. STAND ROLLS MANAGEMENT

A stand roll is seen as a material subject to stock movements and as a tool subject to maintenance interventions. Thus, there are basically four sets of functions supplied for stand rolls management:

- stand roll master data management:
  - stand roll movements,
  - stand roll set-up,
- stand roll maintenance.

H. Roll Maintenance and Posting of Tons

For each stand roll a maintenance history is kept in the system. Initially a stand roll is defined by its minimum diameter, roll ID, roll type, total tons planned to be rolled, total number of machining's planned to be performed, a list of rolling mill stands on which it may be mounted, etc. Contextually, a so-called "initial technical level" of a roll is created, having the following data: state of the roll, diameter, hardness, next planned maintenance intervention, etc. Each time a roll is machined, a user is called to create a new technical level. *Number-of-tons-rolled-till-now* attribute of a groove is updated in real-time (tons posting), thanks to functional integration of the Level 3 with the Level 1 automation.

A stand roll may assume different states throughout its life cycle: new (blank roll), usable, being machined, consumed, etc.

A. Movements

A stand roll may be placed in:

- Stand Roll Stock;
- Stand Roll Workshop;
- Stand-By Location;
- Rolling Mill Train (in-line stands);

The following transactions are provided for roll movements:

- Issue;
- Receipt;
- Workshop Exit - creation of a new technical level is triggered contextually;
- Mount - mounting a roll (with a given groove selected) on an in-line stand;
- Quick Transfer - swap between stand-by locations and in-line stands;
- Location Transfer;

B. Stand Rolls Demand Preview and Usability Check

There are another two important functions which give a "closer look" on demands for stand rolls and on usability of rolls-on-hand for work orders which are about to be processed.

C. Stand Rolls Demand Preview

Demand for stand rolls, raised by work orders having the end-time within a given time interval, or raised by a selected work order is given. Note that, as a work order may be in different states of its life cycle, this function gives preview on both, planned and operative demand for stand rolls.

D. Usability of a stand roll / roll groove

For each work order lot, a stand-roll-type/groove-type configuration is given at work order creation time. This is compared with the actual rolling mill train configuration, i.e. with rolls/grooves actually mounted on in-line stands.

First, the roll type/groove type requested by the work order is checked against the actual rolls/grooves mounted. Second, the rolling capacity of mounted grooves is checked against the rolling quantity of a work order lot.

VIII. MANAGEMENT OF ORDERS

The following groups of orders are managed by the system:

- work orders,
- input orders,
- purchase orders.

It is important to underline that the main attention of the Level 3 system is paid to management of Work Orders, because it is a work order which is the main object of interest of a Manufacturing Execution System (MES).

A. Input Orders

The name "Input Orders" is used to denote orders by which a demand for finish product, in some point in time scale, is defined. Two types of input orders are distinguished by the system:

- Sales Order (demand is raised by a customer);
- Stock Order (demand is raised by the rolling mill plant itself, in order to produce for a stock).

Note that the upper distinction is made in order to be able to handle specific characteristics of both types of upper orders. The system uses, in this context, a so-called "supply type", to specify if a demand for a finish product is to be:

- accomplished by the quantity-on-hand found on the stock, and/or
- accomplished by issuing a dedicated work order (typically used for special products).

A. Purchase Demand

There could be two different life steps of a purchase demand: Purchase Proposal and Purchase Order. Purchase proposals are generated by a simple Material Requirements Planning (MRP) function, which takes part of the Level 3 system. Purchase Proposals express a way in which the MPR suggest to cover demands for materials which do not exist on the stock in quantity and within time interval needed by the production plan. Purchase proposals have to be transformed into purchase orders by a person responsible for this domain of the plant business process. This person may or may not respect the MRP suggestions. Also, the purchase could be created directly by a purchase order. Typically, the Purchase Proposals created by the Level 3 are transferred to the company ERP system for subsequent handling.

B. Work Orders

A work order is a set of data describing completely what, how, how many, and when is to be produced. It may be defined on the basis of input received from a company ERP, the Level 3 planning function (strategic or executive planning) can create it, or it may be created the Level 3 manual input. Work order data is created on the basis of:

- finish product demand (formalized by an input order),
- bill of materials data,
- routings data.

A work order may assume different states during its life cycle, as follows:

- Planned - Created by the planning function, or inherited from the company ERP.
- Open - Made known to the operative domain of the plant. It could be a planned work order which state has been changed to "open", or a manually created work order having initially the state attribute set to "open".
- Released - A part or a whole quantity of the work order is ready for production. Contextually a so-called work order lot (work order detail) is created. Finish product quantity produced by a single work order lot has to be produced with billets of the same heat number. Contextually to releasing, a raw material needed by the work order lot is allocated and the billet picking list is created.
- Initiated - Billets needed by the lot are put in front of the re-heating station and work order lot processing may start.
- Completed - Production process relative to the work order lot in question is completed. The quantity produced is reported to the system.
- Stocked - The whole quantity produced by the work order lot is on the stock.

The system gives all functions necessary for creation, selection and modification of work orders, as well as for control of work order state transition.

IX. PLANNING

The planning function is divided into two steps, so-called Strategic Planning and Executive Planning. The temporal domain of the strategic planning (strategic planning horizon) is usually four to five weeks. The executive planning usually acts upon the first week of the strategic planning horizon, i.e. acts in the temporal domain near to "now". [4]

A. Strategic Planning

Strategic planning is an optional module of the product. It could be used as temporary "surrogate" tool for creation of the main plan of the production, in case a customer does not have any ERP system to perform this function. It gives several simple functions for definition of temporal sequence of work orders, created to cover the net demand for finish products. The strategic planning takes into account:

- Technological sequence of production, and
- Rolling mill capacity.

Strategic planning has no means to substitute ERP functionalities in this field.

B. *Executive Planning*

A simple MRP function, incorporated in the Level 3 system, is one of the main building blocks of the executive planning. Here, gross demands for raw materials are obtained in the first step. Then, net demands for raw materials are calculated, subtracting (from gross demand) a non-allocated quantity existing on the stock, and taking into account quantities in already raised purchase proposals. In the last step, purchase proposals for billets and for auxiliary material are created in order to cover its net demands.

Also Executive Planning is an optional module of the Level 3 system. Basically, if the company-wide planning (ERP) goes into details important for operative handling of the rolling mill plant, than there is typically no need to perform the operative planning in the Level 3 system. In case this is not a truth, the Executive Planning module is a valid tool for plant-level planning. It also presents a congruent tool for integration of the Level 3 with the company ERP.

X. LEVEL 3 - 2 INTEGRATION

The interaction between the two levels is performed by message exchange. Typical scenario is as follows (lot is used as synonym for a lot of work order, i.e. for a work order detail):

- *Level 3 to Level 2 massages*:
  - Lot Download,
  - Lot Abort,
  - Lot Billet Quantity Confirmation,
  - Train Configuration.
- *Level 2 to Level 3 massages*:
  - Lot Rolling Start,
  - Lot Rolling End,
  - Lot Processing End,
  - Lot Sequence Confirmation.

The last message in each of the upper two lists is asynchronous, others are synchronous. It is the Level 2 operator who is supposed to establish the real scheduling sequence (if different from proposed one) for released and not yet initiated work order lots. This sequence is reported to the Level 3 by Lot Sequence Confirmation message. The Level 3 respects the scheduling sequence defined by the Level 2. For instance, in the context of billet picking it suggest picking first the billets of the lots in the head of the list.

XI. CONCLUSION

The paper gave a very brief list of the main features of the product. Some functional groups are not even mentioned, for others just titles are given. Planning is not discussed, integration between the Level 2 and Level 3 is just mentioned, and integration between the Level 3 and the ERP is not analyzed.

LITERATURE

[1] Purdue Reference Model: Williams, T.J., ed., *A reference Model for Computer Integrated Manufacturing (CIM), A Description From the Viewpoint of Industrial Automation*, Minutes, CIM Reference Model Committee, International Purdue Workshop on Industrial Computer Systems, Purdue University, West Lafayette, IN (1988) ISA, Research Triangle Park, NC (1989) (see www.pera.net),

[2] ANSI/ISA-95.00.01-2000, *Enterprise-Control System Integration, Part 1: Models and Terminology*. Copyright by ISA - The Instrumentation, Systems, and Automation Society (for information, visit www.isa.org),

[3] *MES Functionalities & MRP to MES Data Flow Possibilities*, MESA (Manufacturing Execution Systems Association) International White Paper n.2, 1997 (for information see www.mesa.org).

[4] Roger W. Schmenner, *Production/Operations Management. Concepts and Situations*, Science Research Associates, Inc., 1984.

Primjena zelenog računalstva u podatkovnom centru Rafinerije nafte Rijeka

Sergej Horvat*, Boris Žeželj ** i Zoran Grgić*
* Rafinerija nafte Rijeka, Procesna informatika, Kostrena, Hrvatska
** Rafinerija nafte Rijeka, Automatsko vođenje procesa, Kostrena, Hrvatska
sergej.horvat@ina.hr, boris.zezelj@ina.hr, zoran.grgic@ina.hr

Sažetak – Rafinerija nafte Rijeka je pokrenula projekt izgradnje informacijskog sustava za podršku procesima proizvodnje MES eng. manufacturing execution systems. Kao nužan preduvjet za uspostavu informacijskoga sustava za podršku procesima proizvodnje bila je izgradnja računalne infrastrukture odnosno podatkovnog centra. Glavni ograničavajući faktor pri projektiranju podatkovnog centra bila je fizička veličina prostorije u koju se uz već postojeću aktivnu komunikacijsku opremu trebala smjestiti nova računalna infrastruktura koristeći pritom standarde TIA-942, TIA-607 i IEEE 1100 standarde. Usvojen je prijedlog za korištenjem na tržištu dostupnih tehnologija koje su nudile istovremeno veoma visoku dostupnost, fleksibilnost, modularnost, optimizirano iskorištenje računalnih resursa i fizičkog prostora, centralizirano upravljanje a sve zajedno su poznate pod zajedničkim nazivnikom zeleno računalstvo (eng. green computing). Izabrana programska osnovica MES sustava je istovremeno podržavala implementaciju svih elemenata sustava u virtualnoj okolini koji su značajno smanjili investiciju u infrastrukturu podatkovnog centra kao i smanjenu ukupnu potrošnju energije. Prvi pokazatelji ispravno izabrane strategije ogledaju se u implementaciji projekta unutar vremena i planiranih financijskih sredstava te kasnijoj lakoći upravljanja infrastrukturom te skalabilnosti u pravom smislu riječi.

I. UVOD

Sve stroži zahtjevi i nužnost brže prilagodbe poslovnih procesa prema brzo razvijajućem poslovnom okružju neminovno su doveli do potrebe za automatiziranim procesom izrade dnevne materijalne bilance u Rafineriji nafte Rijeka. To je osnova za računovodstvo, kontrolu, praćenje performansi, proizvodno planiranje, strateško planiranje, vrednovanje i kontrolu emisija. Usklađenost s ekološkim i sigurnosnim propisima jedan je od najvažnijih zadataka rafinerije u sadašnjosti. Materijalno bilanciranje je dio strategije koje bi osiguralo pravilno gospodarenje proizvodnim procesima i kao takvo postaje standardno obilježje procesnog informacijskog sustava (eng. plant information system). Analiza postojećeg stanja je pokazala da postoji potreba za unapređenjem rafinerijskog informacijskog sustava i procesa materijalnog bilanciranja prema točkama kako slijedi:

- sustav za prikupljanje procesnih podataka trenutno pokriva samo dio rafinerije i ne postoji automatsko prikupljanje svih potrebnih podataka,
- kretanja materijala među spremnicima se bilježe na papiru za svaki spremnik pojedinačno,
- količine materijala procjenjuje služba za materijalno bilanciranje,
- ručni unos u vlastito razvijenu aplikaciju koja se koristi za potrebe materijalnog bilanciranja,
- kašnjenja sa dostavom izvještaja vezanih za poslovanje rafinerije (proizvodnja, namješavanje, otprema),
- nedostatak odgovarajućih kalkulacija na DCS (eng. distributed control system) razini kao: kompenzacija, akumulacija, VCF (eng. volume correction factor) kalkulacija, kalkulacija entalpije pare.

Glavni ciljevi integralnog projekta jesu uspostava pouzdanih i točnih primarnih mjerenja na spremnicima i granicama postrojenja te implementacija informacijskog sustava za prikupljanje i obradu podataka (PI system 2010) u realnom vremenu i Sigmafine-a alata za materijalno bilanciranje kako bi se:

- poboljšala kvaliteta i brzina izvještavanja,
- poboljšala točnost i kvaliteta podataka,
- stavile na raspolaganje informacije o poslovnom sustavu kao podrška za donošenje odluka,
- osigurale bolje informacije za planiranje, raspoređivanje i donošenje odluka,
- smanjili proizvodni gubici,
- redefinirali radni procesi,
- optimizirali ljudski resursi - određivanje potrebne radne snage.

Izgradnja podatkovnog centra bila je kao nužan preduvjet i zaseban projekt koji je zbog ranije napomenute fizičke ograničenosti prostorom buduće server sobe nametnuo uporabu na tržištu dostupnih tehnologija poznatijih pod zajedničkim nazivnikom (eng. green computing) zeleno računalstvo.

II. SERVER SOBA

Za potrebe podatkovnog centra informacijskog sustava rafinerije nafte Rijeka pristupilo se uređenju prostora i povećanju kapaciteta postojeće sistem sobe u prizemlju centralne kontrolne zgrade u rafineriji nafte Rijeka.

Predmet projekta je bilo projektiranje [1]:

- napajanja svih potrošača podatkovnog centra mrežnim i besprekidnim napajanjem,
- razvod elektroinstalacija te uzemljenja i izjednačenja potencijala,
- sustava za dojavu i gašenje,
- klimatizacijski sustav
- tehnička zaštita
- ostali građevinski i strojarski radovi

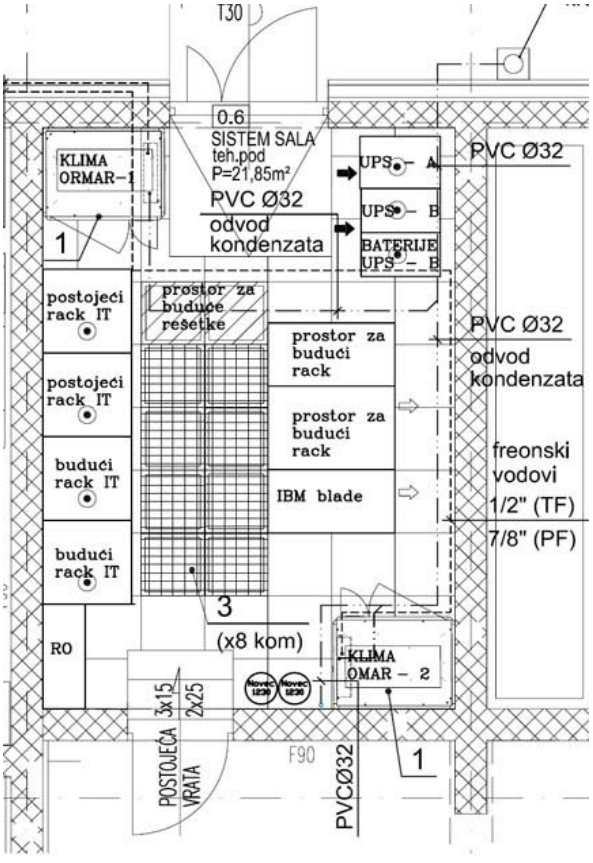
Sistem soba prikazana na slici 1. je smještena u prostoriju površine 22 m2, u koju postoje ulazi iz unutrašnjosti zgrade i iz vanjskog prostora. U prostoriji su se nalazila dva komunikacijska ormara s računalnom opremom, te uređaj za besprekidno napajanje snage 20 kVA u nastavku teksta UPS (eng. uninterruptible power supply). Ugrađena su dodatno 3 komada ormara, od kojih jedan tipa 'blade', te još jedan UPS snage 20 kVA. Organizacijom prostora predviđeno je i pričuvno mjesto za eventualnu naknadnu ugradnju do dva nova standardna ormara. Klimatizacija prostora je izvedena novom ugradnjom dva klima ormara u prostoriji koji upuhuju zrak u tehnički pod, i pripadnih vanjskih jedinica. Postavljen je novi razdjelnik jake struje iz kojeg se napajaju svi već navedeni potrošači. Ovaj se razdjelnik napaja s pričuvnog strujnog kruga u glavnom razdjelniku građevine, koji je smješten u obližnjoj prostoriji u prizemlju. Glavni razdjelnik sistem sobe je samostojeći metalni ormar s ugrađenom opremom za energetski razvod, zaštitu strujnih krugova i praćenje parametara priključka i potrošnje pojedinih trošila.

Energetski distributivni dio razdjelnika funkcionalno je podijeljen u tri cjeline: mrežni, sustav UPS-a 'A', te sustav UPS-a 'B'. Besprekidnim napajanjem s 'UPS' sabirnica razdjelnika napajaju se potrošači u ormarima i PLC (eng. Programmable Logic Controller) uređaj u samom razdjelniku. Svi ostali potrošači, kao rasvjeta, priključnice i klimatizacija, napajaju se s mrežnih sabirnica. Uređaji za besprekidno napajanje (A i B) pojedinačnih su snaga 20 kVA, i svaki od njih je dovoljan da na sebe preuzme svu potrošnju priključenih potrošača u sistem sali, tako možemo reći da je postignuta potrebna sigurnost napajanja, odnosno redundancija '1+1'. Svaki od UPS uređaja ima i unutarnji premosnik (eng. bypass) koji se automatski uklapa u slučaju kvara i na izlaz prenosi mrežno napajanje. Ormari za smještaj računalne opreme u sistem sali su sljedećih vrsta:

- ormari za smještaj računalne opreme u otežanim uvjetima, dimenzija 800x1100x2000 mm ('blade') ,
- klasični ormari za komunikacijsku opremu, dimenzija 800x900x2200 mm .

Kao što je već ranije rečeno svaki se ormar napaja dvostrano, i to preko inteligentnih i modularnih PDU (eng. power distribution unit) jedinica (dva komada po ormaru od kojih je jedna priključena na A, a druga na B stranu UPS razvoda). U glavnom razvodnom ormaru nalazi se elektroenergetska i zaštitna oprema pripadnih strujnih krugova (osigurači, prekidači…), kao i uređaji ION 7600 (Intelligent metering and control device) za mjerenje parametara glavnih priključaka (napona, struje, snage, faktora snage, viših harmonika) i ostalih parametara kvalitete napajanja u skladu s EN 50160, koji su preko ethernet priključka spojeni i na centralno nadzorno upravljački sustav u daljnjem tekstu CNUS. Automatski osigurači/prekidači opremljeni su pomoćnim kontaktima za signalizaciju stanja. Za prikupljanje informacija s opreme u razdjelniku (pomoćni kontakti osigurača, prekidača i sklopnika), analognih informacija s uređaja za mjerenje temperature i vlažnosti prostora u sistemskoj sobi, stanja UPS-a, ugrađen je PLC, koji pak prikupljene informacije prosljeđuje u CNUS.

Radi praćenja veličina struja prema računalskoj opremi, u ormar je ugrađen i analogni mjerni modul tip VERIS H704-42, a u kompletu ima 2*21 strujni transformator s mogućnošću mjerenja struja u rasponu do 50 A, također povezan na CNUS. Za signalizaciju pojave vlage u šupljini poda ugrađen je senzor sa sondom, koji je spojen sa centralnom jedinicom. Na centralnu jedinicu su spojeni i osjetnici temperature u ormarima. Na PLC je spojen osjetnik temperature i vlage u prostoru postavljen na zid server sobe.

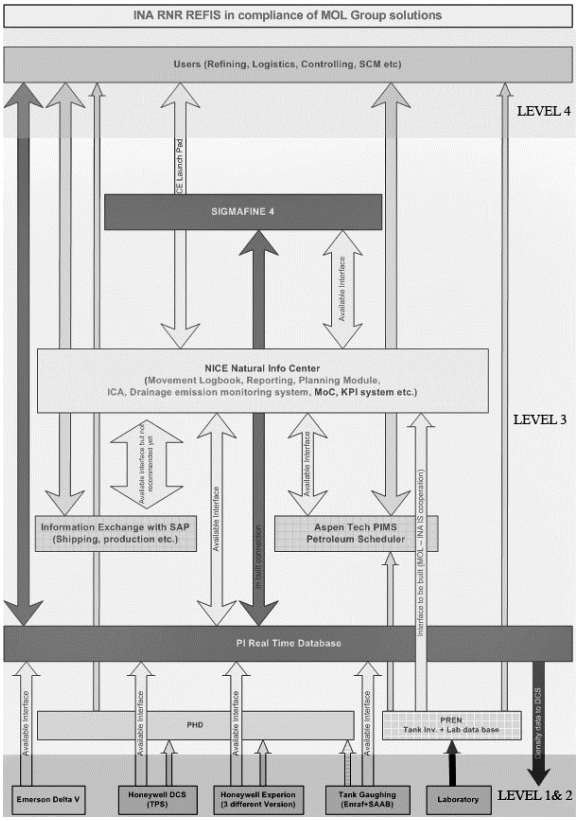


Slika 1. Server soba informacijskog sustava rafinerije

III. RAFINERIJSKI INFORMACIJSKI SUSTAV

Rafinerija nafte Rijeka projektom izgradnje informacijskog sustava kao što je prikazano na slici 2. je željela povezati najniže razine automatizacije proizvodnog procesa sa nadređenim razinama s ciljem prikupljanja i objedinjavanja podataka na centralnom mjestu. U tu svrhu izabrana je platforma renomiranog proizvođača programske podrške tvrtke Osisoft te programsko rješenje PI system 2010 (eng. plant information system). Samo programsko rješenje je industrijski standard za upravljanje podacima i događajima u realnom vremenu koje istovremeno omogućuje pregled povijesnih podataka, razne vrste pretraga, analize, kalkulacije, notifikacije statistiku te vizualizaciju. Komunikacija između najniže razine automatizacije i PI sustava odvija se OPC protokolom. OPC je akronim za OLE (eng. object linking and embedding) što je povezivanje i ugrađivanje objekata za (eng. process control) procesnu kontrolu. Cilj OPC komunikacijskog protokola je osigurati otvoren, fleksibilan, plug and play standard za interoperabilnost programske podrške u industrijskoj automatizaciji. OPC tehnologija se temelji na OLE (ActiveX) tehnologiji i komunikacijskim modelima COM (eng. component object model) i DCOM (eng. distributed component object model). OLE/COM tehnologije definiraju se kako individualne programske komponente koje mogu među djelovati i dijeliti podatke. OPC predviđa standardno sučelje između različitih procesno kontrolnih uređaja, bez obzira na kontrolne aplikacije ili uređaje u procesu. Skraćenica za (eng. component object model) ili COM predviđa standardno sučelje i među-komponentnu komunikaciju na lokalnom PC-u. Preko COM-a, aplikacija može koristiti karakteristike bilo kojeg aplikacijskog objekta ili operativnog sustava, ili dozvoliti dogradnju programske komponente bez smetnji u izvođenju. Kao osnovni standard COM je osnova i srž DCOM, ActiveX i OLE tehnologijama.

U samom projektu izgradnje informacijskog sustava rafinerije nafte Rijeka kao standardno rješenje za prikupljanje podataka sa različitih sustava je uzeto PI OPC DA (eng. data access) sučelje za pristup podacima. Sva OPC sučelja su konfigurirana unutar virtualne okoline koja je smještena na tri fizička računala imenom Blade šasija slot jedan, dva i pet. Procesna postrojenja u rafineriji kontroliraju se sa dva DCS (eng. distributed control system) distribuirana kontrolna sustava i to proizvođača tvrtki Honeywell i Emerson. Procesna komunikacijska mreža kontrolnog sustava Honeywell se sastoji iz dva dijela, stariji je baziran na token bus arhitekturi te noviji na ethernet protokolu koji su međusobno povezani te se stoga razmatraju kao jedna logička cjelina uvažavajući pritom ograničenja brzine komunikacijskog protokola od 5 Mb/s koji sa sobom nosi token bus. Podaci sa Honeywell procesnih računala se prikupljaju sa tri neovisna izvora: 2 Honeywell Experion redundantna računala poslužitelja na postrojenjima Energana i Blending te jednom redundantnom paru ESVT Experion poslužitelja na postrojenju FCC. Izvor podataka Energana prikuplja podatke sa pripadnog kotlovskog postrojenja, postrojenja turbogeneratora te postrojenja za kemijsku pripremu vode.



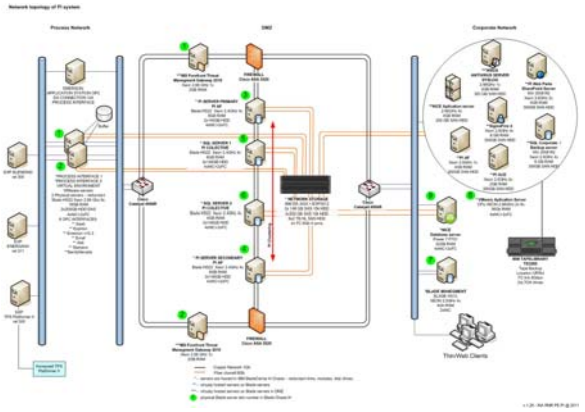
Slika 2. Tokovi podataka informacijskog sustava rafinerije sa pripadnim razinama automatizacije prema ISA-95 standardu

Izvor podataka Blending prikuplja podatke sa postrojenja za namješavanje derivata, izvor podataka FCC prikuplja podatke sa postrojenja atmosferske destilacije, postrojenja vakuum destilacije, postrojenja za odvajanje sumpora, postrojenja fluid katalitičkog krekinga, postrojenja za hidrodesulfurizaciju, postrojenja izomerizacija, postrojenja visbreaking, te postrojenja platforming I i II . Svaki od ranije spomenutih izvora podataka je spojen redundantnim PI OPC DA sučeljem na PI sustav omogućavajući tako kontinuirano prikupljanje procesnih podataka. Sama konekcija ima mogućnost bidirekcionalne komunikacije budući postoji potreba i za upisivanje podataka laboratorijskih analiza prema nazad. Distribuirani kontrolni sustav tvrtke Emerson je baziran na ethernet komunikacijskom protokolu. Kao izvor podataka koriste se dvije aplikacijske stanice koje će osigurati nesmetan i kontinuiran tok podataka sa postrojenja za proizvodnju vodika, postrojenja hidrokreker te postrojenja za odvajanje sumpora. Uz ranije spomenute sustave u rafineriji se prikupljaju podaci sa spremnika koji se baziraju na sustavima mjerenja razine u spremnicima proizvođača Honeywell Enraf i Saab. Saab sustav je izveden na način da su podaci sa sklopovlja spojeni na centralni poslužitelj sa kojega se preko redundantnog PI OPC DA sučelja prosljeđuju na PI sustav. U slučaju Honeywell ENRAF sustava šest samostalnih poslužiteljskog računala prikupljaju procesne podatke sa spremničkih grupa (A, B, C, D, E i S) farme spremnika. Nadalje sva strateški bitna i kritična rotacijska oprema je spojena na poslužitelj proizvođača Bently Nevada s programskom platformom System 1 koja osigurava praćenje stanja i dijagnostiku kritične rotacijske opreme.

Za potrebe informacijskog sustava rafinerije kao što je prikazano na slici 3. implementirano je 17 računala poslužitelja od toga 13 baziranih na Windows 2008 R2 x64 operativnom sustavu te 1 računalo poslužitelj na AIX 7.1 operativnom sustavu za Oracle 11 bazu podataka. Dimenzioniranje Vmware ESXi poslužitelja za virtualnu okolinu napravljeno je prema zahtjevima pojedinih poslužitelja koji se u njoj nalaze dodajući dodatne resurse kako bi se svi poslužitelji sa pojedinog ESXi poslužitelja u slučaju hardwerskog kvara mogli prebaciti na slobodne resurse operativnih ESXi poslužitelja. Dimenzioniran je virtualni klaster od tri fizička poslužitelja od kojih dva mogu preuzeti cjelokupno opterećenje virtualne okoline koristeći Vmware vMotion tehnologiju. U virtualnu okolinu implementirani su svi poslužitelji za koje je ranije utvrđeno ispravno funkcioniranje unutar virtualne okoline. U fizičkoj okolini implementirani su poslužitelji PI sustava (primarni i sekundarni) te poslužitelj za Oracle bazu podataka baziranu na PowerPC arhitekturi i AIX 7.1 operativnom sustavu.

Tablica 1. Raspodjela poslužitelja rafinerijskog informacijskog sustava prema okolini i OS-u

	Server	OS, okolina
1	PI primarni server 2010 MS SQL Server 2008replikacija	Windows 2008 R2 Fizička okolina
2	PI sekundarni server MS SQL Server 2008replikacija	Windows 2008 R2 Fizička okolina
3	Nice Oracle 11g	AIX 7.1 Fizička okolina
4	Blade Management server IBM Tivoli Server	Windows 2008 R2 Fizička okolina
5	MS SharePoint 2010	Windows 2008 R2 Virtualna okolina
6	MS SQL Server 2008	Windows 2008 R2 Virtualna okolina
7	Sigmafine	Windows 2008 R2 Virtualna okolina
8	PI Advanced computation engine	Windows 2008 R2 Virtualna okolina
9	Nice aplikacijski server	Windows 2008 R2 Virtualna okolina
10	Proces interface A	Windows 2008 R2 Virtualna okolina
11	Proces interface B	Windows 2008 R2 Virtualna okolina
12	Proces interface C	Windows 2008 R2 Virtualna okolina
13	Proces interface D	Windows 2008 R2 Virtualna okolina
14	Proces WSUS, antivirus i SYSLOG	Windows 2008 R2 Virtualna okolina
15	MS Threat managment gateway 2010	Windows 2008 R2 Virtualna okolina
16	MS Threat anagment gateway 2010	Windows 2008 R2 Virtualna okolina
17	VmWare Virtual centar	Windows 2008 R2 Virtualna okolina
18	IBM Tivoli Monitor	Windows 2008 R2 Virtualna okolina

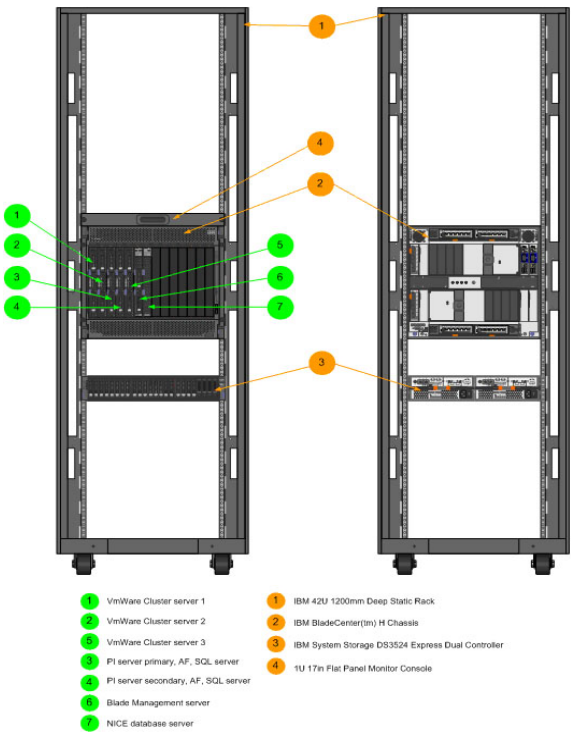


Slika 3. Logička topologija rafinerijskog informacijskog sustava

Tijekom prve godine eksploatacije implementirana su još 3 poslužitelja u virtualnu okolinu jednostavno i bez dodatnih troškova.

IV. HARDWERSKA PLATFORMA

Hardwerska platforma dizajnirana je prema potrebama rafinerijskog informacijskog sustava kako bi se osigurala visoka dostupnost svih servisa. Izvršena je instalacija novog SAN (eng. storage area network) okruženja baziranog na IBM DS3524 diskovnom sustavu, IBM xSeries i pSeries poslužiteljima u IBM H blade šasiji (ukupno sedam poslužitelja, od kojih šest posjeduje FC (eng. Fibre Channel) HBA adaptere), IBM SAN optičkim preklopnicima (Brocade, 2 komada) i pripadajućoj dodatnoj opremi. Prikaz instalacije opreme vidljiv je na slici 4.



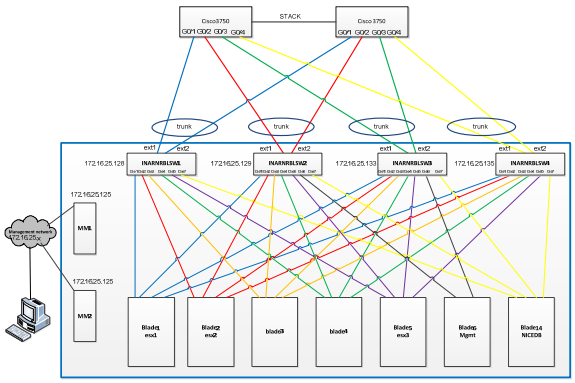
Slika 4. Fizička konfiguracija Blade poslužitelja

Visoka dostupnost se osigurava na hardverskom nivou instalacijom redundantnih komponenti sustava:

- redundantnim napajanjima na poslužiteljima i sustavu za skladištenje podataka,
- sve mrežne konekcije od virtualnih poslužitelja prema fizičkoj komunikacijskoj mreži putuju redundantnim linkovima,
- sve FC (eng. Fibre Channel) konekcije od fizičkih poslužitelja prema fizičkoj infrastrukturi sustava za skladištenje podataka putuju redundantnim FC linkovima,
- za prihvat mrežnih konekcija s blade šasija koriste se dva preklopnika u posebnom modu rada (eng. stack) gdje djeluju kao jedan i svaki blade poslužitelj ima mrežnu konekciju na oba preklopnika.

Svi uređaji koji posjeduju više mrežnih konekcija su spojeni sa više bakrenih kablova sa preklopnicima, a ostali su jednim kabelom ili interno spojeni na preklopnik. Nadzor i upravljanje uređajima vrši se putem lokalne računalne mreže i TCP protokola. Na slici 5. je prikazana fizička shema lokalne računalne mreže vezane uz Blade okolinu.

Za pohranu prikupljenih podataka koristi se SAN sačinjen od 24 SAS (eng. Serial Attached SCSI) diska kapaciteta 600 GB podijeljenih u 3 RAID 5 polja. Dva diska su rezervirana kao 'Hot Spare'. Ostvarena je potpuna redundancija svih poslužitelja u komunikaciji prema diskovnim sustavima (dvostruki optički spojni put od diskovnog sustava na SAN preklopnik). Sva eksterna SAN komunikacija ostvarena je putem optičkih multimodnih kabela i radi na brzini od 8 Gbit/s.



Slika 5. Fizička shema lokalne računalne mreže sustava Blade okoline

V. VIRTUALIZACIJA

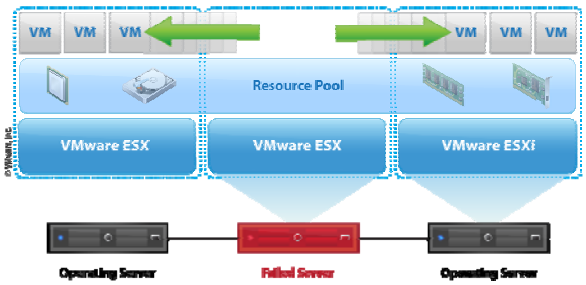
Pod pojmom virtualizacija u računarstvu se podrazumijeva apstraktno predstavljanje pojedinih funkcionalnosti i resursa [3] odnosno, izvana za korisnika (čovjek ili program) nema razlike između stvarnog i virtualnog ostvarenja funkcionalnosti, ali stvarne vrijednosti i aktivnosti u virtualnoj izvedbi razlikuju se od onih prikazanih korisniku. Primjerice, stvarni operacijski sustav komunicira izravno sa sklopovljem računala, dok virtualni operacijski sustav ima za korisnika sva obilježja stvarnog sustava, ali se pokreće u drugom stvarnom sustavu. Dakle, komunikacija se ne obavlja sa sklopovljem već sa drugim sustavom. Pritom taj drugi sustav oponaša sklopovlje u komunikaciji s virtualnim sustavom. Rad sklopovlja se u ovom slučaju simulira programski pa je riječ o virtualnom sklopovlju. Virtualizacija može značiti da korisnik preko virtualnog sučelja skup računala koristi kao jedno računalo, a može i značiti da se na jednom računalu simulira rad nekoliko sustava.

Prednosti u poslovanju koje sa sobom nose tehnologije virtualizacije su:

- uštede na nabavci i održavanju IT opreme,
- oslobađanje kapaciteta postojeće IT opreme, ubrzana implementacija novih IT rješenja ili proširenja postojećih,
- jednostavnije upravljanje konfiguracijama, te razvojnim, testnim i produkcijskim okruženjima.

Uz navedena poboljanja učinkovitosti upravljanja IT infrastrukturom, virtualizacija omogućava jednostavnije planiranje i provođenje upravljanja kontinuitetom poslovanja (eng. Business Continuity), te oporavka od ispada (eng. Disaster Recovery), te jednostavniju implementaciju visokodostupnih (eng. High Availability) sustava.

Virtualizacijsko riješenje za potrebe projekta izgradnje rafinerijskog informacijskog sustava sastoji se od tri IBM HS22 blade poslužitelja, na koje je instaliran Vmware ESXi vSphere 5 poslužitelj. Poslužitelji su konfigurirani u High Availability clusteru koji osigurava redundantnost sustava gdje se u slučaju nedostupnosti jednog ESXi poslužitelja virtualni poslužitelji automatski restartaju i nastavljaju rad koristeći resurse drugog dostupnog ESX poslužitelja. vMotion tehnologija prikazana na slici 6 omogućava nesmetanu migraciju virtualnog poslužitelja sa jednog fizičkog poslužitelja na drugi.



Slika 6. Prikaz vMotion tehnologije

Ova arhitektura omogućava nesmetan rad virtualiziranih servisa za slučajeve održavanja fizičkih poslužitelja. Korištenjem redundantnih preklopnika povezanih sa svjetlovodnim vlaknima osigurava se nesmetan pristup virtualnim poslužiteljima smještenim na sustavu za skladištenje podataka u slučaju eventualnog prestanka funkcionalnosti jednog od njih. Visoka dostupnost mrežnih resursa ESXi sustava ostvarena je uparivanjem mrežnih adaptera i upotrebom redundantnih komunikacijskih preklopnika. Instaliran je Virtual Center poslužitelj koji je centralno mjesto za kontrolu, upravljanje i administraciju ESXi infrastrukture.

VI. ZAKLJUČAK

Implementacijom projekta izgradnje rafinerijskog informacijskog sustava pred stručnjake procesnih informacijskih sustava postavljen je veliki izazov. S jedne strane trebalo je zadovoljiti fizičku ograničenost prostora u koji je trebalo smjestiti sklopovlje novoizgrađenog informacijskog sustava a s druge strane trebalo se je usuditi krenuti potpuno novim putem koristeći na tržištu dostupne tehnologije virtualizacije računala.

Veličina prostorije je sa sobom povlačila problem hlađenja prostora i utrošaka energije u vrijeme kada se pred moderne i odgovorne tvrtke postavlja imperativ brzih promjena, pametnih rješenja sa ciljem ušteda i svakako smanjenja emisija stakleničkih plinova kao što je CO2 budući je poznato koliki je omjer vezan za računala i potrošnju električne energije sa potonjim.

Virtualna platforma se je veoma brzo u ekspoataciji pokazala izuzetno stabilnom, fleksibilnom, jednostavnom za nadzor i upravljanje, dodavanje novih računala poslužitelja. Postignuto je smanjenje fizičkog broja računala poslužitelja u omjeru 1:5 te smanjenje svih potrebnih infrastrukturnih resursa kao i operativnih troškova.

Prednosti primjene tehnologija virtualizacije i blade poslužitelja nad klasičnim (fizičkim) pristupom projektiranja podatkovnog centra iskazani su slijedećim aspektima:

- Vrijeme nedostupnosti rada podatkovnog centra nije postojalo niti prilikom redovnog rada na poslužiteljima jer su poslužitelji migrirani sa jednog fizičkog poslužitelja na drugi
- Ukupan trošak posjedovanja, TCO (eng. total cost of ownership) je smanjen za 20% prvenstveno smanjenjem ulaganja u poslužiteljsku i komunikacijsku opremu te potrošnju električne energije na napajanje i hlađenje poslužitelja. Povećana su ulaganja u SAN infrastrukturu za podršku virtualnoj okolini.
- Za petogodišnje razdoblje eksploatacije sustava iskazane su projicirane uštede u iznosu od 40.900 USD korištenjem virtualizacije u odnosu na tradicionalni pristup odnosno fizička računala, uštede nastale korištenjem virtualizacijom komunikacijske infrastrukture iznose 14.530 USD, itd... Ukupne uštede iznose 104.779 USD dok je period povrata investicije otprilike tri godine.
- Utjecaj na okoliš je smanjen za emisiju CO2 za 300 t u periodu od 5 godina što je ekvivalent 60 auta na cesti koji godišnje prođu 20.000 km sa prosječnom potrošnjom od 10 l/100km ili jednako učinu 1100 stabala da bi apsorbirala proizvedeni CO2.

LITERATURA

[1] INA rafinerija nafte Rijeka, Centralna kontrolna zgrada (objekt 360), Sistem sala sustava za prikupljanje procesnih podataka u realnom vremenu Izvedeni elektrotehnički projekt F.I.L.D. Projekt d.o.o.

[2] PROJEKT POVEZIVANJA DISTRIBUIRANIH KONTROLNIH SUSTAVA U RNR I IMPLEMENTACIJA SERVER PARKA Dokumentacija izvedenog stanja , Combis d.o.o

[3] CCERT-PUBDOC-2009-12-285I., Laboratorij za sustave i signale pri Zavodu za elektroničke sustave i obradbu informacija Fakulteta elektrotehnike i računarstva Sveučilišta u Zagrebu.

Nadziranje i dijagnostika rada rafinerijskih postrojenja

I. Mohler * i N.Bolf*, R.Karlović i I.Šarlija**
* Sveučilište u Zagrebu, Fakultet kemijskog inženjerstva i tehnologije, Zavod za mjerenja i automatsko vođenje procesa, Savska c.16/5a, HR-10 000 Zagreb
**Rafinerija nafte Rijeka, Automatsko vođenje procesa, Urinj bb, 51221 Kostrena
bolf@fkit.hr, i.mohler@fkit.hr, Romano.Karlovic@ina.hr, Ivana.Sarlija@ina.hr

Sažetak - U industrijskim postrojenjima veliki broj procesnih varijabli potrebno je održavati unutar definiranih granica kako bi postrojenja radila zadovoljavajuće. Nadziranje i dijagnostika procesa imaju ključnu ulogu u osiguranju da postrojenje zadovoljava radne zadatke. Ciljevi su im rutinsko praćenje procesa, otkrivanje i dijagnoza kvarova te preventivno nadgledanje. Procesne varijable se tradicionalno nadziru jednostavnim metodama praćenja granica varijabli i analizom djelovanja procesa. Statističko vođenje procesa temelji se na kontrolnim grafovima, a koristi se za praćenje kvalitete proizvoda i slične primjene gdje su periodi uzorkovanja dugački u odnosu na dinamiku procesa. U novije vrijeme sustavi za vođenje procesa postaju ključna komponenata za cjelokupni nadzor postrojenja. Više veličinske metode omogućuju bitna poboljšanja nad tradicionalnim metodama u slučajevima kad su mjerene varijable visoko korelirane. Također, suvremeni softverski alati za nadzor i optimiranje rada regulacijskih krugova omogućuju rad procesa u optimalnim uvjetima. Ovim je radom prikazana analiza rada procesa i optimiranje djelovanja regulacijskih krugova uz primjenu suvremenih metoda i specijaliziranih softverskih paketa visoke razine. Kao rezultat dana je statistička analiza rada regulacijskih krugova vrha destilacijske kolone i optimiranje vođenja industrijskog procesa destilacije tj. poboljšanje kvalitete proizvedenih lakih ugljikovodika.

I. UVOD

U industrijskim postrojenjima veliki broj varijabli i regulacijskih krugova potrebno je održavati unutar definiranih granica kako bi postrojenje radilo u optimalnim uvjetima i bez zastoja. Ukoliko neka od ključnih varijabli poprimi vrijednost izvan zadanih granica to može ozbiljno utjecati na sigurnost postrojenja, okoliš, kvalitetu produkta i isplativost proizvodnje. Odstupanje varijabli izvan zadanih granica javlja se zbog različitih uzroka, a neki od njih su: problemi s opremom, kvarovi instrumenata, loše ugođeni parametri regulatora te neuobičajeni poremećaji.

U osnovne ciljeve nadziranja procesa spadaju:

- Rutinsko nadgledanje kojim se utvrđuje jesu li procesne varijable unutar definiranih granica;
- Otkrivanje i dijagnoza nenormalnih stanja i određivanje ključnih razloga;
- Preventivno nadgledanje, čime se otkrivaju problemi i nenormalna stanja na vrijeme kako bi se poduzele potrebne mjere prije pojave većih problema i zastoja.

Analiza rada regulacijskih krugova svodi se na primjenu statističkih kriterija kojima se ocjenjuje rad pojedinog regulacijskog kruga. Pri tome se analizira dinamičko vladanje sustava, prate se varijable uključene u regulaciju, otkrivaju nestabilnosti i oscilacije u krugu, optimiraju vrijednosti parametara regulatora i kreiraju izvještaji iz kojih se mogu jasno identificirati problemi pojedinih regulacijskih krugova.

Razvojem empirijskih modela moguće je simuliranje djelovanja i optimiranje parametara standardnih, kaskadnih i unaprijednih regulacijskih krugova. Dodavanjem stvarne razine šuma u signalu (generiranje bijelog šuma) mogu se simulirati realne situacije. Polazište je identificiranje i optimiranje parametara empirijskih modela procesa i parametara regulatora. Pri tome je moguće simulirati i optimirati rad nestabilnih regulacijskih krugova, kompenzirati nelinearnosti te ukloniti probleme vezane sa zadrškom ventila i histerezom.

II. OPIS PROCESA

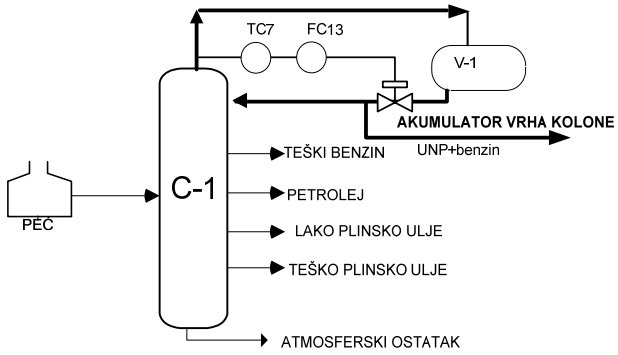
Postrojenje atmosferske destilacije namijenjeno je za preradu sirove nafte i njeno razdvajanje na produkte. Prerada sirove nafte provodi se u destilacijskoj koloni. Sirova nafta je smjesa ugljikovodika koji se pri određenim uvjetima u destilacijskoj koloni razdvajaju na osnovi temperature vrenja u skupine koje imaju tehničku primjenu. Svi produkti atmosferske destilacije zatim odlaze na sekundarna postrojenja radi poboljšanja kvalitete. Sirovu naftu iz spremnika sirovine dostavljaju pumpe i tlače preko izmjenjivača topline u odsoljivač. Odsoljena nafta preko dodatnih izmjenjivača topline dolazi u predfrakcionatorsku kolonu gdje se odvajaju lagani ugljikovodici, približno 8% sirovine. S dna destilacijske kolone tako reduciranu sirovinu se tlači preko izmjenjivača topline u peć na zagrijavanje i nakon toga sirovina ulazi u destilacijsku kolonu[1].

U koloni pri temperaturi 375°C i radnom tlaku 1.3 bara, a na osnovi temperature vrenja dolazi do razdvajanja skupina ugljikovodika na produkte. U destilacijskoj koloni C-1 dobivaju se tri bočna proizvoda: teško plinsko ulje,

lako plinsko ulje, petrolej, na vrhu kolone benzini, a na dnu ostatak. S dna kolone pumpa tlači ostatke putem rebojlera i sirovinskih izmjenjivača topline na uskladištenje bili Vakumsku destilaciju – postrojenje za obradbu atmosferskoga ostatka. Teško plinsko ulje iz stripera tlači se preko izmjenjivača topline u spremnik ili na Bender – postrojenje za rafinaciju petroleja. Nekondenzirani ugljikovodici sa vrha destilacijske kolone C-1 preko izmjenjivača topline dolaze u posudu V-1 odakle se kondenzirani ugljikovodici koriste za suhi refluks C-1 kolone, a nekondenzirani idu preko sekundarnih izmjenjivača topline u apsorber kolonu.

Plinoviti lakši ugljikovodici sa vrha apsorbera odlaze u peć na sagorijevanje. Nestabilni benzin sa dna kolone pumpa preko izmjenjivača topline tlači u debutanizer kolonu gdje se odvajaju ukapljeni plin i laki ugljikovodici od benzina. Ukapljeni plin iz posude V-2 šalje se na rafinaciju u Merox – postrojenje za uklanjanje sumpornih spojeva merkaptana iz UNP-a , dok lakši ugljikovodici odlaze u sustav plina za loženje. Teži ugljikovodici sa dna debutanizer kolone odlaze u deheksanizator kolonu gdje dolazi do razdvajanja lakog i teškog benzina. Laki benzin iz posude V-3 šalje se na uskladištenje kao sirovina za Izomerizaciju – postrojenje čija je namjena podizanje oktanskog broja benzina. Teški benzini sa dna spliter kolone skladište se u spremnike i dalje koriste kao sirovina za Platforming–postrojenje za proizvodnju motornih benzina.

U ovom radu analizirana je regulacija temperature vrha TC7 destilacijske kolone C-1 s obzirom da ona ima najveći utjecaj na kvalitetu i količinu proizvoda vrha kolone. Temperatura vrha kolone određuje sastav vršnog proizvoda, odnosno njegov kraj destilacije. Regulira se količinom vršnog pretoka iz posude V-1 pomoću FC13 koji je pri normalnom radu u kaskadi sa regulacijskim krugom TC7. S obzirom na plan proizvodnje provodi se optimizacija rada primarnog postrojenja atmosferske destilacije pri čemu se podizanjem željene vrijednosti temperature vrha kolone C-1 maksimizira iscrpak do gornje dopuštene granice temperature pazeći pritom da se ne naruši zadana kvaliteta. Imajući sve navedeno na umu izuzetno je bitno odabrati optimalne parametre regulacije krugova TC7 i FC13 kako bi se temperatura vrha destilacijske kolone održavala stabilnom unutar definiranih granica.



Slika 1. – Pojednostavljena shema kolone za atmosfersku destilaciju C-1

III. METODIKA

U specijaliziranim softverskim paketima analizira se rad regulacijskih krugova u zadanom vremenskom razdoblju te se računaju statistički kriteriji i procjenjuje oscilacija i nestabilnosti u regulacijskom krugu. Na temelju dobivenih statističkih pokazatelja moguće je identificirati probleme u radu pojedinog regulacijskog kruga te poboljšati kvalitetu proizvoda.

Pri ocjeni rada regulacijskih krugova koristi se slijedeći kriteriji:

- Integral apsolutne pogreške (IAE)

$$IAE = \int_0^\infty |e(t)|dt \tag{1}$$

- Integral kvadrata pogreške (ISE)

$$ISE = \int_0^\infty e(t)^2 dt \tag{2}$$

- Integral apsolutne pogreške otežan vremenskim koeficijentom (ITAE)

$$ITAE = \int_0^\infty t|e(t)|dt \tag{3}$$

pri čemu e(t) predstavlja razliku između zadane vrijednosti i izmjerene vrijednosti.

Glavni cilj regulacije je da vrijednost odstupanja, odnosno pogreške e(t) u što kraćem vremenu bude što manja, što nadalje znači da se proces nalazi u željenom stanju. Kriterij ITAE ima prednost pred IAE i ISE kriterijem jer se njime otkrivaju dugotrajne pogreške. Za otkrivanje velikih pogreški pouzdaniji je ISE kriterij [2]. Ostali kriteriji koji govore o kvaliteti djelovanja regulacijskog kruga su:

- *Čvrstoća regulacije* (engl. *control tightness*) – učestalost prelaska vođene veličine (PV) preko radne točke (SP). Pokazatelj čvrstoće ugođenog regulatora,
- *Disbalans* (engl. *imbalance*) – omjer vremena vođene veličine provedenog s obiju strana radne točke. Indicira loše vođenje, veliki vanjski poremećaj ili probleme s regulacijskim ventilima,
- *Nestabilnost* (engl. *unstable*) – pojava kontinuirane oscilacije. Pokazatelj agresivnog djelovanja regulatora,
- *Rastuća oscilacija* (engl. *hunting*) – pojava rastuće oscilacije u regulacijskom krugu. Pokazatelj agresivnog djelovanja regulatora,
- *Zasićenje* izlaza iz regulatora (engl. *Saturation* OP) – prikazuje koliko je vremena (u postocima - %) izlaz iz regulatora bio iznad 90% ili ispod 10%,
- *Zasićenje* PV-a (engl. *saturation* PV) – prikazuje koliko je vremena (u postotcima - %) vođena veličina bila unutar 3% od njezine gornje ili donje granice,

- *Pik* OP-a (engl. *spike* OP) – detekcija najveće promjene izlaza iz regulatora. Pokazatelj pojačanja regulatora,
- *Pik* PV-a (engl. *spike* PV) - prikaz najveće promjene vođene veličine. Pokazatelj nestabilnosti ili poremećaja u procesu,
- *Zamrznuti* PV (engl. *frozen* PV) – brojač koliko je puta singal vođene veličine bio zamrznut,
- *Vrijeme rada* (engl. *onstream factor*) – prikazuje koliko je vremena (u postocima %) regulator u automatskom, kaskadnom ili ručnom režimu rada,
- *Intervencije* (engl. *intervene*) – brojač prelaska iz jednog režima rada u drugi (man/auto/cascade),
- *Kretanje ventila* (engl. *rope lenght*) – prikaz ukupnog kretanja ventila tijekom perioda,
- *Kolebanje ventila* (engl. *vacillation*) - brojač promjene smjera rada ventila,
- *Razina šuma* (engl. *noise level*) - prikaz količine bijelog šuma u signalu vođene veličine,
- *Preklapanje* (engl. *match*) – prikaz tagova/regulatora u sustavu kojima se preklapaju frekvencije,
- *Asimetrija* (engl. *skew*) – omjer pozitivnog i negativnog integrala pogreške. Pokazatelj tromosti rada djelovanja regulatora, pojava poremećaja u procesu ili problemi na ventilu,
- *Tromost* – prikaz omjera akumulirane promjene radne točke i akumulirane promjene vođene veličine. Pokazatelj tromosti ili agresivnosti djelovanja regulatora.

IV. REZULTATI I RASPRAVA

Primjenom specijaliziranih softverskih paketa provedena je analiza rada dvaju kaskadno povezanih regulacijskih krugova vrha destilacijske kolone, slika 1. Regulator temperature vrha kolona (primarni krug) kaskadno je povezan s regulatorom protoka gornjeg refluksa destilata u kolonu (sekundarni krug). Analiza rada temelji se na podacima iz rafinerijske baze podataka: trenutne vrijednosti vođenih varijabli (PV) temperatura i protoka, radne točke (SP) te izlazi iz regulatora (OP). Ostale važne informacije su: način rada regulacijskog kruga, mjerno područje i mjerni opseg te vrijednosti konstanti regulatora za svaki od regulacijskih krugova, tablica 1. Također je potrebno poznavati informacije o frekvenciji prikupljanja podataka iz baze podataka, algoritam regulacije i vrijeme proračuna.

Prilikom analize, tablica 2, ustanovljeno je da su integrali pogreške (IAE, ISE i ITAE) kod regulacijskog kruga protoka vrlo mali, dok vrijednost apsolutne pogreške za regulacijski krug temperature treba razmotriti [3]. Kriterij čvrstoće ukazuje da se svi mjerljivi i nemjerljivi poremećaji u procesu uspješno potiskuju. Kod regulacijskog kruga temperature čvrstoća regulacije je ponešto manja što može ukazati na sporo regulacijsko djelovanje uzrokovano velikom integracijskom konstantom.

Tablica 1 – Osnovne informacije o regulacijskim krugovima

Regulacijski krug	**FC 13**	**TC 7**
Broj podataka	5759	5759
Način rada	Kaskadni	Automatski
Pojačanje regulatora ($m^3/h°C$)	0,9	3
Integracijska konstanta (min)	1,2	6
Derivacijska konstanta (min)	0	0,4
Vrsta algoritma regulatora	A	B
Vrijeme uzorkovanja (s)	15	15
Vrijeme proračuna regulatora (s)	1	1
Radno područje	0-278m^3/h	0-300°C

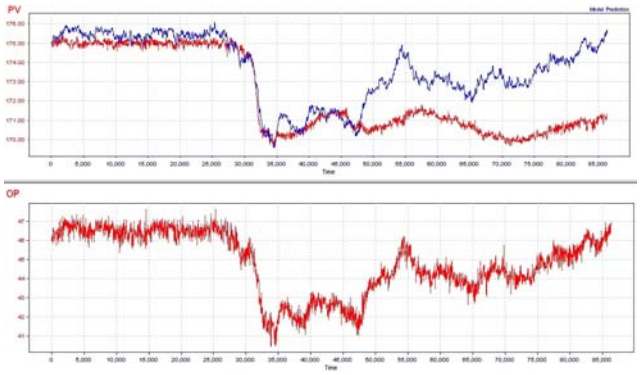
Iz kriterija disbalansa, nestabilnosti i rastuće oscilacije može se zaključiti da regulacijski krugovi ne osciliraju toliko da bi mogli uzrokovati probleme u radu. Spektralnom analizom signala i kriterijem preklapanja vidljivo je da je frekvencija rada regulacijskog kruga protoka tri puta veća od frekvencija rada regulacijskog kruga temperature, što je očekivano pri ovakvom načinu regulacije. Iz analize šuma može se vidjeti da je razina šuma dosta velika za oba regulacijska kruga što može prouzročiti probleme u regulaciji. Preporuča se filtriranje signala. Kod regulacijskog kruga temperature treba obratiti pozornost na razinu zamrznutosti signala koja može biti uzrok neželjenog djelovanja regulatora. Iz kriterija kretanja i kolebanja ventila, zasićenja i pika OP-a može se zaključiti da ventil ispravno radi i na njemu nema mehaničkih problema u radu. Intervencija operatera nije bilo, što je znak da automatska regulacija djeluje dobro. Zaključno se može reći da je regulacijski krug protoka dobro ugođen, dok je za regulacijski krug temperature preporučljivo optimirati parametre regulatora kako bi se donekle poboljšalo djelovanje ovog kruga.

Stoga je slijedeći korak određivanje parametara regulatora temperature. Pri tome je potrebno simulirati regulacijski sustav, odrediti dinamičke karakteristike i optimirati parametre regulacijskog kruga. Kod simuliranja se upotrebljava bijeli šum za simuliranje stvarne razine šuma u signalu. Identificiranje empirijskog modela procesa bitan je korak za određivanje optimalnih vrijednosti parametara regulatora.

Pri identificiranju modela uzet je isti skup podataka kao i za analizu rada regulacijskog kruga temperature. Slika 2 prikazuje djelovanja regulatora. Regulator djeluje direktno (pozitivna promjena izlaza regulatora daje pozitivnu promjenu u temperaturnom odzivu). Shodno tome statička osjetljivost empirijskog modela imat će pozitivnu vrijednost. Pri analizi pretpostavljeno je da se radi o procesu prvog reda.

Tablica 2 – Pokazatelji rada regulacijskih krugova

Regulacijski krug:	**FC 13**	**TC 7**
IAE (%)	0.0459	**0.2187**
ISE (%)	0.0807	0.004
ITAE (%)	0.0065	0.0044
Varijanca (°C)^2	0.0807	0.004
Standardna devijacija (°C)	0.2841	0.0634
Čvrstoća regulacije	132.7105	**44.6328**
Disbalans	1.0010	1.0444
Asimetrija	1.0271	**1.1258**
Tromost	0.6462	0.0461
Nestabilnost	0	0
Porast oscilacija	0	0
Spektralna analiza	9.5714	2.6377
Preklapanje	1	3
Razina šuma	**5**	**5**
Zasićenje OP (%)	0	0
Pik OP (%)	1.8402	0.8720
Zasićenje PV (%)	0	0
Pik PV (°C)	0	0
Zamrznutost PV	0.0695	**6.2164**
Kretanje ventila	0.2385	0.1302
Kolebanje ventila	0.5886	0.2480
Proporcionalno djelovanje(%)	54.1031	37.0035
Integracijsko djelovanje (%)	45.8969	16.9546
Derivacijsko djelovanje (%)	0	46.0419
Intervencije	0	0
Vrijeme rada (%)	100	100



Slika 2. – Identificiranje modela procesa

Na y-osi nalaze se PV(°C) i OP(%) regulacijskog kruga temperature, dok x-os prikazuje vrijeme u minutama. U prvom dijelu grafičkog prikaza vidljivo je da empirijski model procesa (plava linija) dobro opisuje promjenu temperature (crvena linija). Donji graf prikazuje kretanja izlaza regulatora. U daljnjem dijelu jasno je izražen poremećaj koji se javlja u sustavu te empirijski model procesa odstupa od stvarnih vrijednosti. Poremećaj u sustavu javlja se zbog promjene temperature gornjeg refluksa ili zbog promjena temperature okoline.

Za identificiranje parametara modela primjenjena je geometrijska metoda bez filtriranja podataka [4]. Ova metoda se primjenjuje za procese u kojima nema ili su prisutni manji nemjerljivi poremećaji. Dobiveni parametri modela 1. reda su:

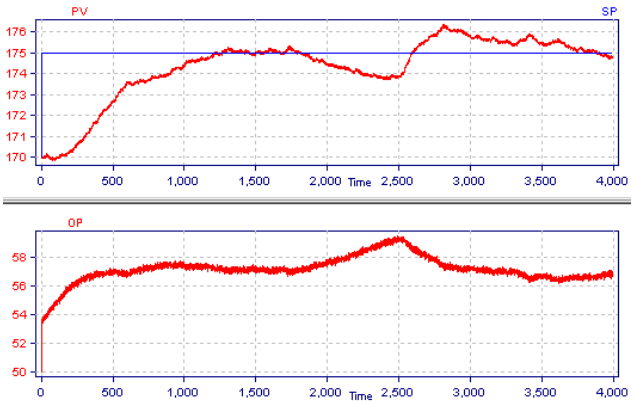
- statička osjetljivost: k = 0,97 m^3/h°C
- mrtvo vrijeme: θ= 23 s
- vremenska konstanta procesa: τ= 404 s

Nakon određivanja parametara modela procesa pristupilo se optimiranju parametara PID regulatora. Osnovne informacije i početni parametri regulacijskog kruga, tablica 1, uzeti su kao polazište za optimiranje parametara regulacijskog kruga. Osim osnovnih informacija, za optimiranje rada regulacijskog kruga potrebno je identificirati poremećajne veličine koje utječu na rad kruga, te modelirati šum koji se javlja pri mjerenju. Analizom procesa utvrđeno je da je poremećajna veličina koje utječe na rad kruga temperatura povrata destilata u kolonu (TI 146), a mjereni signal (TC 7) je šumovit s rasponom od 0,2°C.

Za optimiranje parametara PID regulatora, u ovom slučaju, rabi se kriterij integrala kvadrata pogreške (ISE) zbog danih ograničenja vezano s reguliranom temperaturom [5]. Slika 3 prikazuje djelovanje PID regulatora na zadanu promjenu radne točke (SP). Na gornjem grafu plavom bojom prikazana je promjena radne točke sa 170°C na 175°C, dok je temperaturni odziv (PV) označen crvenom bojom. Vrijeme potrebno da temperatura dosegne radnu točku procesa je 20 min. Uzrok većeg odstupanja od radne točke može biti jedino poremećaj u procesu. Na donjem grafu crvenom linijom prikazan je izlaz regulatora. Promjena radne točke od 5°C uzrokuje maksimalnu promjena izlaza regulatora od 3,5%. Nakon promjene radne točke intenzivno je proporcionalno djelovanje, zatim dominira integracijsko djelovanje čiji je zadatak smanjiti regulacijsko odstupanje. Oscilacije izlaza regulatora javljaju se zbog derivacijskog djelovanja koje reagira na šum u procesnom signalu. U vremenu od 25 do 60 minute prikazano je djelovanje regulatora na utjecaj nemjerljivog poremećaja, konkretno promjene temperature protoka refluksa. Nakon 60. minute regulator uspješno potiskuje poremećaj i nastavlja održavati temperaturu vrha kolone.

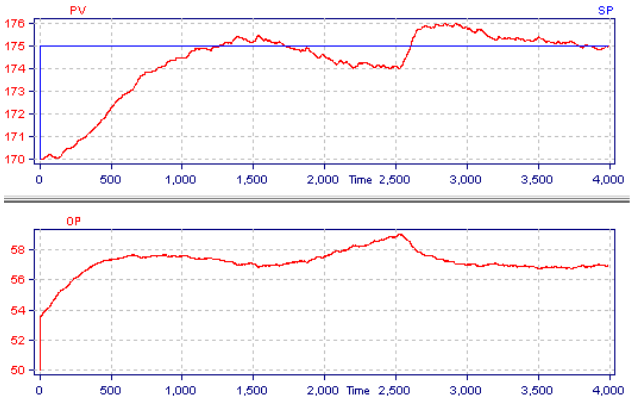
Parametri PID regulatora dobiveni postupkom optimiranja su:

- pojačanje regulatora: K_C = 2,10 m^3/h°C
- integracijska konstanta: τ_i = 4,50 min
- derivacijska konstanta: τ_d = 0,05 min



Slika 3. – Odziv vođene temperature i djelovanja PID regulatora nakon ugađanja parametara regulatora

Šum koji se javlja pri mjerenju temperature potrebno je filtrirati. Stoga se primjenjuje filtarska konstanta iznosa $\tau_f = 0{,}06$ min. Na slici 4 vidljivo je da se primjenom filtra ostvaraje uglađeniji izlaz regulatora što produljuje vijek trajanja regulacijskog ventila.



Slika 4. – Posljedica filtriranja signala temperature na izlaz PID regulatora

V. ZAKLJUČAK

U ovom radu statistički su analizirani i optimirani ključni regulacijski krugovi na postrojenju atmosferske destilacije Rafinerije nafte Rijeka. Podaci na temelju kojih je provedena analiza prikupljeni su iz internih sustava i baza podataka koje obuhvaćaju različite vremenske periode.

Prikazani su statistički pokazatelji kvalitete djelovanja regulacijskih krugova putem proračuna odgovarajućih kriterija kvalitete djelovanja.

Zaključno se može istaknuti da je primjenom metoda statističke analize i specijaliziranih suvremenih softverskih alata moguće nadzirati i dijagnosticirati stanje procesa i regulacijskih krugova. Time se omogućava kvalitetnije vođenje industrijskih postrojenja, poboljšanje kvalitete proizvedenih lakih ugljikovodika te pravovremene intervencije kako bi se postrojenja održala sigurnim i ekonomičnim.

LITERATURA

[1] N. Prijić, "Tehnološki priručnik", 2. izdanje, Rafinerija nafte Rijeka, Rijeka, 2003.

[2] Seborg, D. E., T. F. Edgar & D. A. Mellichamp (2010). Process Dynamics and Control, 2nd ed., John Wiley & Sons; New York.

[3] Statistica Electronic Manual, version 7.1, Statsoft Inc., Tulsa, 2006.

[4] Ljung, L., System Identification: Theory for the User, 2nd ed., Prentice Hall, New Jersey, 1999.

[5] http://www.mathworks.com/help/toolbox/ident/ref/idnlhw.html Statistica Electronic Manual, version 7.1, Statsoft Inc., Tulsa, 2006.

[6] N. Bonavita, J. C. Bovero, R. Martini, "Tuning loop: control performance and diagnostics", ABB Process Solutions & Services

[7] T. Olsen, N. Ito, „Implement an Effective Loop Tuning Strategy", 2013

Korištenje zalihosne opreme poslužitelja u VHF DSC sustavu Plovputa

J. Vrlić*, N. Grgić** i I.Ružić**
* Plovput d.o.o., Split, Hrvatska
** Sveučilište u Splitu, Sveučilišni odjel za stručne studije, Split, Hrvatska
josip.vrlic@plovput.hr
nikola.grgic@oss.unist.hr
ivica.ruzic@oss.unist.hr

Sažetak - VHF DSC (Digital Selective Call) sustav Plovputa je dio Svjetskog pomorskog sustava za pogibelj i sigurnost (GMDSS) na moru. S obzirom da navedeni Sustav ima funkciju spašavanja ljudskih života i imovine na moru od iznimnog je značaja da Sustav ima što veću robustnost, raspoloživost i pouzdanost. Navedeni zahtjevi, su riješeni raspodijeljenom arhitekturom samog Sustava, ali i najvažnije, implementacijom zalihosne opreme poslužitelja na kojoj se nalaze poslužiteljski primjenski programi i baze podataka. U ovom radu je opisan način implementacije zalihosne opreme poslužitelja, statistička raspoloživost samog Sustava, te prijedlog poboljšanja istog.

I. UVOD

DSC (Digital Selective Call) sustav je dio svjetskog pomorskog sustava za pogibelj i sigurnost GMDSS (Global Maritime Distress and Safety System) [1], a koji je važan dio IMO (Internacional Maritime Organisation) SOLAS (Safety Of Life At Sea) konvencije [2].

ITU (International Telecommunication Union) preporuka ITU-R M.541-9 [3] definira operativne procedure za korištenje DSC opreme, dok preporuke ITU-R M.493-13 [4] i dodatak ITU-R M.821-1 [5] definiraju tipove DSC poruka i način kako se iste kodiraju.

DSC se koristi kao primarni sustav za uzbunjivanje na VHF (Very High Frequency), MF (Medium Frequency) i HF (High Frequency) frekvencijskim pojasima.

S obzirom da je Republika Hrvatska ugovornica SOLAS konvencije, preuzela je obvezu implementacije GMDSS sustava na morskom teritoriju, koji spada u GMDSS A1 područje plovidbe (20-30 NM od obale) za što je dovoljno koristiti VHF frekvencijski pojas. VHF kanal za DSC sustav je VHF Ch70 (156.525 MHz)

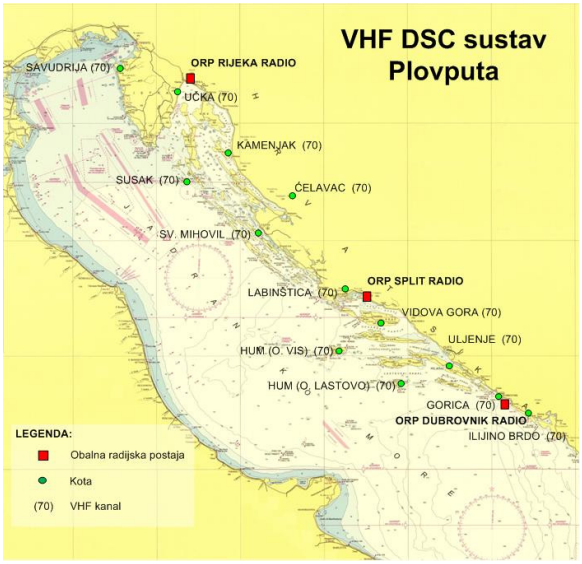
Stoga je tvrtka Plovput d.o.o. Split koja je u stopostotnom vlasništvu Republike Hrvatske, još tijekom 1995. g. kupila i instalirala VHF DSC sustav talijanskog proizvođača Cimat S.p.A. U međuvremenu, 2008. g., zbog otežanog održavanja računalne opreme i programske podrške uslijed napretka tehnologije, ali i zbog prestanka postojanja proizvođača (Cimat), tvrtka Plovput je odlučila nabaviti/izraditi novi VHF DSC sustav. Novi VHF DSC sustav je nabavljen/izrađen na način da je tvrtka Plovput izradila Specifikaciju sustava, te po njoj naručila izradu dijela programske podrške (simulatori, baza podataka, poslužiteljski primjenski program i operatorski primjenski program), a preostali je dio programske podrške izradila samostalno (koder i dekoder DSC poruka, alarmne jedinice i primjenski program za upravljanje VHF baznim stanicama). Tijekom 2011. g. obavljena je nadogradnja programske podrške VHF DSC sustava Plovputa, povećanjem funkcionalnosti istog.

Prilikom odabira računalne opreme, vodilo se računa da se izaberu renomirani proizvođači, ali i da tehnologija proizvoda bude prilagođena uvjetima okoline.

U ovom radu opisan je VHF DSC sustav Plovputa i opis implementacije zalihosne opreme poslužitelja. Posebno je prikazana statistička raspoloživost VHF DSC sustava Plovputa kroz kontinuirani rad od 2008. g..

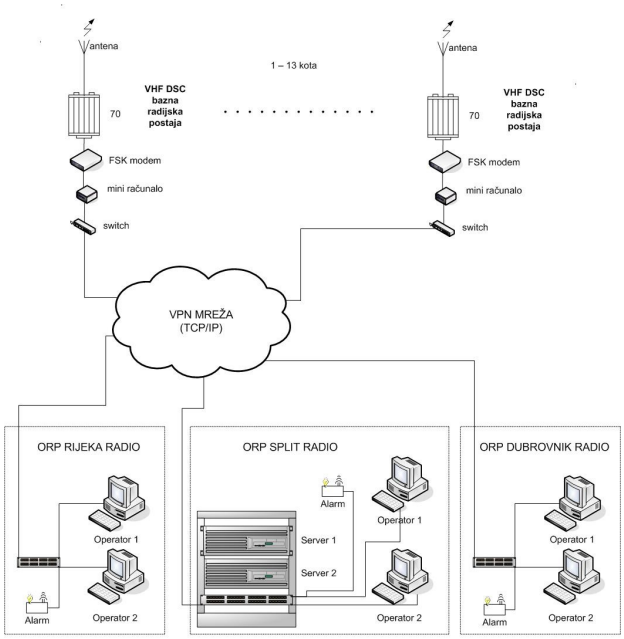
II. OPIS VHF DSC SUSTAVA PLOVPUTA

VHF DSC sustav Plovputa uključuje opremu smještenu na tri obalne radijske postaje, te na trinaest kota (po jedna radijske postaja na svakoj koti: VHF DSC kanal 70) – slika 1. Obalne radijske postaje su s VHF DSC baznim radijskim postajama povezane VPN mrežom baziranoj na TCP/IP tehnologiji – slika 2.



Slika 1. Prikaz obalnih radijskih postaja, kota i VHF DSC kanala

Sponzor: Plovput d.o.o. Split



Slika 2. Oprema na obalnim radijskim postajama, kotama i način povezivanja

Kote s VHF radijskim postajama su geografski raspoređene po Jadranu kako bi svako morsko područje bilo minimalno dvostruko pokriveno VHF radijskim signalom, s tim da je većina područja i višestruko pokrivena. Na taj način osigurana je robusnost VHF DSC sustava Plovputa, jer kvarom jedne kote, isti i dalje može potpuno ispravno funkcionirati.

Osnovne funkcije VHF DSC sustava Plovputa su:

- kontinuirano bdijenje za slučaj pogibelji na VHF DSC kanalu 70;
- pozivanje i primanje poziva u slučaju pogibelji, hitnosti, sigurnosti i rutinskih poziva na relaciji brod-kopno i kopno-brod.

Tehničke mogućnosti VHF DSC sustava Plovputa su:

- zalihost sustava, odnosno neovisan rad oba poslužitelja koji rade paralelno uz međusobnu sinkronizaciju baze podataka, kao i neovisnost rada svakog operatorskog računala i alarmne jedinice;
- osiguranje pristupa svim VHF DSC radijskim uređajima sa svih operaterskih mjesta, bez obzira na opterećenost sustava;
- kontinuirano praćenje rada svih dijelova sustava, uključivši i provjeru komunikacije s VHF DSC radijskim uređajima na kotama, te automatsko dojavljivanje greške u sustavu;
- jednostavno održavanje pojedinih uređaja sustava uz kontinuirani rad sustava;
- mogućnost daljinskog pristupa sustavu, a u svrhu održavanja programske podrške cjelokupnog sustava;
- mogućnost jednostavnog proširenja rada sustava dodavanjem operaterskih mjesta i VHF DSC radijskih uređaja na kotama.

A. VHF DSC sustav Plovputa na obalnim radijskim postajama

Oprema na VHF DSC sustava na obalnim radijskim postajama uključuje:

- dva operaterska mjesta opremljena osobnim računalom s monitorom, tipkovnicom i mišem;
- alarmnu jedinicu za zvučni i svjetlosni alarm o primitku DSC poruke,

s dodatkom da se na obalnoj radijskoj postaji Split radio dodatno nalazi:

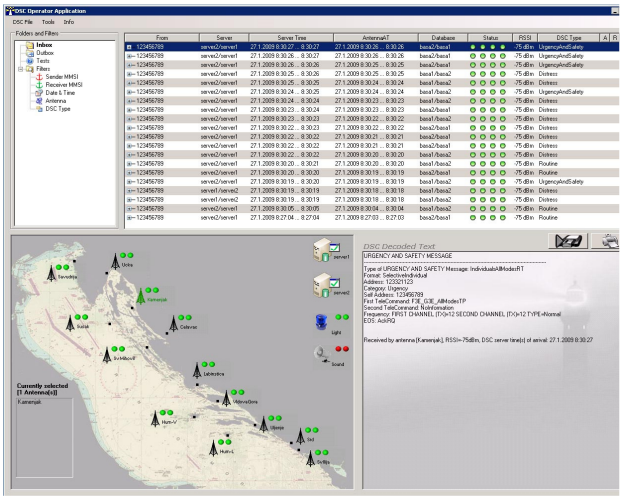
- upravljačko-komunikacijski ormar u koji su smještena dva poslužitelja.

Na operatorskom računalu je cijelo vrijeme prikazano:

- stanje svih objekata u VHF DSC sustavu uz slikoviti prikaz smještaja kota na karti Jadranskog mora;
- pregled svih primljenih i poslanih VHF DSC poruka;
- rezultati testiranja objekata u VHF DSC sustavu;
- stanje dostupnosti svih objekata u VHF DSC sustavu (kota, poslužiteljski primjenski program, baza podataka, alarm).

Programska podrška na operatorskim računalima (operatorski primjenski program) ima sljedeće mogućnosti – slika 3.:

- pozivanje jedne, grupe ili svih radijskih postaja, radijske postaje unutar nekog geografskog područja, te primanje poziva na VHF DSC kanalu;
- testiranje kota s ciljem dobivanja informacije o stanju opreme na koti;



Slika 3. Prikaz osnovnog prozora programske podrške VHF DSC sustava

- testiranje VHF DSC sustava uz pomoć stvarnih VHF DSC poruka;
- pretraživanje primljenih ili poslanih VHF DSC poruka po raznim kriterijima;
- ispis i pohranjivanje VHF DSC primljenih i poslanih poruka;
- ispis i pohranjivanje pogrešaka u radu svih objekata u VHF DSC sustavu radi sustavnog praćenja rada VHF DSC sustava;
- dobivanje informacije o RSSI (Received Signal Strength Indication) vrijednosti primljenih VHF DSC poruka u svrhu određivanja pozicije odakle je poslana VHF DSC poruka;
- prikaz informacije o emitiranoj i reflektiranoj snazi za sve poslane VHF DSC poruke;

Sva oprema VHF DSC sustava spojena je na uređaje za neprekidno napajanje, s tim da se u slučaju dužeg prekida električne energije napajanje prebacuje s javne električne mreže na napajanje s diesel agregata.

B. VHF DSC sustav Plovputa na kotama

Oprema VHF DSC sustava na kotama uključuje:

- baznu radijsku postaju za VHF DSC kanal 70;
- antenski sustav;
- mini računalo;
- komunikacijske uređaje;
- uređaj za neprekidno napajanje.

C. Komunikacija između VHF DSC sustava Plovputa na obalnim radijskim postajama i kota

VHF DSC sustav na obalnim radijskim postajama i kotama je međusobno povezan VPN mrežom baziranoj na TCP/IP tehnologiji – tablica 1., čime je omogućen daljinski pristup svakom objektu unutar VHF DSC sustava.

TABLICA I. PREGLED LOKACIJA SA PRIKAZOM BRZINE PRIKLJUČNE TOČKE VPN MREŽE

	Lokacija	***Brzina priključne točke***
1.	ORP[1] Rijeka radio	1 Mbit/s
2.	ORP Split radio	2 Mbit/s
3.	ORP Dubrovnik radio	1 Mbit/s
4.	PS[2] Savudrija	256 Kbit/s
5.	RRP[3] Učka	256 Kbit/s
6.	PS Susak (o. Susak)	256 Kbit/s
7.	RRP Kamenjak	256 Kbit/s
8.	RRP Sv. Mihovil (o. Ugljan)	256 Kbit/s
9.	RRP Čelavac	256 Kbit/s
10.	RRP Labinštica	256 Kbit/s
11.	VP[4] Hum (o. Vis)	256 Kbit/s
12.	RRP Vidova Gora (o. Brač)	256 Kbit/s
13.	RRP Uljenje (po. Pelješac)	256 Kbit/s
14.	RRP Hum (o. Lastovo)	256 Kbit/s
15.	ORP Gorica Sv. Vlaha	spojeno na ORP Dubrovnik radio
16.	RRP Sv. Ilija	256 Kbit/s

[1]ORP – Obalna Radijska Postaja
[2] PS – Pomorski Svjetionik
[3] RRP – Radio Relejna Postaja
[4] VP – Vojna Postaja

III. OPIS IMPLEMENTACIJE POSLUŽITELJSKE RAČUNALNE OPREME U VHF DSC SUSTAVU PLOVPUTA

Prilikom definiranja specifikacije VHF DSC sustava Plovputa, najveća je dilema bila odabrati primjeren način rada poslužiteljske računalne opreme u svrhu postizanja zalihosti.

Nužan uvjet za postizanje zalihosti računalne opreme poslužitelja je da postoje minimalno dva poslužitelja. U VHF DSC sustavu Plovputa odabrano je rješenje s točno dva poslužitelja. Razmatrana su sljedeća rješenja njihove konfiguracije:

- sekundarni poslužitelj ima funkciju hladne pričuve – „cold backup“;
- sekundarni poslužitelj ima funkciju tople pričuve – „warm backup“;
- sekundarni poslužitelj ima funkciju vruće pričuve – „hot backup“;
- oba poslužitelja rade paralelno, odnosno ne postoje primarni i sekundarni poslužitelj.

Pod hladnom rezervom podrazumijeva se da sekundarni poslužitelj bude instaliran i konfiguriran za rad, ali stoji isključen do trenutka greške na primarnom poslužitelju, kada ga je potrebno ručno uključiti, kako bi preuzeo funkciju primarnog poslužitelja. Ovo rješenje je neprikladno za VHF DSC sustav Plovputa jer isti među ostalim služi i za spašavanje ljudskih života te je neprimjereno da je potreban ljudski faktor za oporavak Sustava zbog greške primarnog poslužitelja.

Topla rezerva, je sličnog načina rada, kao i hladna rezerva, osim što je sekundarni poslužitelj potrebno povremeno uključiti, kako bi se sinkronizirali podaci (npr. baze podataka) s primarnim poslužiteljem. Razlog neprikladnosti ovog rješenja su slični kao kod hladne rezerve.

Rad sekundarnog poslužitelja, kao vruće rezerve omogućuje da se automatski preuzme cjelokupna funkcionalnost sustava od strane sekundarnog poslužitelja uslijed greške primarnog poslužitelja. Međutim, razlog zašto je procijenjeno da ovaj način rada nije najpogodniji za VHF DSC sustav Plovputa stoji u činjenici da je za ovakav način rada poslužitelja potrebno posebno definirati i programirati uvjete kada sekundarni poslužitelj treba preuzeti funkcioniranje sustava od primarnog poslužitelja, odnosno vratiti funkcioniranje sustava primarnom poslužitelju. Takvi uvjeti bi bili kompleksni, jer bi ovisili o velikom broju parametara (poslužiteljski primjenski program, operacijski sustav, baza podataka, poslužiteljska računalna oprema, komunikacijska oprema), te bi sigurno unosili nepouzdanost u funkcioniranje VHF DSC sustava Plovputa.

U konačnici, je odlučeno koristiti rješenje da oba poslužitelja rade paralelno, potpuno neovisno jedan o drugome. Takav način rješenja pojednostavnjuje arhitekturu i način implementacije pojedinog poslužitelja,

ali uvjetuje da svi ostali objekti u sustavu (npr. operatorski primjenski program, kote, alarmi) povećaju svoju složenost jer moraju istovremeno komunicirati s dva poslužiteljska primjenska programa i baze podataka. Navedena složenost ostalih objekata u sustavu nije opasna jer je raspodijeljena na više lokacija, odnosno svaki objekt je višestruko zalihostan (postoji više operatorskih računala, alarma i kota).

Na svakom od poslužitelja se nalazi poslužiteljski primjenski program i baza podataka u kojoj su pohranjeni svi podaci i događaji generirani u VHF DSC sustavu Plovputa (slanje i primanje VHF DSC poruka, testiranje objekata, testiranje komunikacije, itd.).

Kako bi se povećala sigurnost spremljenih podataka u bazi podataka, ugrađena je replikacija baze podataka svakog poslužitelja u bazu podataka drugog poslužitelja, što je omogućilo da se svaki podatak u normalnom radu Sustava nalazi zapisan na četiri mjesta:

1. bazi podataka na prvom poslužitelju;
2. bazi podataka na drugom poslužitelju;
3. replikaciji baze podataka prvog poslužitelja u bazi podataka na drugom poslužitelju;
4. replikaciji baze podataka drugog poslužitelja u bazi podataka na prvom poslužitelju.

Za replikacijski mehanizam korištena je P2P (Peer-to-Peer) transakcijska replikacija, što je zapravo inačica transakcijske replikacije koja je predstavljena u paketu SQL Server 2005 [7]. Kod standardne transakcijske replikacije, te kod transakcijske publikacije sa obnovljivim pretplatama (Transactional Publication with Updatable Subscriptions) relacijski odnos je uvijek hijerarhijske prirode, ali u P2P formatu, relacijski odnos dvaju ili više poslužitelja temelji se na jednakopravnom parterskom modelu (topologija se pritom često naziva i master-to-master, budući da su svi poslužiteljski čvorovi jednaki).

Posao upravljanja bazama podataka, raspodijeljen na više poslužitelja po P2P modelu omogućuje povećanje performansi čitanja rezultata upita. Budući da svaki čvor u P2P topologiji igra ulogu i poslužitelja (eng. Server) i publikatora (eng. Publisher), promjene koje se dogode na bilo kojem čvoru neće se aplicirati na ostalim čvorovima više od jedanput. P2P transakcijska replikacija iznimno je pogodna za sustave koji za svoj rad zahtijevaju „hot backup“ baze podataka, te za sustave koji zahtijevaju visoku dostupnost (raspoloživost) mehanizama za upravljanje bazom podataka. Zbog navedenih činjenica, opisani model replikacije je korišten u VHF DSC sustavu Plovputa.

Odabrani način paralelnog rada oba poslužitelja može prouzročiti privremeni nedostatak povijesti poruka, zbog greške drugog poslužitelja za vrijeme obavljanja replikacije nakon povratka u rad prvog poslužitelja. Takva situacija je zanemariva, jer je vremensko trajanje replikacije vrlo kratko zbog brze komunikacijske veze između poslužitelja i relativno male količine podataka koja se generira u Sustavu. Također, vjerojatnost takvog scenarija je vrlo slična vjerojatnosti da se na oba poslužitelja istovremeno dogodi kvar.

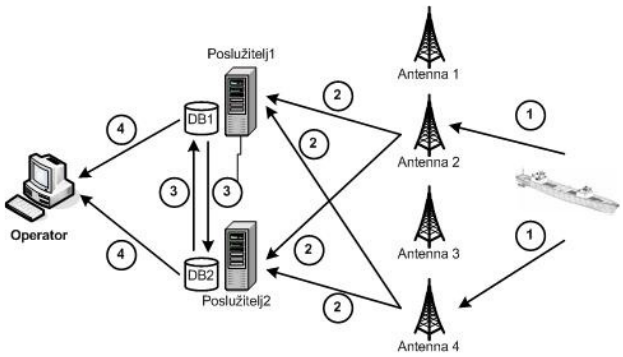
Paralelnim radom oba poslužitelja omogućeno je:

- višestruka zalihost svih podataka u Sustavu;
- dvostruka zalihost rada poslužiteljski primjenski program Sustava;
- održavanje poslužiteljske računalne opreme bez utjecaja na rad Sustava;
- održavanje baza podataka i poslužiteljskog primjenskog programa bez utjecaja na rad Sustava

Kao što je ranije navedeno paralelnim radom oba poslužitelja složenost je raspodijeljena na ostale objekte u VHF DSC sustavu Plovputa, a prikazana je u opisu primanja i slanja VHF DSC poruka.

A. Primanje VHF DSC poruka – slika 4.

Svaka emitirana VHF DSC poruka (1) u sustavu može biti primljena sa jedne ili više VHF radijskih stanica na koti što ovisi o vremenskim uvjetima, području pokrivenosti signalom te geografskoj lokaciji izvora signala. Svaka kota, nakon što primi emitiranu poruku, pošalje odgovarajuću poruku na oba poslužitelja (2). Po primitku poruke poslužiteljska aplikacija (svakog od poslužitelja) zapisuje sve relevantne informacije u bazu podataka koja se nalazi na istom poslužitelju. Replikacijski mehanizam osigurava da se sve promjene u jednoj bazi podataka prenesu u drugu bazu podataka zadržavajući konzistentnost podataka, bez posredstva vanjske aplikacijske logike (3). Osim upisa informacija o primljenim porukama u bazu podataka, poslužiteljska aplikacija je zadužena i za aktiviranje odgovarajućih alarma. Procedura u operaterskoj aplikaciji koja periodički provjerava prisutnost eventualno novo pristiglih informacija u bazi podataka se koristi za prikaz primljene VHF DSC poruke na zaslonu operatorskog računala (4).

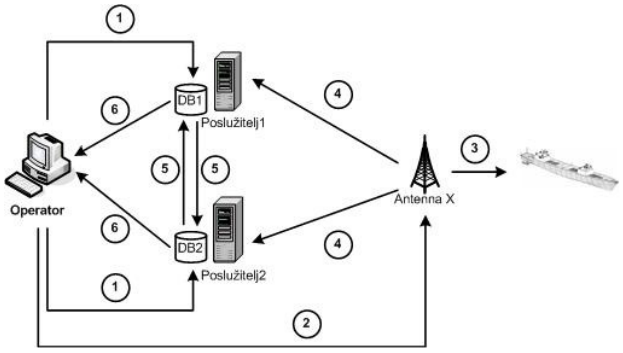


Slika 4. Prikaz primanja VHF DSC poruke

B. Slanje VHF DSC poruke – slika 5.

Prilikom slanja VHF DSC poruke, operater putem sučelja operatorske aplikacije zadaje nalog slanja VHF DSC poruke s određene kote (Antenna X). Procedura započinje direktnim upisom poruke koja se namjerava poslati u posebnu tablicu obje baze podataka koja ne podliježe replikacijskom mehanizmu (1). Prethodni korak je nužan zbog pamćenja informacije o identifikatoru

operatera koji pokreće proces slanja VHF DSC poruke, te će upisani podaci izbrisati iz posebne tablice baze podataka, nakon uspješnog slanja VHF DSC poruke. Nakon toga, operatorska aplikacija šalje dotičnu VHF DSC poruku na definiranu kotu bez posredstva poslužitelja(2). Po primitku poruke od operatorske aplikacije, kota je emitira u sustav putem VHF radijske postaje (3). Nakon obavljenog procesa emitiranja poruke, informacije o poslanoj poruci kota povratno šalje poslužiteljskoj aplikaciji na oba poslužitelja (4). Tada poslužitelji ubacuju dobivenu poruku u bazu podataka, a baza može/zna razlučiti da se radi o potvrdi emitiranja VHF DSC poruke koju je inicijalno kreirao dotični operater. Replikacijski mehanizam će osigurati da se zapisi u obje baze podataka na poslužiteljima razmjene (5). Promjena u bazama podataka biti će „objavljena“ operatorskoj aplikaciji koja će ih operateru učiniti vidljivima (6).



Slika 5. Prikaz slanja VHF DSC poruke

IV. STATISTIČKI PODACI O RADU VHF DSC SUSTAVA PLOVPUTA

Rad VHF DSC Plovputa se prati i vrednuje kroz kvartalne, polugodišnje i godišnja izvješća o radu tvrtke Plovput, gdje se posebno navodi broj kvarova na VHF DSC sustavu kada se isti nije mogao koristiti.

U tablici II. su navedeni svi evidentirani kvarovi po godinama, dok su u tablici III. kvarovi razvrstani po uzrocima kvarova.

TABLICA II. BROJ KVAROVA NA VHF DSC SUSTAVU PLOVPUTA OD 2008. G. DO KRAJA 2012. G.

	Godina	*Broj kvarova VHF DSC sustava Plovputa*
1.	2008. g.	0
2.	2009. g.	1
3.	2010. g.	11
4.	2011. g.	5
5.	2012. g.	8
	UKUPNO	**25**

Analizom navedenih podataka o broju i uzroku kvarova VHF DSC sustava Plovputa dolazi se do zaključka da niti jedan kvar nije uzrokovan zbog arhitekture VHF DSC sustava Plovputa, kao niti zbog odabranog načina implementacije zalihosne opreme poslužitelja.

TABLICA III. PREGLED UZROKA KVAROVA NA VHF DSC SUSTAVU PLOVPUTA OD 2008. G. DO KRAJA 2012. G.

	Uzrok kvara	*Broj kvarova VHF DSC sustava Plovputa*
1.	Prekid priključne točke VPN mreže na ORP Split radio	13
2.	Greška napajanja na ORP Split radio	6
3.	Radovi na ORP Split radio	3
4.	Greška u konfiguraciji baze podataka	3
	UKUPNO	**25**

Uzrok kvarova greške napajanja na ORP Split radio je tijekom 2012. g. riješen nabavom novog uređaja za neprekidno napajanje.

Nadalje, većina kvarova VHF DSC sustava Plovputa se ne bi dogodila da su poslužitelji dislocirani na različitim lokacijama, npr. jedan poslužitelj na ORP Split radio, a drugi poslužitelj na ORP Rijeka radio. Na taj način bi bilo spriječeno 22 od ukupno 25 grešaka u prethodnom razdoblju.

Stoga je u planu u bliskoj budućnosti (tijekom prvog kvartala 2013. g.) poboljšati VHF DSC sustav Plovputa na način da se jedan od poslužitelja premjestiti na ORP Rijeka radio, jer na toj lokaciji već postoji sva potrebna infrastruktura (priključna točka VPN mreže od 1Mbit/s, neprekidno napajanje, agregatno napajanje, klimatizacija).

V. ZAKLJUČAK

Slijedom svega navedenog u ovom članku, zaključuje se da je odabrani način implementacije zalihosne opreme poslužitelja u VHF DSC sustavu Plovputa, odnosno paralelni rad oba poslužitelja, u potpunosti opravdan, te se isti način rada može pouzdano iskoristiti za sve „near real time“ sustave.

Paralelan rad oba poslužitelja osigurao je robusnost, raspoloživost i pouzdanost VHF DSC sustava Plovputa, te na taj način povećao sigurnost plovidbe Jadranom, što je primarni cilj VHF DSC sustava.

Analizom statističkih podataka uočene su mogućnosti poboljšanja VHF DSC sustav Plovputa, koji se može realizirati na način da se poslužitelji razmjeste na različite lokacije. Takvom dislociranošću poslužiteljske računalne opreme spriječilo bi se 88% grešaka VHF DSC sustava Plovputa koje su se dogodile od početka eksploatacije (2008. g.) do kraja 2012. g.. U bliskoj budućnosti će se obaviti realizacija razmještanja, na način da će se jedan poslužitelj premjestiti na lokaciju ORP Rijeka radio, a drugi će ostati na postojećoj lokaciji ORP Split radio.

LITERATURA

[1] International Maritime Organization, „Master Plan of Shore-based facilities for the GMDSS“, web izvor: http://www.imo.org/blast/mainframemenu.asp?topic_id=822 (13.02.2013)

[2] International Maritime Organization, „International Convention for the Safety of Life at Sea (SOLAS), 1974“, web izvor: http://www.imo.org/about/conventions/listofconventions/pages/international-convention-for-the-safety-of-life-at-sea-(solas),-1974.aspx (13.02.2013)

[3] International Telecommunication Union, “M.541: Operational procedures for the use of digital selective-calling equipment in

maritime mobile service", web izvor http://www.itu.int/rec/R-REC-M.541/en (14.02.2013)

[4] International Telecommunication Union, "M.493: Digital selective-calling system for use in maritime mobile service", web izvor http://www.itu.int/rec/R-REC-M.493/en (14.02.2013)

[5] International Telecommunication Union, "M.821: Optional expansion of the digital selective-calling system for use in the maritime mobile service", web izvor http://www.itu.int/rec/R-REC-M.821/en (14.02.2013)

[6] Pro SQL Server 2005 Replication, Sujoy P. Paul, Apress 2006, Berkley CA, USA

Primjena Microsoft Windows Embedded operativnog sustava na ugrađenim računalnim platformama

S. Braica*, N.Grgić*
*Sveučilišni odjel za stručne studije, Split, Hrvatska
stipe.braica@oss.unist.hr
nikola.grgic@oss.unist.hr

Sažetak–Primjena računala za POS (*Point of Sale*), web kiosk, informacijske panele, navigacijsku primjenu u plovilima i sl. karakterizirana je dominantnim korištenjem samo jedne aplikacije. Ovaj način korištenja znatno smanjuje zahtjeve na složenost programske i sklopovske opreme računala, ali ovisno o primjeni računala i programa može donijeti i neke nove specifične zahtjeve. Ovaj rad opisuje primjenu *Microsoft Windows Embedded* operativnog sustava, modularnog operativnog sustava, prilagođenog radu na računalima male procesorske snage sa specifičnim sklopovskim komponentama. Modularnost sustava omogućuje da operativni sustav sadrži samo komponente nužne za konfiguraciju i pokretanje glavne aplikacije ugrađenog računala. Ugrađena računala tradicionalno se baziraju na specifičnim sustavima više cijene temeljenim na *Linux* operativnim sustavima. Korištenje *Windows Embedded* operativnog sustava relativno je nova tehnologija koja omogućuje korištenje postojećih Windows aplikacija na ugrađenim računalima.

U radu se opisuje sustav bežičnog pristupa Internetu i prikaz plovnih informacija broda/motora na čarter plovilima.

I. UVOD

Zahvaljujući dobroj pokrivenosti GSM signalom obale i otoka Hrvatske računala s ekranima velikih dijagonala postaju standardni dio ponude čarter plovila. Osim pristupa Internet mreži, praćenja parametara rada motora i broda, meteo informacija moguće je (uz korištenje vanjskog modula za GPS pozicioniranje) korištenje i naprednih aplikacija za navigaciju.

Uvođenje računala u čarter plovila ima određene specifičnosti, koje se poklapaju s tržišnom ulogom *Windows Embedded* operativnog sustava. Korisnik s računalom komunicira ekranom osjetljivim na dodir. Nakon uključenja računala i pokretanja operativnog sustava, pokreće se glavna aplikacija računala, koja sadrži svu funkcionalnost koja je dostupna korisniku. Sklopovski i programski zahtjevi na računalo definirani su glavnom aplikacijom i unaprijed su poznati.

U opisanom računalu, glavna aplikacija omogućuje pristup Internetu i prikaz različitih parametara plovidbe (brzina i smjer broda i vjetra, parametri motora, meteo informacije). U odnosu na postojeće uređaje koji pružaju sličnu funkcionalnost, korištenje elemenata klasičnih računala nudi očite prednosti: nižu cijenu, jednostavnije korištenje, neusporedivo veći ekran i mogućnost nadogradnje programskog ili sklopovskog dijela računala.

Računala na plovilima izložena su različitim nepogodnim utjecajima okoline: velikoj vlazi, velikoj temperaturi i vibracijama. Utjecaj vlage i temperature može se smanjiti korištenjem kvalitetnih komponenti s malom disipacijom topline. Matična ploča malih je dimenzija, a kućište (slika 1) računala je izrađeno od aluminija s rebrima za dodatno hlađenje i mogućnošću pričvršćivanja vijcima na šasiju plovila.

Slika 1 – Mini ITX kućište računala

Vibracija je problem koji je najznačajniji kod korištenja računala na plovilima, i zahtijeva posebnu pažnju.

Mehanički dijelovi računala na koje vibracije imaju štetno djelovanje su ventilatori za hlađenje i magnetski čvrsti disk. Kvalitetni ventilatori relativno su otporni na vibracije, a ukoliko su priključeni na matičnu ploču računala uz korištenje senzora broja okretaja moguće je programsko detektiranje otkazivanja ventilatora i zaštitno isključivanje računala. Treba napomenuti da procesori manje snage mogu raditi i bez ventilatora, posebno ukoliko je matična ploča postavljena u kućište s dovoljnim efektom hlađenja, ali radi sigurnosti korištenje ventilatora je preporučen opcija.

Magnetski čvrsti disk komponenta je računala koja se ne smije koristiti na plovilima jer bi se brzo pokvarila. Čak i najmanje vibracije uzrokuju nepovratno oštećenje diska, pa iako *Windows embedded* operativni sustav koristi čvrsti disk neusporedivo manje u usporedbi s klasičnim Windows sustavom, njegovo korištenje nije opcija.

SSD (engl. *solid state drive*) diskovi bazirani su na FLASH memorijskim ćelijama i imaju niz prednosti pred klasičnim magnetskim diskovima, u prvom redu veću brzinu i nižu potrošnju energije ali i znatno veću cijenu i niži kapacitet, te posebno važno svojstvo otpornosti na vibracije zbog nepostojanja mehaničkih dijelova.

Posebnost računala korištenog u opisanom radu je korištenje CompactFlash diskova malog kapaciteta (256 MB) u vremenu prije masovnog uvođenja SSD tehnologije u stolnim računalima.

Naime, SSD diskovi dimenzijama su prilagođeni ugradnji u standardna kućišta računala, što ugradnju u Mini ITX kućište prikazano na slici 1, čini nemogućom. Osim toga, performanse pristupa disku nisu takve da opravdavaju njegovo korištenje: disk se koristi za učitavanje operativnog sustava i aplikacije, te zapisivanje male količine podataka tijekom gašenja računala, pa se kod računala korištenog u opisanom radu koristi CompactFlash disk male veličine (256 MB).

CompactFlash je dugo vremena bio dominantni standard za spremanje podataka na elektroničkim uređajima, npr. foto i video uređajima. Posebnost koja ga čini prilagodljivim za spajanje na računala je korištenje ATA sučelja koje je signalizacijski/elektronički identično ATA sučelju na klasičnim magnetskim diskovima, pa je spajanje moguće uz korištenje jednostavnog pasivnog adaptera. Dimenzije diska i adaptera ne predstavljaju problem, kako se vidi iz slike 3: adapter i disk svojim dimenzijama omogućuju jednostavnu ugradnju u mala kućišta, poput kućišta prikazanog na slici 1, okvirnih dimenzija 18 X 22 X 7 cm, što odgovara dimenzijama potrebnim za montažu matične ploče i sklopa napajanja, a visina omogućuje i nesmetanu ugradnju pasivnog adaptera i CF diska.

Windows Embedded je prvi Microsoftov operativni sustav koji podržava FLASH tehnologije za trajnu memoriju kroz prilagodbe za rad na FAT16 particiji (nedostupno na 32-bitnim Windows instalacijama) i RAMDISK rad (omogućuje minimalno zapisivanje na FLASH disk, čime se produljuje njegov životni vijek). Modularnost operativnog sustava omogućuje iznimno malu veličinu potrebnog prostora na FLASH disku (oko 220 MB je dovoljno za pokretanje Windows aplikacije napisane za Microsoft .Net framework).

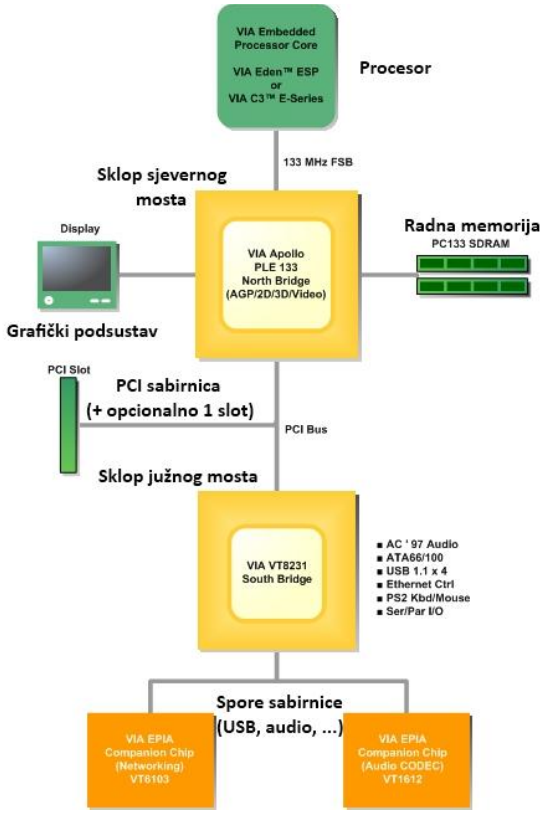
Brzina čitanja CompactFlash diskova je približno 50MB/s u čitanju i pisanju, što je s obzirom na veličinu operativnog sustava dovoljna brzina. Ostale karakteristike uređaja (rad na temperaturi do 85°C i manja potrošnja u odnosu na SSD) omogućuju njegovo korištenje bez posebnih ograničenja.

Algoritmi za inteligentno zapisivanje na FLASH uređaje nisu potrebni jer *Windows embedded* FLASH uređaje koristi na način da se zapisivanje na uređaj obavlja samo kod isključivanja uređaja, te je trošenje FLASH memorije minimalno.

II. SKLOPOVSKA PODRŠKA

A. *Mini ITX platforma*

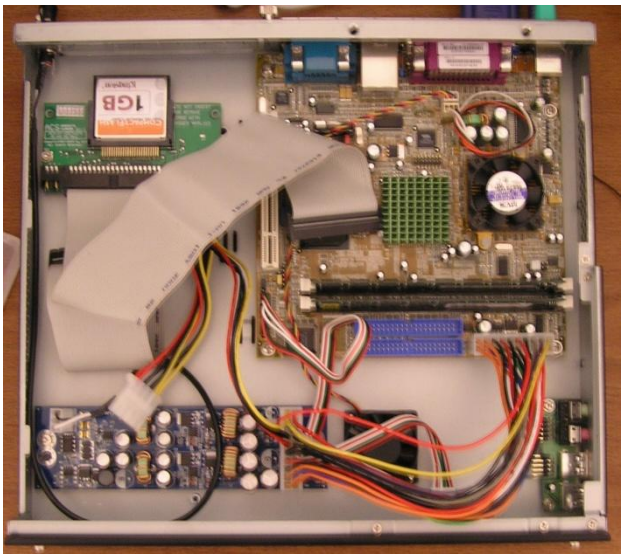
Dominantna dimenzija matičnih ploča za ugrađena računala na *Windows Embedded* platformi je Mini ITX (170 x 170 mm), prvenstveno zahvaljujući tajvanskoj tvrtki VIA, koja je serijom Intel-kompatibilnih matičnih ploča s vlastitim procesorima Via C3, C7 i Eden serija popularizirala ovaj format ugrađenog računala (slika 2).



Slika 2 – Arhitektura VIA C3 računalnog sustava

Osnovne karakteristike matične ploče slijede filozofiju male potrošnje koja je glavna karakteristike procesora: bez ili samo jedan utor za proširenje, procesor i krug mosta bez hladnjaka, grafički podsustav male snage. Kako se radi o računalima male snage, korištene su komponente koje su duže vrijeme dostupne na tržištu (sporija memorija, jednostavni audio sustav) te određene pojednostavljenja matične arhitekture (dijeljena glavna i grafička memorija). Za primjenu koja je namijenjena ovim računalima, navedene karakteristike ne predstavljaju ograničenje.

Napajanje matične ploče redovito se obavlja s istosmjernog izvora 6 ili 12V, s značajnim oscilacijama u vrijednosti napona tijekom rada, koje su često prisutne električnim instalacijama vozila i plovila. Za matične ploče male snage, postoje kvalitetna rješenja generatora potrebnih napona matične ploče po ATX standardu.



Slika 3 – Mini ITX računalo s ispravljačem i CompactFlash diskom

Slika 3 prikazuje jednu od varijanti računala korištenu tijekom razvoja koji se sastoji od:

1. Matične ploče (desno, gore)
2. CompactFlash-IDE adaptera i 1GB flash kartica (gore, lijevo)
3. Adaptera za napajanje 12V ATX (dolje, lijevo)

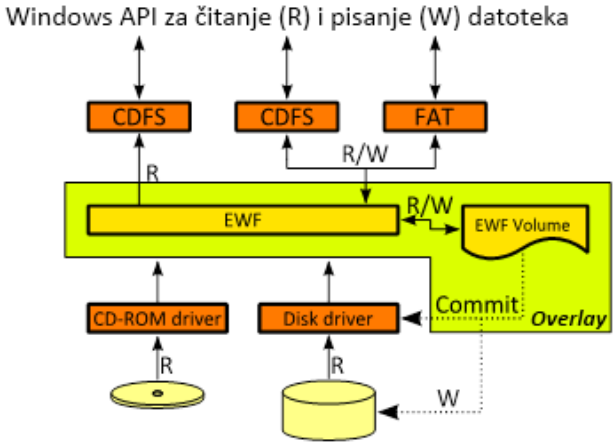
CompactFlash kartice početkom 21. stoljeća masovno se koriste kao medij za pohranu fotografija na digitalnim foto aparatima. Digitalno sučelje ovih kartica u Compact Flash izvedbi sačinjeno je kao 40-pin IDE sučelje +4 pina za napajanje +5V (ponekad i 3.3V), pa se isključivo mehaničkom adaptacijom može povezati na klasično IDE sučelje računala, kako je i prikazano na slici.

B. *Primjena FLASH memorije*

Problem trajnosti FLASH memorija još uvijek je predmet istraživanja a u prikazanom slučaju riješen je korištenjem *overlay* funkcionalnosti *Windows embedded* sustava.

Naime, *Windows Embedded* omogućuje osnovnu prilagodbu radu s FLASH memorijom na način da operativni sustav stvara sliku diska u radnoj memoriji računala. Kako je operativni sustav iznimno male veličine (izuzet od svih, osim nužnih modula, potreba za pisanjem u *swap* datoteku operativnog sustava je malo vjerojatna, posebno ako je radna memorija dovoljne veličine. Korištenjem 2GB radne memorije i 1GB FLASH memorije (iskorištenje približno 256 MB) postignuto je da se cijeli disk može premapirati u radnu memoriju, a da aplikacija i operativni sustav i dalje imaju dovoljno memorije za svoje potrebe.

Situacija pri kojoj aplikacije ne trebaju nikakav trajan zapis na disku je jako rijetka, stoga je potrebno osigurati mogućnost zapisivanje podataka na FLASH disk, ali eliminirati svaku upotrebu FLASH diska u *swap* modu. Rješenje unutar Windows Embedded sustava je korištenje *EWF (Enhanced Write Filter)* funkcionalnosti, a u osnovi se sastoji od posebnog memorijskog područja (*overlay*) u koji se preusmjeravaju sve operacije zapisivanje namijenjene FLASH disku. Ovaj model ne zahtijeva značajne promjene glavne aplikacije. Slika 4 prikazuje EWF arhitekturu *Windows Embedded sustava.* Operacije čitanja i pisanja operativnog sustava, ne prosljeđuju se izravno pokretačkim programima uređaja (*device driver*), već novouvedenom *Overlay* području. Ova arhitektura omogućuje korištenje postojećih aplikacija bez izmjena, odnosno aplikacije nisu svjesne postojanja *Overlay* područja. EWF arhitektura podržava sve najčešće korištene datotečne sustave (FAT16, FAT32, CDFS, itd.)



Slika 4 – EWF arhitektura *Windows Embedded* sustava

Overlay područje nalazi se u radnoj memoriji i spremanje na stvarni CompactFlash disk obavlja se komandnolinijskom aplikacijom *EWF Manager* na kraju rada računala, ili korištenjem API funkcionalnosti dostupne kroz operativni sustav. Regularno gašenje glavne aplikacije odgovara gašenju računala, stoga glavnu aplikaciju računala treba modificirati tako da uključuje i *Commit* operaciju, čime se podaci konačno iz EWF područja spremaju na FLASH disk. Broj zapisivanje na CompactFlash disk je ograničen, pa ga *Windows Embedded* operativni sustav ograničava na dva načina:

1. Zapisivanje se obavlja samo jednom tijekom jednog ciklusa rada računala, tijekom gašenja.
2. Algoritam zapisivanja zapisuje samo nove ili promijenjene podatke.

Opterećenje jedne ćelije flash memorije stoga je jednako broju gašenja računala u promatranom razdoblju. U odnosu na regularan rad računala koje ima aktivan *swap* zapis ovo je zanemariva brojka.

Problem koji se može pojaviti je nenadano gašenje računala uslijed greške programa ili operativnog sustava. Dobro napisana glavna aplikacija može prepoznati vlastite greške i reagirati na njih uključujući i *Commit* operaciju prije prestanka rada. Greška operativnog sustava vodi prisilnom gašenju glavne aplikacije i u tom slučaju *Commit* operacija se neće izvršiti.

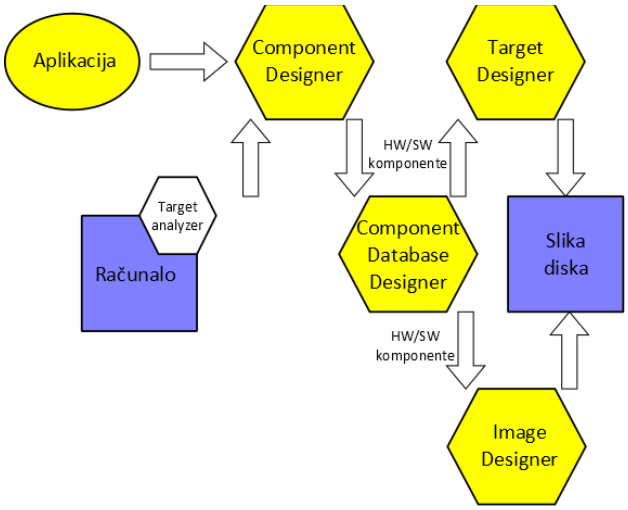
III. PROGRAMSKA PODRŠKA

A. *Uvod*

Operativni sustav pruža podršku korištenja raznovrsnog sklopovlja raznovrsnim aplikacijama. U cilju

postizanja minimalne veličine operativnog sustava (***footprint)*** operativnog sustava, potrebno je eliminirati sve suvišne pomoćne programe za hardver koji ne postoji, kao i sve module operativnog sustava koji nisu nužni za pokretanje aplikacije uređaja.

Windows Embedded omogućuje ostvarenje ovog cilja korištenjem alata *Windows Embedded Designer* i važnog pomoćnog alata *Target Analyzer*. Proces stvaranja finalne slike diska prikazan je na slici 5.



Slika 5 – Proces stvaranja slike FLASH diska

Kako bi uklonio nepotrebne upravljače uređaja iz konačne slike, *Designer* treba što bolji uvid u elemente računala. *Target Analyzer* je aplikacija koja postoji u dvije varijante: komandnolinijska se može pokrenuti s DOS uređaja (disketa, bootable USB flash disk), ali zbog realnih ograničenja DOS sustava daje realno malo informacija, dok Windows varijanta aplikacije zahtijeva instalaciju pravog Windows sustava na računalo i ispravno namještanje svih uređaja, ali daje potpun i ispravan popis svih uređaja u računalu kao i potrebne upravljačke programe za njihov rad. Prikupljene podatke *analyzer* sprema u *pmq* datoteku koja se koristi kao najvažniji ulazni parametar izrade slike.

```
<?xml version="1.0"?>
<!DOCTYPE HIB SYSTEM "devices.dtd">
<!--                                          -->
<!--                                          -->
<HIB>
 <TOOLINFO>
  <PLATFORM></PLATFORM>
  <TOOL>Target Analyzer Probe Version 1.1.222.0</TOOL>
  <TOOLOPTIONS>"C:\WES7\Tools\x64 Tap.exe"</TOOLOPTIONS>
  <TIMESTAMP>2013-02-09,11:31:56PM</TIMESTAMP>
 </TOOLINFO>
 <DEVICES>
  <CATEGORY Name="ACPI">
   <DEVICE ConfigFlags="0">
    <DEVICEDESC>MS AC Adapter;</DEVICEDESC>
     <HARDWAREIDS>
      <DEVICEID Order="1">ACPI\ACPI0003</DEVICEID>
      <DEVICEID Order="2">*ACPI0003</DEVICEID>
     </HARDWAREIDS>
     <COMPATIBLEIDS>
     </COMPATIBLEIDS>
    </DEVICE>
```

Prikazani odsječak prikazuje početak pmq datoteke i jedan zapis karakteristike jednog uređaja. Na isti način opisuju se svi uređaji unutar računala.

Podešavanje programskog dijela operativnog sustava ovisi o vrsti glavne aplikacije koja će se primarno koristiti na ugrađenom sustavu i potrebama konfiguracije računala. *Windows embedded* razvojno okruženje olakšava ovaj postupak korištenjem paketa određene funkcionalnosti (jezični paketi, paket za mrežnu funkcionalnost, multimedija, itd). Uključivanje određenog paketa rezultira dodavanjem svih potrebnih biblioteka tražene funkcionalnosti u finalnu sliku.

Točno utvrđivanje potrebnih modula operativnog sustava znatno je olakšano ukoliko se aplikacija izvodi na virtualnom stroju (npr. .NET) jer *Windows Embedded* sam zna razlučiti sve potrebne komponente za rad virtualnog stroja.

IV. ZAKLJUČAK

Područje korištenje ugrađenih računala prolazi kroz fazu značajnih promjena, iz faze usko specijaliziranih sustava u inteligentne, na Internet povezane uređaje. Ovaj trend poznat je pod nazivom „pervasive computing“ (sveprisutno računarstvo).

Windows Embeded inačica *Windows* operativnih sustava, za razliku od mnogih drugih ugradbenih operativnih sustava omogućuje jednostavnu prenosivost aplikacija pisanih na jezicima visoke razine pripremljene za osobna računala na ugrađene računalne sustave.

Ugrađena računala danas nalazimo u sasvim novim primjenama, od MP3 uređaja do foto i video uređaja. Većina komunikacijskih uređaja danas standardno koristi i Internet mrežu i IP kao protokol za prijenos podataka te ima relativno kratak vijek na tržištu uslijed potrebe za stalnim povećanjem performansi koje nije izvedivo bez poboljšanja sklopovlja računala. Primjena ugrađenih računala danas često zahtijeva podršku cijelog niza protokola (IP kao spomenuti najočitiji primjer), te je razvoj tako složenih računala samo za pojedinu specifičnu primjenu i relativno kratak životni vijek neisplativ.

Svi navedeni razlozi čine proizvođačima takvih uređaja mogućnost korištenja standardizirane platforme mnogo isplativijom od vlastitog razvoja sklopovlja i programske potpore, čiji je vijek na tržištu relativno kratak. Pouzdana i standardizirana platforma omogućuje brz izlazak na tržište novog modela proizvoda, kompatibilnost s različitim podržanim standardima i protokolima i iskoristivost prethodno razvijenog softvera.

Windows Embeded operativni sustav omogućuje korištenje većine funkcionalnosti klasičnog *Windows* sustava uz prilagodbu veličine slike diska i sklopovlja radu na ugrađenim platformama, npr. najjednostavnije

varijante FLASH memorije, ili podizanje operativnog sustava s CD diska umjesto hard diska.

Korištenje mrežnih funkcionalnosti, grafičkog i sigurnosnog podsustava potpuno je identično kao u postojećim aplikacijama.

Navedene prednosti čine *Windows Embeded* zanimljivim predstavnikom trenda sveprisutnog računarstva, posebno u novim područjima primjene ugrađenih računala, gdje je moguć prijenos već razvijenih aplikacija s osobnih računala. Razvoj aplikacija za *Windows Embedded* sustav karakteriziran je činjenicom da je ugrađeno računalo (*target*) jako slično, ponekad i identično računalu na kojem se obavlja razvoj aplikacije, što nije slučaj kod mnogih novih platformi za ugrađena računala (npr. Android), pa je razvoj testiranje i otklanjanje pogreški jako olakšano.

Programska kompatibilnost i jednostavnost razvoja složene programske podrške glavni je razlog odabira Windows Embedded operativnog sustava u prikazanom uređaju.

Široka dostupnost matičnih ploča, kućišta i ostale sklopovske opreme prilagođene za korištenje u plovilima, automobilima i sl. omogućuje relativno jednostavan i jeftin razvoj ugrađenog računala za primjenu u takvim uvjetima.

LITERATURA

[1] Via Technologies, "EPIA SPseries mini-ITX mainboard operating guide", web izvor:http://www.viaembedded.com/servlet/downloadSvl?id=261&download_file_id=12674 (15.1.2013)

[2] Microsoft Windows Embedded OS, web izvor: http://www.microsoft.com/windowsembedded/en-us/windows-embedded.aspx

[3] Microsoft Windows Embedded IDE, pomoćni tekst I dokumentacija

Upotreba mini računala s Linux operativnim sustavom na lokacijama s zahtjevnim uvjetima okoline

N. Grgić*, J. Vrlić** i I. Ružić*
* Sveučilišni odjel za stručne studije, Split, Hrvatska
** Plovput d.o.o, Split, Hrvatska
nikola.grgic@oss.unist.hr
josip.vrlic@plovput.hr
ivica.ruzic@oss.unist.hr

Sažetak - Često postoji potreba korištenja računala na lokacijama koje u potpunosti ne zadovoljavaju standardne uvjete okoline (temperatura, vlaga, nečistoće u zraku, udaljenost od ostalih uređaja, ...), a potrebna snaga računala nije kritična. Korištenjem mini računala (mini-ITX) bez mehaničkih pokretnih dijelova s instaliranim Linux operativnim sustavom koji koristi isključivo RAM memoriju za svoj rad, može se dobiti potpuno funkcionalno računalo, koje i u zahtjevnim uvjetima okoline nije potrebno posebno održavati. U ovom radu opisan je strojni i programski dio mini računala, te statistika rada 13 računala u četverogodišnjem razdoblju s 24/7/365 radnim režimom.

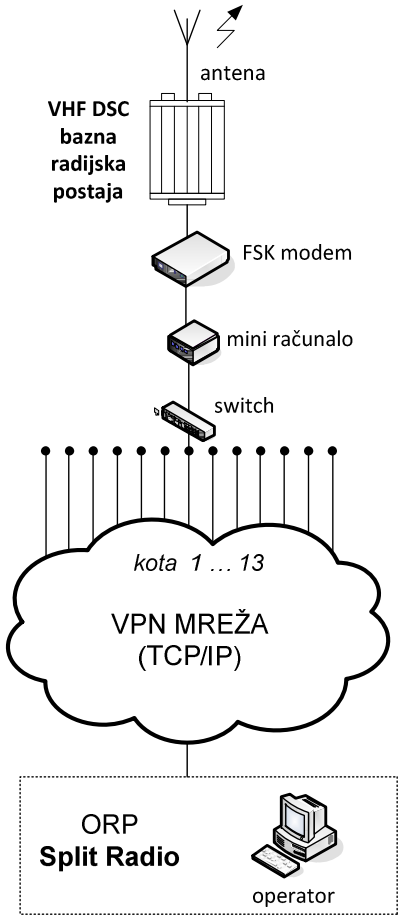
I. UVOD

Plovput d.o.o. je tvrtka čija je djelatnost vezana za sigurnost plovidbe, te između ostalog obuhvaća postavljanje objekata sigurnosti plovidbe na hrvatskom teritorijalnom moru i osiguranje njihovoga pravilnog rada, obavljanje poslova radijske službe na pomorskim plovnim putovima i obavljanje istraživanja i projektiranja radi obavljanja navedenih poslova.

U sklopu povećanja sigurnosti plovnog prometa na Jadranu i usklađivanja s međunarodnim standardima o sigurnosti plovidbe, Plovput je izgradio i održava VHF DSC sustav kao dio svjetskog pomorskog sustava za pogibelj i sigurnost (engl. *Global Maritime Distress Safety System,* GMDSS). Sustav se koristi za automatsko pozivanje u svrhu dojave o pogibelji, hitnosti i sigurnosti, te za pozivanje zbog ostvarivanja ostalih vrsta komunikacije. VHF DSC sustav Plovputa uključuje opremu smještenu na obalnoj radijskoj postaji Split Radio, te na 13 kota duž Jadranske obale.

Pojedine kote smještene su na lokacijama koje su zahtjevne za elektroničku opremu i uvelike ovisne o trenutnim klimatskim uvjetima, pri čemu je svaka lokacija specifična, a najveći izazov predstavlja izloženost visokim temperaturama, prašini, posolici i vlazi.

U sklopu izgradnje VHF DSC sustava pokrenut je projekt povezivanja obalne radijske postaje iz Splita s pojedinačnim VHF DSC baznim radijskim postajama putem virtualne privatne mreže (VPN) bazirane na TCP/IP tehnologiji – slika 1. U sklopu projekta bilo je potrebno na svaku kotu postaviti računalo koje će usmjeravati komunikacijsko-upravljačke poruke prema VHF DSC baznoj radijskoj postaji. Gotova rješenja koja su postojala na tržištu i koja bi zadovoljila opisane uvjete bila su skupa, pa se krenulo u izgradnju vlastitog sustava od komponenti dostupnih na tržištu. Tom prilikom je sastavljeno 13 mini računala koja su zajedno sa baznim postajama postavljena na lokacije sa kojih se ostvaruje optimalna pokrivenost hrvatskog teritorijalnog mora i tamo do danas obavljaju opisane zadatke.



Slika 1 – VHF DSC sustav Plovputa

Svaka od 13 kota sastoji se od mrežne opreme, mini računala, FSK modema i VHF DSC bazne radijske postaje s antenom.

Na mini računalo postavljena je višenitna aplikacija koja prima i usmjerava poruke između obalnih radijskih postaja sa pripadnim poslužiteljima i bazne radijske postaje na koti. Veza između računala i radijske postaje ostvarena je FSK modemom koji modulira DSC poruke prije slanja u eter, odnosno demodulira DSC poruke koje pristignu na postaju. Modulacija i demodulacija se obavljaju sukladno ITU-R preporuci M.493 [1].

II. TEHNIČKE KARAKTERISTIKE RAČUNALA

A. *Osnovna konfiguracija*

Korištena mini računala sastavljena su oko matične ploča VIA EPIA serije SP u formatu Mini ITX. U vrijeme sastavljanja računala (rujan 2008.), ploča se proizvodila u dvije različite varijante: SP8000E s integriranim VIA Eden procesorom čije su osnovne karakteristike izrazito niska potrošnja i zagrijavanje, te SP13000 s procesorom VIA C3 koji se više zagrijava, ali je pogodniji za uporabu u procesorski zahtjevnijim zadaćama. S obzirom da je za hlađenje potonjeg potreban aktivni hladnjak, a jedan od osnovnih zahtjeva kod izgradnje sustava je bio da oprema nema mehaničkih pokretnih dijelova, te da zahtjevi za procesorskom snagom računala nisu bili toliki da opravdaju upotrebu jačeg procesora, izabrana je inačica ploče SP8000E.

Na ploči se nalazi integrirana AGP grafička kartica, te jedan DDR400 utor u koji je ugrađeno maksimalno podržanih 1 GB radne memorije. Od korištenih ulazno-izlaznih konektora na ploči, valja istaknuti:

- 2 serijska porta
- ethernet port
- VGA port
- PS2 port za tipkovnicu i miša.

Ostale su specifikacije dostupne na [2].

Osim nabrojenih konektora, odabrana matična ploča ima podršku za široki skup općeprihvaćenih tehnologija i standardnih ulaza i izlaza, te se po tome ne razlikuje od standardnih ploča za osobna računala, što joj omogućava primjenu u gotovo svim uobičajenim zadatcima za koje nije potrebna velika procesorska snaga.

Matična ploča je dimenzija 17x17 cm, te je zajedno s ostalom opremom ugrađena u robusno aluminijsko kućište proizvođača Mini Box (model VoomPC) koje je dizajnirano u obliku pasivnog hladnjaka s rebrima koja povećavaju toplinsku disipaciju sustava. Ukupna težina sastavljenog računala je 3,25 kg, od čega 2,85 kg otpada na kućište.

B. *Napajanje*

U kućište je ugrađeno nestandardno ATX napajanje proizvođača CarTFT, model M1-ATX 6-24V DC/DC sa istosmjernim ulaznim naponom. Na lokacijama na kojima je dostupna javna elektroenergetska mreža, ugrađen je 230/12 V ispravljač MeanWell S-150 i neprekidno napajanje (UPS) proizvođača GE Digital Energy, model NetPro 1000.

Izmjerena prosječna potrošnja računala u mirovanju je 18,3 W/h, od čega na ploču i integrirani procesor otpada 12,5 W/h [2]. Izabrano neprekidno napajanje može za tu razinu potrošnje osigurati preko pet sati autonomije u radu [3].

C. *Diskovni podsustav*

Klasični tvrdi diskovi izgrađeni su od magnetskih ploča koje u radu rotiraju velikom brzinom, dok iznad njih lebdi glava za čitanje/pisanje. Zbog osjetljivosti opisanog mehanizma upotreba takvih diskova nije moguća na mjestima gdje je računalo izloženo vibracijama (npr. automobili, brodovi, industrijska postrojenja). Osim toga, za pokretanje navedenih dijelova potrebna je energija koja uzrokuje dodatno zagrijavanje sustava, što je redovito nepoželjno.

Kako bi se izbjegle navedene poteškoće, u opisano računalo je kao medij za pohranu ugrađena *CompactFlash* kartica, pri čemu je korišten adapter s CF na standardni 40-*pinski* IDE, s podrškom za podizanje operativnog sustava (*boot*). Valja napomenuti da su se u međuvremenu na tržištu pojavile mini ITX matične ploče s integriranim CF konektorom, što će dodatno olakšati izgradnju ovakvih sustava.

Unatoč prednosti koje ima korištenje memorijskih *flash* kratica u ovakvim sustavima, kod dugotrajne eksploatacije se pojavljuje problem ograničenog životnog vijeka *flash* memorije koji je posljedica tehnike zapisivanja podataka i fizičkih ograničenja medija. U računalo je ugrađena CF kartica veličine 2 GB tipa *Extreme III* proizvođača *Sandisk* proizvedena u NAND tehnologiji [4]. Trajnost jednom pohranjenih podataka u takvu vrstu memorije procjenjuje se na više od 10 godina. Pri tome proces čitanja značajno ne degradira fizička svojstva kartice, ali je broj pisanja u pojedinačni blok podataka ograničen, te je preporučeno da se kod dizajniranja sustava baziranih na NAND memoriji ne planira više od 10^6 upisa po istom bloku [5].

Navedeno ograničenje posebno dolazi do izražaja kod rada operativnih sustava koji intenzivno koriste straničnu datoteku (*swap*) te zbog toga nekontrolirano zapisuju velike količine podataka na iste fizičke blokove kartice.

Opisana ograničenja moguće je zaobići učitavanjem cjelokupnog operativnog sustava u virtualni disk smješten u radnoj memoriji računala (*RAM-boot*).

III. PRIPREMA OPERATIVNOG SUSTAVA ZA POKRETANJE IZ RADNE MEMORIJE (RAM-BOOT)

A. *Općenito*

Ukoliko se sve datoteke operativnog sustava učitaju u virtualni RAM-disk, te se s njega izvrši podizanje sustava, svako kasnije čitanje i zapisivanje podataka vršit će se unutar radne memorije, pa će se tako znatno produžiti

radni vijek memorijske *flash* kartice. Osim toga, na taj način se postižu daleko bolje performanse računala. Pri tome treba voditi računa o tome da radna memorija nije medij za trajnu pohranu podataka, pa će se u slučaju neočekivanog gašenja računala (nestanak struje, softverski reset i slično) izgubiti svi podatci koji su bili pohranjeni na virtualnom disku.

Gubitak podataka iz radne memorije računala kod VHF DSC sustava Plovputa ne predstavlja problem jer se opisana računala koriste isključivo kao komunikacijska i upravljačka oprema za VHF bazne radijske postaje, bez potrebe za trajnom pohranom podataka.

Iako je računalo spojeno na jedinicu neprekidnog napajanja (UPS), postoji mogućnost neočekivanog isključenja napajanja ili prekida rada uslijed više sile, pa je u postavkama BIOS-a namješteno automatsko paljenje računala prilikom ponovnog priključenja na napajanje, nakon čega se ponovno učitava OS i sustav nastavlja s radom.

B. *Postavljanje operativnog sustava*

Kao operativni sustav na računalima je postavljena Debian v4.0 distribucija Linuxa, s kernelom 2.6.18. OS je instaliran na memorijsku karticu uobičajenim postupkom, izuzevši instalaciju softverskih paketa za koje je procijenjeno da su suvišni. Potom je pripremljena RAM-disk inačica OS-a prema sljedećim koracima [6]:

1. Napravljena je `fstab` (engl. *file system table*) datoteka `/etc/fstab.ramboot` s listom uređaja za pohranu, točaka za montiranje (engl. *mount point*) i pripadnih datotečnih sustava. Za učitavanje sustava u RAM posebno je važna linija u kojoj se datotečni sustav točke za montiranje / definira kao *tmpfs* (engl. *temporary file storage facility* – privremeni datotečni sustav u radnoj memoriji) [7].
2. Pokrenuta je skripta kojom su kopirane sve datoteke operativnog sustava u privremeni direktorij, gdje su izbrisani nepotrebni dijelovi sustava (nekorišteni pogonski programi i dokumenti), te je sve komprimirano u tar arhivu veličine 208 MB.
3. U direktoriju `/boot/` je pomoću alata *mkinitramfs* pripremljena *initrd* (engl. *initial ramdisk*) datoteka `initrd.img.ramboot` koju će kernel za vrijeme podizanja operativnog sustava raspakirati u RAM i koristiti kao početni *root* datotečni sustav. Navedeni postupak odvija se u ranoj fazi podizanja sustava (*early user space*), koja je dostupna u kernelu od verzije 2.5, a omogućava učitavanje RAM-diska iz *boot loadera*. Na taj način je omogućeno pravovremeno učitavanje pogonskih programa za specifične uređaje, a kako se tu ujedno izvršava i `/sbin/init` inicijalizacijski proces, posljedično je omogućeno i pokretanje proizvoljnih programa i skripti [8]. U toj fazi raspakirat će se pripremljena tar arhiva sa slikom operativnog sustava u *tempfs*, gdje će *init* montirati „pravi“ *root* datotečni sustav. S njega se pokreće i novi `/sbin/init` koji nastavlja s uobičajenim postupkom podizanja sustava, a *initrd* iz prve faze se uklanja.
4. Promijenjena je konfiguracijska datoteka GRUB *loadera* `menu.lst` kako bi se prilikom paljenja računala omogućio izbor između „glavnog“ operativnog sustava koji radi na memorijskoj kartici i njegove inačice učitane u RAM-disk.

Valja primijetiti da bi korištenje stranične datoteke na ovako postavljenom sustavu uzrokovalo nepotrebno prebacivanje blokova memorije iz jednog dijela RAM-a u drugi, pa je stranična datoteka isključena.

C. *Omogućavanje daljinskih izmjena na sustavu*

Uzevši u obzir opisani način rada, jasno je da će bilo kakve promjene u konfiguraciji operativnog sustava ili pripadnih programa biti poništene prilikom sljedećeg gašenja računala. Za trajno spremanje izmjena u konfiguraciji potrebno je pokrenuti računalo s memorijske kartice izborom odgovarajuće stavke u *boot loaderu*, te nakon izvršenih promjena pripremiti novu inačicu tar arhive sa izmijenjenom slikom sustava koja će se učitati u memoriju prilikom sljedećeg paljenja računala.

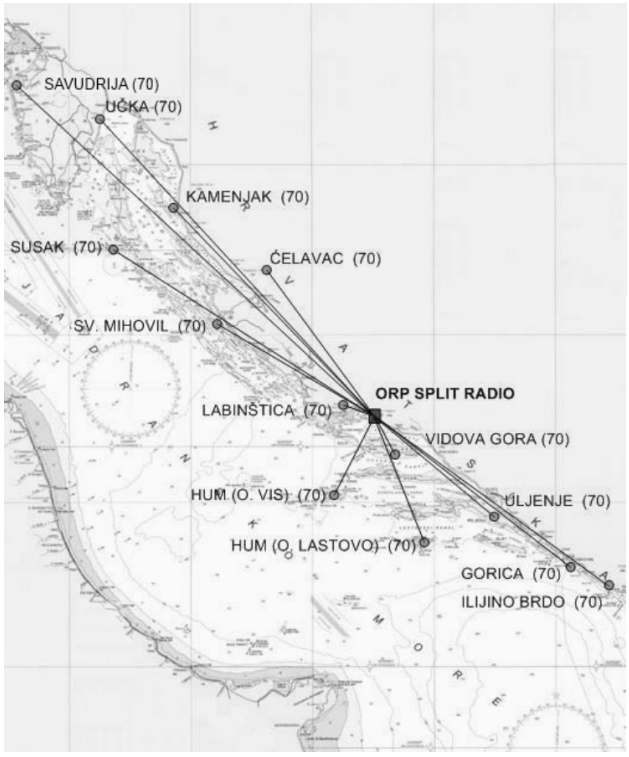
Taj jednostavni zadatak otežan je činjenicom da su računala fizički smještena na udaljenim lokacijama, a GRUB *loader* ne omogućava jednostavno daljinsko upravljanje. Stoga su na memorijskoj kartici unutar direktorija `/boot/grub/` napravljene dvije različite GRUB konfiguracijske datoteke, `menu.CF.lst` i `menu.RAM.lst`, u kojima su kao početni sustav za podizanje postavljeni CF kartica, odnosno RAM.

Na taj način je na računalu koje je podignuto s RAM-diska dovoljno pokrenuti pripremljenu skriptu koja će montirati CF karticu i presnimiti datoteku `menu.CF.lst` preko izvorne GRUB konfiguracijske datoteke (`menu.lst`) kako bi se prilikom sljedećeg pokretanja računalo podiglo s memorijske kartice umjesto iz RAM-a.

IV. STATISTIKA RADA RAČUNALA

A. *Radni uvjeti*

Sustav je pušten u pogon krajem 2008. godine, pri čemu je oprema razmještena na sljedeće lokacije duž Jadranske obale: Savudrija, Učka, Kamenjak, Susak, Ćelavac, Sv. Mihovil, Labinštica, Hum na Visu, Hum na Lastovu, Vidova Gora, Uljenje, Gorica i Ilijino Brdo – slika 2.



Slika 2 – Razmještaj kota VHF DSC sustava na Jadranu

Neke od navedenih lokacija valja izdvojiti zbog izrazito nepovoljnih uvjeta za rad elektroničke opreme:

- Kota Savudrija – oprema je smještena na vrhu svjetionika, unutar kamene kule, gdje ljetne temperature u prostoriji prelaze 50°C.
- Kota Susak (o. Susak) – računalo je smješteno u svjetioniku izloženom udarima bure, koja u određenim uvjetima podiže morsku vodu i zasipa posolicom svjetionik. Računalo nije direktno izloženo kapljicama mora, ali indirektno dolazi u kontakt sa soli.
- Kota Vidova Gora (o. Brač) – računalo se nalazi u prostoriji koja uslijed jakih udara južnog vjetra i kiše stvara veliku vlažnost zraka.
- Kota Kamenjak (o. Rab) – računalo se nalazi u potkrovlju koje nije u potpunosti izolirano, te se po kišnom vremenu događa da se voda slijeva na mrežni ormar u kojem se nalazi računalo.

Sve lokacije su pobrojane u Tablici I.

TABLICA I. PREGLED KOTA SA PRIKAZOM UVJETA RADA

Kota	***Opis***	***Klima-tizacija***	***Venti-lacija***	***Prašina (0-5)***
PS[a] Savudrija	Vrh kamene kule svjetionika	ne	nije potrebno	4
PS Susak (o. Susak)	Prostorija u svjetioniku	ne	ne	3
RRP[b] Učka	Prostorija s radijskom opremom	da	da	1
RRP Kamenjak (o. Rab)	Neizolirano potkrovlje	ne	nije potrebno	5
RRP Ćelavac	Prostorija s telekomunikacijskom opremom	da	ne	2
RRP Sv. Mihovil (o. Ugljan)	Prostorija s telekomunikacijskom opremom	da	ne	2
RRP Labinštica	Prostorija s radijskom opremom	ne	ne	3
RRP Vidova Gora (o. Brač)	Prostorija s radijskom opremom	da	da	4
VP[c] Hum (o. Vis)	Prostorija u neizoliranoj osmatračnici	ne	da	3
RRP Uljenje (po. Pelješac)	Prostorija s telekomunikacijskom opremom	da	da	3
RRP Hum (o. Lastovo)	Kontejner s telekomunikacijskom opremom	da	da	2
Gorica Sv. Vlaha	Prostorija s radijskom opremom	da	da	1
RRP Ilijino Brdo	Kontejner s radijskom opremom	da	ne	1

a. pomorski svjetionik

b. radio relejna postaja

c. vojna postaja

B. Kvarovi

Kroz četiri godine rada (od prosinca 2008. do siječnja 2013.) na sustavu su zabilježena dva kvara:

- kolovoz 2011. – na koti Vidova Gora (o. Brač) došlo je do prestanka rada računala uslijed pregrijavanja. Pregrijavanje je nastupilo kao posljedica kvara na klimatizacijskom uređaju koji nije pod kontrolom Plovputa. Zbog gusto postavljene elektroničke opreme na malom prostoru radijske postaje, temperatura u prostoriji je znatno porasla, što je uzrokovalo gašenje mini-računala i ispad kote. Pritom nije uzrokovana trajna šteta na opremi, pa je kota nastavila s radom nakon što je popravljen klimatizacijski uređaj.
- siječanj 2012. – na koti Vidova Gora (o. Brač) uslijed kiše i jakog juga, voda se preko antenskih kabela uvukla u prostoriju s opremom i izlila se po računalu, što je prouzročilo izgaranje modula za napajanje.

C. Srednje vrijeme između kvarova

U navedenom razdoblju računala su ostvarila preko 450.000 radnih sati, pri čemu su zabilježena samo dva kvara. Pritom je ostvaren MTBF (engl. *Mean time between failures* – srednje vrijeme između kvarova) od 225.000 sati. Važno je napomenuti da su oba kvara

nastala uslijed vanjskih utjecaja koji nisu direktno vezani za nedostatke u samom sustavu.

D. Trenutno stanje sustava

U vrijeme pisanja članka, sva računala su ispravna i sustav je pod punim radnim opterećenjem. Neka od računala u memoriji imaju učitan operativni sustav duže od šest mjeseci, a izuzevši navedene kvarove, sva dosadašnja gašenja računala su izvršena zbog održavanja i nadogradnji sustava.

U vrijeme redovnog obilaska lokacija, pojedina računala su otvarana kako bi se izvršila vizualna dijagnostika njihove unutrašnjosti. Pritom nisu uočene potencijalno problematične komponente, a posebno je zanimljivo da je unutrašnjost računala uvijek bila vrlo čista i bez tragova prašine. Navedeno je posljedica izbora komponenti bez mehaničkih pokretnih dijelova i izostanka strujanja zraka kroz kućište, pa unatoč zahtjevnim uvjetima okoline, komponente računala gotovo i ne dolaze u dodir sa prašinom.

V. ZAKLJUČAK

Odabirom PC kompatibilnih i tržišno dobavljivih komponenti za izgradnju VHF DSC sustava Plovputa postignuta je lakša nadogradnja i održavanje sustava i ušteda u nabavnoj cijeni. Pritom je postojao rizik od nepredvidivih kvarova jer su PC komponente generalno namijenjene radu u okruženju koje je daleko povoljnije od terenskih uvjeta koji vladaju na kotama Plovputa.

S vremenom se pokazalo je odluka bila ispravna, a povećana pouzdanost sustava se postigla pažljivim izborom komponenti i opisanim rješenjima prilikom implementacije sustava. Ipak, došlo je do dva neželjena ispada za koje se može reći da su posljedica više sile, no svejedno je potrebno uložiti maksimalan napor kako bi se smanjila mogućnost ponavljanja takvih događaja. Nažalost, prevencija opisanih kvarova zahtijeva znatne investicije koje do daljnjeg nisu planirane (ulaganje u odgovarajući klimatizacijski sustav na kotama, građevinski zahvati na pojedinim objektima), pa će biti potrebno dodatno obratiti pažnju na navedene lokacije kako se navedeni kvarovi ne bi ponovili.

Osim toga, za lokacije koje su izložene potencijalnom prodoru vode potrebno je dodatno razmisliti o premještanju cijelog računala zajedno sa napajanjem u ormar sa IP44 zaštitom od prodora prašine i vode, čime bi se dodatno povećala otpornost sustava na vanjske utjecaje.

Zaključno, sustav je kroz četverogodišnji rad dokazao svoju pouzdanost, a ostvareno srednje vrijeme između kvarova ispunjava zahtjeve koji su postavljeni prije projektiranja. Uzevši u obzir udaljenost lokacija i uštedu ostvarenu zbog smanjenog broja interventnih odlazaka na teren, sustav se do danas višestruko isplatio.

LITERATURA

[1] International Telecommunication Union, “M.493: Digital selective-calling system for use in maritime mobile service”, web izvor http://www.itu.int/rec/R-REC-M.493/en (19.03.2013)

[2] Via Technologies, "EPIA SP series mini-ITX mainboard operating guide", web izvor: http://www.viaembedded.com/servlet/downloadSvl?id=261&download_file_id=12674 (15.1.2013)

[3] General electric Company, “Netpro UPS user manual”, web izvor: http://sg.is/gogn/netpro_operation_manual_0k6___1k5_va.pdf (25.1.2013)

[4] SanDisk Corporation, “SanDisk compact flash cards” , web izvor: http://www.sandisk.com/products/memory-cards/compactflash/ (15.1.2013)

[5] Samsung Electronics, "Application note for NAND flash memory, rev. 2.0", web izvor: http://www.datasheetarchive.com/SAMSUNG%20NAND%20Flash%20Qualification%20Report-datasheet.html (25.1.2013)

[6] Debian user forums, „HOW-TO: New and improved RAMboot for speed and silence“, web izvor: http://forums.debian.net/viewtopic.php?t=29774 (15.1.2013)

[7] The linux cross reference, “Tempfs”, web izvor: http://lxr.linux.no/#linux+v3.7.5/Documentation/filesystems/tmpfs.txt (28.1.2013)

[8] The Linux Kernel Archives, „Using the initial RAM disk (initrd)“, web izvor: http://www.kernel.org/doc/Documentation/initrd.txt (28.1.2013)

CIS

International Conference on

INTELLIGENT SYSTEMS

Steering Committee

Chairs: Slobodan Ribarić, University of Zagreb, Croatia
Andrea Budin, Ericsson Nikola Tesla, Zagreb, Croatia

INVITED PAPER

Let the Machines Do. How Intelligent is Artificial Intelligence?

Invited paper

Hans-Dieter Burkhard
Institute of Informatics, Humboldt University Berlin, Germany
hdb@informatik.hu-berlon.de

Abstract - "Intelligent" systems are present around us. Such machines are not only tools in our hands, they are able to make decisions and perform actions directly in the real world or in the artificial worlds of the internet. But even in the latter case, their decisions have consequences to our life. Many of them still act as "assistance systems", leaving the final decision to the human user. The trend goes to more autonomy of the machines, even in critical situations when humans become overloaded by complexity. Additionally, humans are more and more willing to accept the proposals of the machines. But is this technique mature enough to guide or even to replace human decision-making? Especially, perception appears to be a hard problem for technical equipment. How accurate, how safe can decisions be in the case of incomplete and unreliable data? The paper gives some overview about recent developments in Artificial Intelligence and Robotics, their capabilities and serious problems from a general point of view.

I. INTRODUCTION

Twenty years ago, it was common understanding that Artificial Intelligence has failed to reach its goals, and its usefulness was considered to be moderate. Nowadays, intelligent robots are announced that might replace and even outperform humans in many ways. Civil life still keeps many restrictions to the autonomy of machines, but the borders are moving as machines become better and more reliable. A lot of new methods have been developed in the last decades [14], and we are investigating the design of natural systems in order to copy them with growing success [12].

We belief in the information collected and presented by computers, we depend on the functioning of traffic control systems, we trust in the results of medical devices. But we are still waiting for intelligent household aids – the existing "intelligent" vacuum cleaners have less intelligence than bacteria. Especially, perception appears to be a hard problem for technical devices. The main problem is the processing of incomplete, unreliable and even inconsistent data, which has to be performed in short time with restricted computational resources. How safe are decisions in case of unsound information? When we copy natural systems, we know that all natural systems are fallible. Are machines allowed to make mistakes by their design? What are the consequences of building and using more efficient but still fallible technical systems?

As in the past, the new technique can change social and private life. Ethic debates are necessary about possible consequences of the already existing and of the forthcoming technical developments. While a self development of the technology without human influence might be still far away, the usage of the techniques by humans need to be investigated and discussed on a wide spectrum. It will affect the future of mankind, but the directions can be controlled. Recent intelligent machines can be used to survey and to control humans – but for purposes of other humans, not by intentions of the machinery itself. Intelligent machines will furthermore overtake human jobs. If this will be beneficial does not depend on the technique, it depends on the usage of the automation profit.

The paper gives some overview about recent developments in Artificial Intelligence and Robotics, its capabilities and its problems from a general point of view. After a discussion on the notion of intelligence, the paper will follow the main aspects of machine intelligence: action (mainly locomotion), perception and control of intelligent machines. Hopefully, the paper can provide some insight into certain trends in Robotics and Artificial Intelligence as basics for technically informed discussions and for further investigations of consequences.

II. ON THE NOTION OF INTELLIGENCE

It is difficult to define, what an "intelligent machine" is like, especially because there are no unique definitions of natural intelligence. Some old fashioned arguments have stated that tasks cannot be intelligent if machines are able to do it. Actually, machines are better in tasks that are commonly accepted as "intelligent behavior" like complex calculations, playing chess, memorizing huge amounts of information. Even in language translation (e.g. [6]), recent systems outperform humans with respect to the number of available languages. But the quality of translation is still behind the performance of humans. The machines don't understand the meaning, they simply do statistical calculations over huge multilingual corpora. They are not able to do conversations in a human style. Such skills were requested by the well known Turing Test [16]. By that, the performance of a machine is comparable to human intelligence if it is not distinguishable from humans by conversations. It was often questioned if

systems which do only symbolic calculations can have human like intelligence (cf. e.g. [15]). The arguments, that computing machines by their physical properties can do only binary decisions are obviously wrong as the various approaches in Soft Computing have demonstrated. Machines can go far behind symbolic calculations, but up to now they miss the grounding in the physical world. It might be the case that this is the key to real intelligence

Many concepts of human intelligence are focused only on mental capabilities, on reasoning and knowledge representation – just similar to the concepts of classical symbolic Artificial Intelligence. Intelligence is understood more or less as using the mind to do rational actions, i.e. only on "higher" cognitive levels. Crossing a street, playing with a ball or simple conversation are not understood as intelligent behavior. Fast running, powerful striking, climbing a tree, catching an insect can be done by animals, even better than by humans. Hence it is often not included into intelligent behavior – especially when intelligence is considered as the distinguishing characteristics of human beings. But researchers have learned while developing Artificial Intelligence, that it is comparably easy to implement journey planners, electronic market places, computer games etc. But daily tasks like observing the environment by looking around, speech understanding, bicycling etc. are much more difficult for machines and offer many still unsolved challenges.

Bodily tasks need much more than forces and muscles, they need complex controls which actually must be trained over long times e.g. by dancers, athletes, surgeons, craftspeople. This constitutes the bodily intelligence (cf. [5]). At the same time, the performance of bodily tasks is shaped by the “design”, by good interactions between the physics of the body, the environment, and the controls, such that optimal behavior can emerge by their interplay.

The "cybernetic turtle" requires only light-sensitive sensors and separately driven wheels. The sensors and the drives are connected via cross: If the left sensor gets more light, then the right wheel moves faster and the turtle moves to the left. If the right sensor is closer to the light, the vehicle moves accordingly to the right. Overall, the machine always moves towards the light. If the light moves, the vehicle will pursue it. If sensors and motors are connected left to left and right to right, then the turtle will move away from the lights. Now several turtles, some of them with lights on their backs, can be placed together in one room. What happens looks quite complex: Some turtles seem to like each other (those with crossed couplings and lights on the back), others (with parallel connections) seem to be afraid from light carrying turtles and so on. Human observers tend to ascribe even further mental capabilities like tactics and learning, or affections like altruism, brotherhood, revenge. Some primitive intelligence seems to rule the behavior, while only a suitable coupling of sensors and actuators is used. Nature has many examples of organisms that realize very complex behavior in a simple manner. Primitive behavior can be built following the related principles [3].

Artificial Intelligence and Robotics are just going to learn about “intelligent designs” which are necessary for complex bodily intelligence. Moreover, recent AI and Robotics claim that intelligence needs a body to literally grasp the world. The body is necessary to have physical groundings of meanings, and to learn about the own body and its environment by bodily experiences. In the following, we will discuss the capabilities of recent artifacts for locomotion, sensing, and control, respectively.

III. LOCOMOTION

Differences in locomotion concern unmanned air / ground / surface / underwater vehicles (UAV, UGV, USV, UUV). Locomotion in the open air and in the open water is comparatively easy, as long as the conditions do not change fundamentally: as long as direct movement to a destination is possible. The respective control for the compensation of disturbances like drift or wind is relatively easy. It can be performed similarly to the principles of the cybernetic turtle. The situation becomes more complicated if obstacles have to be avoided by appropriate maneuvers. This needs more efforts on perception and control. When humans are not needed on board for control, the machines can be much smaller. There are both insect-like machines and large unmanned vehicles.

Miniaturization is especially suitable for reconnaissance in the vicinity because of their limited reach, e.g. for the search of victims in a devastated environment. Deficient equipment of the individual machines can be compensated by simultaneous transmission of whole robot swarms: Just like in a state of insects, only the success of the system as a whole is relevant, the individual counts nothing. Successful behavior can emerge from the interaction of primitive individuals like in the following example.

The search for victims in a disaster area is difficult because usual routes are destroyed or blocked by debris. Only miniaturized machines can move through cracks, gaps and holes which have remained. Their way is more or less by chance because of missing landmarks and broken connections. But the random movement is even desired for a good coverage of the terrain. The machines may prefer certain directions. Some individuals are attracted by sound, others are attracted by heat, the third kind may follow the concentration of carbon dioxide. If they come to a place where their "lure" is sufficiently strong, they send signals to the outside world (possibly by broadcasting with the help of the swarm). Many individuals will get lost hopelessly, others will achieve unimportant goals. But in all places where sufficiently many arrive, the probability is high that there are still living persons. The strength of commonly sent signals marks relevant places.

The example shows important elements of “swarm intelligence”: The goal is achieved by simple individual behaviors, but the single individual needs not to know the actual goals. The individual just follows simple rules and acts by chance. Not all individuals need to be successful. Thus, different options can be pursued, even if some individuals are lost.

Locomotion on the ground is not only more demanding than that in the air or in the water, it is also

more diverse. "Normal" vehicles require appropriate paved ways. To follow their course, a good perception and control are required. The DARPA (Defense Advanced Research Projects Agency) has held highly regarded competitions in the United States. The DARPA Challenges and the DARPA Urban Challenge took place between 2004 and 2007 in order to test the state of the art for autonomous vehicles [4]. The recent progress in this area is enormous, and unmanned test runs are performed even on regular roads during normal traffic at different places in the world. The perception is still the most difficult problem. We will discuss it in more detail below.

Tracked vehicles allow rides on rough terrain up to a certain extent. There are also mostly small machines with correspondingly limited range of action that move through rough terrain by rolling, snaking, crawling, creeping or jumping using surprisingly simple but efficient mechanics. Related videos can be found e.g. on the web presence of Boston Dynamics [2].

Special benefits are offered by machines that are equipped with two or more legs following models from nature. Humanoid robots have human shape and human size. It is their advantage that they can act in the environments of people with stairs, doors, switches, faucets and so on. Equipped with suitable manipulators, they could take appropriate actions just like humans. So far their skills are very limited. They suffer under the lack of orientation, and they miss the necessary sensibility for objects and for their own bodies. It has not yet been possible to realize a reasonably robust walk for a humanoid robot. Four legs offer much more stability and robustness even in rough terrain, as for example the "BigDog" from Boston Dynamics already shows [2].

The task of the just ongoing “Robotic Challenge” by the DARPA [4] concerns robots that can act in a disaster area: The robot should come to the disaster site using a utility vehicle, then traverse through devastated area and open a blocked door. Inside the building it must climb a ladder, traverse an industrial walkway and break through a wall using appropriate tools. Finally, it has to locate and replace a defect component of a leaking pipeline. All in all a very challenging task, which needs new hardware and new control software. The exact description of the final competition at the end of 2014 is still open, but teams from prominent institutions are already preparing for the qualifications.

IV. Sensors and Perception

Passive sensors record only information that is created in the environment, as for example by light. Thus, they depend on environmental conditions, e.g. road markings are no longer visible in case of rain, snow or fog. Active sensors send signals (sonar, laser, radar, infrared, ...) and measure the reflections.

Touch sensors, close range distance sensors, position, force and acceleration sensors are useful for gripping and manipulating objects. Visual supervision is possible, too. Simple, repetitive movements (such as in automated manufacturing) are readily manageable, but complex movements are only rudimentarily possible with the recent techniques.

Sensor networks with spatially distributed acoustic, optical, seismic, magnetic and other sensors can be used for a comprehensive monitoring of large areas. The devices themselves can be kept very simple. The global picture emerges by the methods of swarm intelligence, or can be calculated from data of different sources by integration methods.

The perception, i.e. the processing of sensory data into meaningful interpretations, is - perhaps surprisingly to some - the biggest problem of machine intelligence. What appears easy for natural beings, is actually a very complicated process dealing with diverse sensory impressions, expectations, attention, and experience of the world. About half of our brain is involved in visual perception. Much of what we belief to see is in fact extrapolated from our knowledge: We detect the driver in his seat behind the wheel of his car passing by. But our eyes have only sensed a side view of a head, a shoulder and a bit of rest behind a car window. We have learned to interpret incomplete and uncertain information (mostly) correctly. We take advantage that information is redundant in multiple ways and complements each other: The incoming information is linked together and synchronized accordingly.

The same applies to the bodily sensations: Our whole body is equipped with sensors, we feel the position of the joints, the tension of the muscles, but usually this remains unconscious. The constant interplay between these sensations and the muscles, tendons and joints allows us to move safely even in difficult terrain. These skills must be acquired through training, and it is not yet managed to transfer it to (humanoid) robots. The previous approaches of robot walking have instead tried to plan in advance the movement of the legs and feet using physical models [17]. Examples are the well known Asimo from Honda [7] and the Kawada HRP-series [8].

The implementation of such planned motions requires fairly precise movements. Elasticity as in natural beings would be a hindrance. The current problems concerning better motions are probably related to the fact that the robots have not sufficiently many sensors or that they cannot analyze the received sensor data fast enough with sufficient quality. It is not irrefutable that machines need to copy the principles of nature - airplanes and cars have no models in nature. Actually, researchers in the field of (humanoid) robots and Artificial Intelligence try to understand more and more the natural model in order to copy it. It may help to construct machines with better performance in complex environments, but the price may be a loss of reliability. Natural systems are prone to errors.

For the perception of the outside world, there are many sensors in technical use, many of them, but not all, with examples in nature such as for vision, hearing or echo measurement of distances. Different information can be combined and analyzed by linking it together. The processing and interpretation of such information are, despite successes already achieved, still very imperfect in comparison to nature. A digital image consists of many individual points which allow an interpretation only by regarding their mutual relations. Millions of pixels must be compared with their neighbors in order to find certain

features, contours, structures etc. One and the same object appears differently if seen from different viewpoints and under different lighting, it may be partially obscured and it has different relations to other objects in different environments. To identify an object, it must be known in advance with its characteristics. From an image taken by a camera, the features must be extracted and compared with stored characteristics in a database or they must be analyzed by appropriate algorithms. However, not all features need to be known in order to identify an object. It is enough if it is sufficiently well distinguished from other objects by some features found in the image. Nevertheless, it always remains the possibility that there is an object not known by the machine. Then there is no chance for correct identification. Moreover, if someone wants to cheat, it is not so difficult to mislead the machines.

In special cases, there is already astonishing success. It manages quite well to recognize faces in pictures and assign them to known persons. First, regions in an image are determined that show certain general characteristics of faces. For these regions, assignments to persons are made using a data base of tagged images: Specific characteristics (e.g., distances and relationships of distinctive parts) are compared for similarity. Then the people are determined which are assigned to the found similar images. This is achieved with high hit rates in frontal shots very well, as one can easily convince themselves for commercial photo-editing programs, or on social networks.

In principle, such methods can be applied for detection and identification of other image content, too, but the creation of such databases is an endless task. Based on related collections of data, methods of Machine Learning can be used to find efficient methods for classification (we will discuss Machine Learning in short in the next section).

Basically, the machine can detect objects only according to its knowledge. Similar things can be recognized up to a certain extent. The complexity increases with the number of objects and with the desired accuracy.

If scenes are to be analyzed, the objects in the scene must be identified first (maybe in a simpler way), and relations between recognized objects have to be detected. Next, their position or vice versa, the position of the robot relative to the objects are interesting. This can be calculated using the information from the image in different ways. Alternatively, the spatial relationships can also be determined by distance measurements: laser scanners provide very good values for distances. Distance sensors are used for example in road traffic collision avoidance. The cheaper (but less accurate) ultrasonic sensors are sufficient in many cases.

To be sufficiently accurate in scene interpretation, considerable computing power has to be available. Potentially inaccurate and incomplete information needs to be analyzed and evaluated. Sensor and processing failures, insufficient information and ambiguous data can lead to wrong identifications and to wrong interpretations. High accuracy is not at least a question of resources and available time. Up to now, machine performance in the interpretation of situations is still far behind human capabilities, even in case of extensive use of technology.

For the detection of movements, consecutive observations are analyzed. The appearances of the same object in the images (or other sensed data) have to be determined at first. Then the changes in the consecutive images allows for speed calculations. The positions and sizes in the images or the distance measurements change, e.g. for an approaching vehicle. The own motion of the observing robot can be calculated from changes in the relationships to other objects. Alternatively - or in addition - the own movements can be determined from motor commands, acceleration sensors and measurements of footsteps or wheel rotations. In general, all these data are flawed and often even contradictory. The algorithms usually calculate only probabilities of possible hypotheses. Again, high precision comes by high expenses: depending on the speed, several measurements must be evaluated per second.

In order to determine the own position of a robot, e.g. according to a given map, reference objects are required. The satellites of GPS are used as globally available landmarks. Navigation using GPS needs a map with GPS references, e.g. for roads and destinations. The limits are reached where the satellite signals cannot be received or where the precision is not sufficient.

Other opportunities for orientation are perceivable reference points in the area. Again, it needs given knowledge about the environment, e.g. about the location of buildings along a planned route. In the experiments with autonomous vehicles by Google [11], comprehensive material is used that combines maps, tagged views, and GPS data together. In particular, the positions of traffic lights are known exactly. The recognition of traffic lights from a sea of lights in a city at night would exceed today's technology. Conversely, this also means for the next future that machines in an unfamiliar terrain or in unexpected situations may not adequately respond.

V. Control and Autonomy.

The control has to run activities according to the demands of the perceived situation. This requires both the ability to change behaviors actively, and an adequate interpretation of the situation. Simple controls require little effort. As with the cybernetic turtle, a measured deviation from the course can lead to a correction corresponding to an intended destination: While the destination is seen in the middle of an image, the direction is correct (assuming that the view is pointed straight to the front). If it is not in the middle, the direction must be corrected accordingly. The strength of the correction can be determined by methods from control engineering. Other sensors can be used instead of a camera as long as they can identify deviations from the intended direction. To avoid obstacles, they must be recognized (e.g. by distance sensors) such that evasive maneuvers become possible. Afterwards the adjustment to the destination is required again. Such controls are robust but not very flexible. Errors are easily possible.

It is useful to distinguish between short-term and long-term tasks. Short-term tasks concern the holding of a chosen course, stabilization, obstacle avoidance if needed, etc. This requires rapid decisions. Remote controls are only of restricted usefulness. In many cases, modern technology is superior to human beings with respect to accuracy and speed, - always assuming that the necessary data can be collected and processed fast enough.

For development and programming of such controls, the possible situations and required responses must be known in advance by the developer. They must be implemented into the hardware and software of the machine. Alternatively, the machines can be trained using machine learning methods. A supervisor presents examples of situations and actions to be performed. By certain generalization features, the machine "learns" to react accordingly in situations which are similar to the presented ones. Other methods provide the machines with some fitness measures. Such measures allow them to experiment autonomously and to learn from the achieved fitness values. For example, a humanoid robot can evaluate its walking speed, its stability and its energy consumption for different kinds of motions and learn to use the optimal ones after a sufficient number of trials.

The generalization allows an extension of trained behavior to similar situations, but it is usually difficult to predict accurately how far the generalization will follow the intentions of the teacher, i.e. of the developer of the robot. In principle, the machine will act as required only in situations which are known in advance (or at least similar enough to known situations). If unforeseen situations occur, the behavior is usually unpredictable.

Longer-term tasks of control relate to the setting of objectives, to planning and activities for achieving goals. There are basically two possibilities: control by humans and complete autonomy, while recently often a compromise by semi-autonomous controls is preferred. Humans decide about tasks and subtasks, while the machines perform sensor processing and motions by their built-in controls. The human operator may work remotely, like e.g. for space missions or dangerous environment. The above mentioned DARPA Robotics Challenge allows semi-autonomous control by a "non-expert" human operator.

The trend to more autonomy of the machines seems to be unavoidable as soon as it appears technically feasible with sufficient accuracy. There are several reasons. One disadvantage of remote control is the fact that the data paths are another source for unreliability by technical problems or manipulation. Highly qualified people are usually needed for the remote control of one high-tech machine. This results in a bottleneck if numerous machines are to be controlled. Several machines controlled by one human team will necessarily require more autonomous action of the machines.

But currently, more autonomy holds still many technical risks. The reason for this is mainly the above-discussed problem with perception. The correct interpretation of complex situations is difficult enough for people, for today's machines it is not practically feasible. We have already autonomous robots in certain open environments, e.g. for transportation, cleaning, and entertainment. These robots simply must stop when people come close (as far as they are big enough to cause harms to people). More complex interaction with humans would need more differentiated and precise reactions. Humans as social beings are trained from early childhood to understand the actions and intentions of others based on communication, bodily signals, emotions, and conventions. For machines that is not yet foreseeable.

Overall, the division of labor between man and machine will probably remain for a long time - combined with the tendency to assign more tasks to the machine. But even if man is still involved, he becomes increasingly dependent on the machine. It is difficult to ensure that humans are the actual masters of decisions: The man pressing a button depends heavily on the information provided by the machines. He is forced to trust in the machines, because he is usually in compliance with his orders if he follows the provided scenarios. The responsibility is in fact up to those who direct the use of machines and to those who construct them. But there can be great local and mental distances between the users of machines and those who are responsible on higher levels, and there is little chance for sanctions. At the same time, if sufficient information is available, machines could follow well defined rules more precisely than humans could.

The problems and advantages of the machines should be known and regarded, but other interests like economical or political power may overwrite related demands. Protagonists of war machines even use the questionable argument that ethical behavior can be programmed into lethal machines making them more ethical than human soldiers [1]. The argument is questionable because it is not foreseen how the necessary information can be provided under the conditions of warfare, and the interests in hesitant machines on a battle field might be low. Such a machine could then refuse to follow the advices of man - they could simply refuse the command. Such a behavior is unique to Stanislaw Lem's "The Trap of Gargantius" [10].

VI. Design, Construction, Programming, and Testing

Prior to the use of machines there is a long process in which hardware and software are designed, implemented, and tested. For all related steps, there are numerous methods and tools for supporting the work to be done. Classical approaches are based on clear ideas and models for their implementation. Before a product is released, it has to pass numerous tests on its functionality and compliance with the specifications. This is a long-term and expensive process. For machines operating in the environment of people, there are extensive regulations. However, there is no total guarantee for secure functioning even in the classical fields of application.

For (intelligent) machines used in the real world, the classical methods of development and testing are confronted with additional limitations. In apparently similar situations, the sensors can produce different values. Even with high redundancy, there are always borderline cases. The examination of those problems and

their effects to security has just started a few years ago, e.g. in the programs on "Cyber Physical Systems" [9].

It has already been discussed that complicated behaviors cannot be programmed directly into (testable) algorithms, they have to be learned by means of training examples. Some evidence for the correct work of trained behavior can be achieved only by additional test examples. For untested situations, in particular for unpredictable situations, it is therefore not clear how the machine will respond.

The existing doubts in more complex intelligent systems is still expressed by the fact that some systems (such as in traffic) are available only as "Assistance Systems" - the responsibility remains with the person that uses them.

To be consistent with law, autonomous systems would need to know about such laws. If the respective behavior and conventions could be implemented, it could even be possible to verify the consistence with the law by third parties. But as testing is difficult in principle, they same would hold for such evaluations. Violations of rules would be difficult to verify – it could be interpreted as unexpected misbehavior e.g. because of missing sensor data. Despite this pessimistic view, related conventions should be discussed. They can help to clarify the situation with autonomous systems.

VII. CONLUSION

Modern machines are efficient. They do not get tired, they are not guided by emotions. It is assumed that they act strictly according to their implementation, but this cannot be verified exactly. Also in terms of intelligent behavior, they are superior in some respects to humans: They can collect and process a lot of data, perform calculations quickly and react instantly. But still they have significant deficits in perception, in evaluating situations and in making decisions in ambiguous or unpredictable situations.

But even if the autonomy of the machine is constrained by final human decision, the activity of the human actor depends on the machine. Actually, there is already a distribution of activities between the human actor and the machine as a technical actor. Both have to be considered as a unique hybrid system [13]. There is a strong tendency to let the machine do, what it proposes. If it is obvious, that the proposal is bad, then the human actor has no problem to cancel it. But usually there is no clear doubt – otherwise the machine is not really useful. For the human actor it is difficult to ignore the proposal as long as there are no convincing counterarguments. He may need such arguments later for justification to the supervisor. Thus, if humans are to keep the responsibility they need the necessary information behind the proposals of a machine. For that, the machines must identify and present potential problems with the proposed decisions, it should suggest alternatives. This could result in problems with efficiency.

Moreover, the machines must be designed in a related way. It means in its final consequences that the designer of machines receives the burden of responsibility. That goes far beyond the usual tasks of design and implementation. What kinds of machines can be built according to such responsibilities? The designer should be able to predict what situations are possible and what information a person needs to make good decisions. He should know exactly how the effects of distributed activities between the human actor and the machine actor work. The related problems are still under investigation (cf. e.g. [18]).

Actually it is more usual that machines do not decide completely autonomously, there is still a person that can affect the decisions. Nevertheless, the trend goes to more autonomy of the machines, but up to now, the proper functioning cannot be guaranteed for such complex machines.

REFERENCES

(Web references as from March 2013)

[1] R.C.Arkin, Governing Lethal Behavior in Autonomous Robots, CRC Press 2009.

[2] Boston Dynamics, http://www.bostondynamics.com

[3] V. Braitenberg, Vehicles. Experiments in Synthetic Psychology. MIT Press, Cambridge, Mass. 1984.

[4] DARPA Challenges,
Urban Challenge: http://archive.darpa.mil/grandchallenge/
Robotics Challenge: http://www.theroboticschallenge.org/

[5] H. Gardner, Frames of Mind: The Theory of Multiple Intelligences, Basic Books 1983.

[6] Google Translation: http://translate.google.com/

[7] HONDA: http://world.honda.com/ASIMO/

[8] KAWADA: http://global.kawada.jp/mechatronics/index.html

[9] E. Lee, "Cyber physical systems: design challenges," University of California, Berkeley, Technical Report No. UCB/EECS-2008-8 (January 23, 2008).

[10] S. Lem, "The First Sally, or the Trap of Gargantius," in: The Cyberiad. MarinerBooks, 2002, pp. 31-42. (Original Polish 1965).

[11] J. Levinson, J. Askeland, J. Dolson, and S. Thrun, "Traffic light mapping, localization, and state detection for autonomous vehicles," IEEE International Conference on Robotics and Automation (ICRA) 2011, pp. 5784 – 5791.

[12] R. Pfeifer and J. C. Bongard, How the Body Shapes the Way We Think – A New View of Intelligence. MIT Press 2006.

[13] W. Rammert and I. Schulz-Schaeffer, „Technik und Handeln - Wenn soziales Handeln sich auf menschliches Verhalten und technische Artefakte verteilt," in: W. Rammert, I. Schulz-Schaeffer (Eds.): Können Maschinen handeln?, Frankfurt/M., Campus 2002, pp. 11-64. (in German).

[14] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd Edition. Prentice Hall 2009.

[15] J. R. Searle, "Minds, brains and programs," Behavioral and Brain Sciences 3, pp. 417–424 (1980).

[16] A. Turing, "Computing machinery and intelligence," Mind 59(236), pp. 433–460. (1950).

[17] M. Vukobratovic and B. Borovac, "Zero-moment point – thirty five years of its life," International Journal of Humanoid Robotics Vol. 1, No 1, pp 157-173, 2004.

[18] H. Wandke and J. Nachtwei, "The different human factor in automation: the developer behind vs. the operator in action," in D. de Waard, F.O. Flemisch, B. Lorenz, H. Oberheid, and K.A. Brookhuis (Eds.), Human factors for assistance and automation, pp. 493-502. Maastricht. Shaker Publishing, 2008.

PAPERS

Some Security Considerations on Crowd-Sourcing an Ontology

Prof. Mauro Migliardi
Centro Ingegneria Piattaforme Informatiche (CIPI)
University of Genoa and University of Padua
Via Opera Pia 13
16145 Genoa, Italy
mauro.migliardi@cipi.unige.it

Prof. Eva Riccomagno
DIMA
University of Genoa
Via Dodecanneso
16100 Genoa, Italy

***ABSTRACT* – The 4w project has demonstrated the feasibility of a prospective memory support systems based on a mobile client and an ontology based server side reasoning engine. However, the experiments also showed that the level of efficiency of the system was largely dependent on the precision and richness of the ontology used to autonomously identify the places where users may efficiently perform their tasks. Thus, in past works we have proposed a social, crowdsourcing approach to the task of enriching the ontology. In this paper we analyze the security aspects of the crowdsourcing approach, we investigate some possible attacks to a crowd-sourced ontology and we devise some measures to protect the system from this kind of attacks.**

Keywords- Green Networking, Distributed Intrusion Detection, Simulation, Evaluation

I. INTRODUCTION

Memory is one of the most important components of human self-consciousness and intelligence. As such, it has been the subject of study in biology, cognitive science, as well as in psychology. A first level categorization, usually divides memory into memory of the past [1] (retrospective memory, the most commonly known conceptualization of memory) and memory of the future also known as prospective memory [2]. Prospective memory is the memory function that allows a human to remember to perform a specific task in a given time frame (time based prospective memory) or in presence of a specific situation (event based prospective memory) [3].

Prospective memory is fundamental in everyday life as it is the function that allows performing both trivial tasks such as remembering to put back the cork on the water bottle or buy the bread for dinner and critical ones such as taking a life-saving medication or restocking for it when it has run out. It is well-established that healthy elderly adults have difficulties with episodic memory [1] and with other cognitive abilities related to executive functions, such as working memory, attention, and task switching [4]. Elderly adults are also more prone to distraction during cognitive tasks. Such difficulties may all contribute to reduce prospective memory capacity [5], [6].

We argue that in our aging society [7][8] a system supporting prospective memory may enhance the capability of elder adults to live independently.

We have designed, prototyped and tested a system capable of providing users with geo-referenced and time-referenced reminders. The system can simply act has a traditional container of to-do lists, in fact it can collect and store user defined tasks and needs through a touch and voice interface; moreover, the system is capable of autonomously scanning Geographical Information Systems to investigate the surroundings of the user and, associating tasks relevant for the user to newly discovered locations, suggest nearby opportunities to efficiently complete the tasks. This latter capability requires reasoning functions that we have obtained leveraging ad-hoc ontology. A first campaign of tests of the system showed that our initial hypothesis is correct and that the provision of timely and geographically contextualized hints improves personal memory efficiency. (see [9], [10], [11], [12]).

Furthermore, our experiments showed a significant dependency of the accuracy of the hints from the richness of the ontology. Thus, in [13] we described our crowd-sourcing approach to the extension and enrichment of this ontology. In this paper we formalize it from a statistical viewpoint and we refine our initial implementation. We suggest that the updating of the ontology should be based not only the values of a point-wise estimate of the truthfulness of a relation but also on a so-called confidence interval.

This paper is structured as follows, in section II we provide a brief introduction to the 4W project prospective memory support system; in section III we briefly summarize why and how we crowd-sourced the enrichment of the ontology; we describe our security concerns, a possible solution and we discuss suitable confidence intervals; finally, in section IV we discuss our future works to improve the system and we provide some concluding remarks.

II. THE 4W PROJECT SYSTEM

The functional flow of our system may be described as follows:

1. Obtain/Receive the input from user in form of natural language;
2. Convert input into structured forms and assign suitable priority
   a. Understand which task the user has to do
   b. Provide a suitable mechanism to assign priority
3. Track user space / time context
4. Query GIS about the possibility for the user to accomplish tasks in his current location
   a. Query the GIS
   b. Collect replies from GIS
5. Present/suggest solutions to the user
   a. Aggregate replies from GIS
   b. Filter the results and present them to user

In order to save one of the most limited resource of mobile clients, namely battery power, we have designed our architecture so that the server is responsible of handling most of the computationally intensive tasks, while multiple clients, connected to the server via HTTP, mainly manage sensorial tasks (e.g. positioning) and interface tasks (e.g. gathering user input on new task and presenting hints to the user). The server is connected both to GIS and to third party service providers (e.g. Google). This connection allows our application to provide additional services such as saving user's tasks in Google Calendar, and retrieving maps from Google Maps. All access to third party providers has to go through the server, therefore, the server needs to be able to handle multiple requests concurrently. In order to do this, we chose to implement the server as a web service, more specifically a RESTful web service.

The clients are designed to be run on a mobile device that can support wireless internet and has a GPS module that devises the user's position with a reasonable precision (about ten meters). Most current mid and high-end mobile devices support these requirements. The client is mostly used to display and manipulate the data that is retrieved from the server and save any changes back to the server. To select the mobile development platform, we mainly take into account the market growth trend and the ease of integration with external GIS services. Android has shown a very fast growth in the global market and provides a high level of integration with Google Services, which can be leveraged to ease the application development. For these reasons, we choose the Android operating system. Because of this choice, we adopt design paradigms specific for development in Android platform, such as Activity (A unit of screen view in Android platform. Every screen that the user can see is an instance of Activity Object) and Service (A background process that keeps running quietly together with the main application, a concept similar to the idea of a service in PC operating systems) are integrated in our system.

Although recent studies (e.g. [14]) suggest that there may be different dimensions to innovate HCI in order to simplify inserting new tasks in the system, we decided to introduce a pseudo-natural language processing interface into our system. Natural Language Processing presents many open problems quite often extremely computationally complex [15].

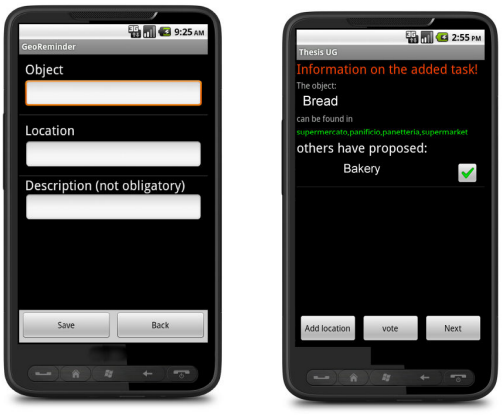
However, the Natural Language Processing group from Mozilla Foundation had a project called Ubiquity [16] that aimed at bringing a natural language interface to a web browser. This project, already ended, contributed a unique way of dealing with complexity in natural language processing. Instead of treating input as just a plain string, the system processes it as structured data. This allows multiple formats or views to be passed around as a single object, which is more flexible and ease further processing. Furthermore, this object can return a textual representation when needed.

Some of the ideas behind the development of Mozilla Ubiquity Parser 2 are documented in the official pages, and developers' blog posts. We take inspiration from it for the workflow of our system.

The approach proposed by Ubiquity does not try to guess the grammatical structure of sentences as many NLP systems do; on the contrary it matches the user sentence against a set of predetermined grammatical structure and chooses the best match. Obviously, this technique cannot cope with unexpected structures; however it is robust enough to fulfill our needs, i.e. to grasp the meaning of requests that are, in most cases, small variations of a limited set of structures.

Further explanations of the theory behind their design considerations can be found in official pages of Mozilla Ubiquity Parser 2, while a detailed description of how we used it can be found in [12].

To test and evaluate our prototype, we performed an experiment with some real users. The experiment consisted in providing a course and a set of tasks to be performed in a given time to two groups of users: the first group had no support system, the second group had our support system. Examples of task are buying goods, finding a post office to send a letter, making a reservation for theater tickets, etc. The system provided the users in the second group with time and location contextualized hints as support to efficient performance of the tasks. The experiments clearly showed that, even in a limited space-time frame, our system enhanced users efficiency in terms of number of remembered tasks, time needed to perform them, and distances walked to perform them. The complete analysis of the results of our experiments has been performed in



a) b)

Figure 1 a) Mobile interface used to request to the user an Object-Location association; b) Voting Interface

[12] and is beyond the scope of this paper, here we will focus only on the different level of efficiency that the system demonstrated in relation with the richness and completeness of the part of the ontology related to the tasks the user had to complete.

It was clear, as a matter of fact, that the system delivered significantly different levels of support to different tasks. In some cases, the system was capable of providing suggestions that allowed an almost immediate (both in time and space) completion of the task; in other cases, however, the systems provided suggestions that were less frequent (both in time and space) and required to spend more time and walk further distances.

To identify the source of this behavior, we analyzed and profiled the flows of data inside the system itself. This process allowed us to tie the phenomenon to the different levels of completeness of the ontology with regards to different types of tasks. In fact, the ontology represents the core of the technique that the system adopts to classify a user need and to relate it to one or more categories of site.

This fact showed us that, in order to efficiently find suggestions to the user, our system needs an ontology that is both as close as possible to completeness and continuously evolving to accommodate changes in the real world. The problem of ontology enrichment is a hard one [17] and far from being completely solved. The problem is usually tackled either by leveraging the support of experts task-forces [18], [19], [20] or through statistical modeling of large text corpora [21], [22], [23].

Neither of these approaches, however, suits our needs. In fact, the large text corpus of user requests onto which training would be performed is not available yet. On the other side, it is impossible to build a task force of experts because it is not possible to identify a single, clearly defined domain inside which users' needs shall be expressed.

For these reasons, we decided to try a different approach and tap into cooperative, user generated ontology enrichment.

III. CROWD-SOURCING AN ONTOLOGY

Crowd-sourcing is a concept that first appeared on the Wired magazine in 2006 and was first precisely defined in [24], and it has often been referred to as the process of tapping into the collective wisdom of large bodies of human being [25]. In practice, it is a form of outsourcing that delegates a task to an undefined (both in identity and in size) group. Recent examples of this approach can be found both in social network project (e.g. foursquare, trip advisor) and in industrial experiences such as the one described in [26].

In our project, we decided to keep the process of enriching the ontology integrated into the system instead of making it a separate component. Thus, whenever a user expresses a need or task that the system cannot find in the ontology an association for that kind of task with a category of places to search in the GIS, the system presents to the user an interface that asks the user himself to suggest such an association. As an example, consider the case where a user expresses the need to buy a cake. If the system cannot associate such as task by means of the ontology, than it asks the user to suggest a category of places where he thinks he can buy a cake. The suggestion, in our case it could be "I can buy a cake in bakery", is both used to find places where the current user can fulfill his needs and it will generate a new association in the ontology for future usage. In Figure 1a we can see the interface for the insertion of Object/Location suggestions by the user. The system also supports the insertion of Action/Location associations, e.g., Jogging at the park, swimming at the pool, etc.

Through this mechanism, the system's capability can be enhanced incorporating the knowledge provided by the users.

III.1 SECURITY AND SAFETY CONCERNS: A VOTING MECHANISM

Our approach, though, has an obvious downfall: not all the suggestions of users can be directly incorporated into the ontology. In fact, users' suggestions may be unfit to be incorporated into the system ontology because they are:

1) Bona-fide mistakes;
2) Malicious attempt to pollute the system's knowledge;
3) True only in a very specific context.

The capability to grasp a clear, unerring distinction between these three cases could greatly help in profiling users; however, while misleading suggestions due to very

specific context may be identified, it is extremely difficult to distinguish between smart malicious sabotage attempts and bona-fide mistakes. Furthermore, it is also true that historical records of users' behavior may provide a first level of profiling and enhance the system capability to categorize the different situations. For these reasons, we have decided to first introduce a mechanism targeted at preventing the inclusion of misleading suggestions inside the system's ontology with no judgment about the nature of the mishap. In future works, once a significant historical record for user behavior has been built, we will introduce user ranking and profiling.

The above mentioned mechanism is simply based on consensus among all users about the correctness of a suggestion. When a user provides a suggestion, this is marked as under evaluation and submitted to all users that may benefit from its use. As an example, consider the case in which a user has suggested that "You may buy bread in a bakery". Each user that needs to buy bread is presented with the voting interface shown in Figure 1b. He may tick the suggestion as correct or mark is down as wrong. In the next section we describe the statistical model we have adopted to evaluate the correctness of a suggestion according to the users' votes.

III.2 A SIMPLE FIXED THRESHOLD MECHANISM

A relation R, e.g. (buy bread at a pharmacy), is included in the ontology if the proportion of true votes is close enough to one and deleted from both database and ontology if the proportion of false votes is high. A simple algorithm proposed in [13] is as follows:

a relation R can be either true or false. Let n be the number of votes received by R and set p1 and p2 to be two nominal values with p2<p1. Let p be the number of true votes divided by n. Then

if p>p1 then move R from the database to the ontology,

if p<p2 then delete R from both database and ontology,

if p2<p<p1 wait for R to receive more votes.

The proportion p will extremely rarely be exactly equal to zero or one, not only because of malicious voters who intentionally votes wrongly e.g. to sabotage the system, but also because there are rare circumstances when unlikely propositions like "bread can be found in a pharmacy" hold true e.g. for the general store of a small village. If, as mentioned above, the choice of a threshold at which promote or delete a relation is rather arbitrary, it can be advised by suitable confidence intervals (IC), a statistical measure of the confidence associated to parameter estimation. In many applications and amply in medicine, parameter estimation is presented in a form of ICs (see e.g. [27]). The larger the IC, the lower our confidence is in our estimation of the proportion and ultimately in the allocation of R to the ontology or otherwise.

A most popular choice of IC for proportion is

$$\left[\hat{p}-z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\hat{p}+z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

where z_c denotes the 100(1−c)th percentile of the standard normal distribution, $\hat{p}$ =X/n is the number of true votes, X, divided by the number of votes, n received by R. This is an interval centered in our estimate of the proportion $\hat{p}$ and whose length depends on the variance of $\hat{p}$ computed under suitable assumptions. Improvements on this basic IC are presented in the literature by varying the center and considering better estimate of this variance. We discuss this next.

A main assumption behind the IC above is that the number of voters, n, is small compared with the number of potential voters. Furthermore, when data are gathered from complex survey, a normal approximation is considered. This is valid when $np \geq 10$ and $n(1-p) \geq 10$ where p is the true probability to be estimated. In our case p and n might be such that those inequalities do not hold as p can be very small or large. In the literature different approaches have been proposed for estimation of small (equivalently high) proportions in order to improve coverage probability and provide more reliable ICs. Recall that coverage probability is the proportion of the times that the interval contains the true value of interest. Ample comparative simulation studies for two-sided CI based on length, coverage probability, asymptotics are in e.g. [27], [28], [29].

In particular, the Agresti–Coull interval [31], [31] performs better than other proposed intervals in coverage and is quite conservative for p near 0 or 1, but is also longer on average [29]. This safeguards us from including in the ontology wrong propositions. Its formulation has the same simple form as the standard CIs above, but with a different centre (to improve coverage) and a modified value for n

$$\left[\tilde{p}-z_{1-\frac{\alpha}{2}}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}},\tilde{p}+z_{1-\frac{\alpha}{2}}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}\right] (*)$$

where $\tilde{p}=\frac{X+z^2/2}{n+z^2}$ with $z=z_{1-\alpha/2}$ and $\tilde{n}=n+z^2$. For a 95% CI $z_{1-\frac{\alpha}{2}}=1.96\approx 2$ and the midpoint of the interval becomes $\frac{X+2}{n+4}$.

A peculiarity of our application is that we are testing a one-sided (null) hypothesis p>p1 (p<p2). Hence we consider a standard modification of (*) substituting $z=z_{1-\alpha}$ to $z=z_{1-\alpha/2}$. This gives the lower and upper one-sided CI

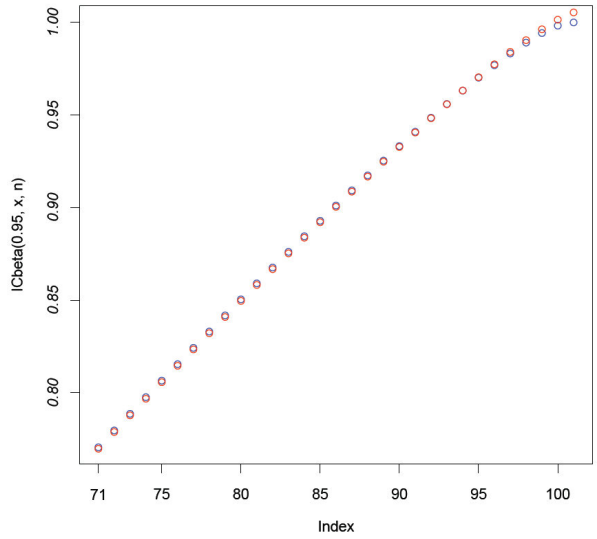


Figure 2 Comparison between a frequentist IC and Bayesian IC

$$\left[0, \tilde{p}+z_{1-\alpha}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}\right] \text{ and } \left[\tilde{p}-z_{1-\alpha}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}, 1\right],$$

respectively. Another strategy advocated in [32] is to consider the one-sided Jeffrey interval, that is one among the possible Bayesian intervals,

$$[0, B_{1-\alpha, X+\frac{1}{2}, n-X+\frac{1}{2}}] \text{ or } [B_{\alpha, X+\frac{1}{2}, n-X+\frac{1}{2}}, 1]$$

where $B_{c,x,y}$ is the 100(1−c)th percentile of a Beta distribution with parameters x and y. Figure 1 shows the left limit of the upper CIs for both cases: in red $\tilde{p}-z_{1-\alpha}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$ and in blue $B_{\alpha, X+\frac{1}{2}, n-X+\frac{1}{2}}$ for n=100, $\alpha = 0.95$ and $X > n/3$. Note that for very high proportion estimates (larger than $\frac{n-2}{n}$, n up to 1000), this lower limit is a poor estimate being larger than one. The relations for which this happens will be included in the ontology with degenerate IC of length zero.

IV. FUTURE WORKS AND CONCLUDING REMARKS

In the present work we suggested to update an ontology not only on a point estimate of the success rate of a relation, but also on confidence intervals. The choice of the type of confidence interval is based on a well-informed literature and a small example has been provided. ICs are very simple to compute and do not provide extra computational cost for the system.

Our proposal can be refined in various directions, at a time and computational costs to be evaluated. We list a few. A relation might have more that two truth values, not only true/false but also maybe with a degree of confidence associated to the length of the CI. Automatic mechanisms can be set up in order to decide the threshold values p1 and p2, based on the number of voters and a categorization of the relations. There is an interesting literature that considers how to evaluate the reliability profile of voters and the robustness of the distributed system to threat scenarios.

In particular the present paper suggests a method to start filling in an ontology and associating to each included relation a measure of the trust in it, namely a CI. For the updating of the ontology, we would like to adopt more refined methods from available reputation and trust systems for online survey [33], as well as techniques from survey theory, decision support system, Bayesian statistics, risk theory. Of course this will be at a computational cost to be evaluated with suitable simulation studies.

REFERENCES

[1] Tulving, E. (1983). Elements of Episodic Memory. New York, NY: Oxford University Press.

[2] McDaniel, M. A., & Einstein, G. O. (2007). Prospective memory: An overview and synthesis of an emerging field . Sage Publications Ltd.

[3] McDaniel, M., Robinson, B., & Einstein, P. (1998). Prospective remembering: Perceptually driven or conceptually driven processes? Memory & Cognition, 26, 121-134.

[4] Grady, C. (2012). The cognitive neuroscience of ageing. Nature Reviews Neuroscience, 13(7), 491–505.

[5] Einstein, G. O., McDaniel, M. A., Manzi, M., Cochran, B., & Baker, M. (2000). Prospective memory and aging: Forgetting intentions over short delays. Psychology and Aging, 15(4), 671–683.

[6] McDaniel, M. A., Einstein, G. O., Stout, A. C., & Morgan, Z. (2003). Aging and Maintaining Intentions Over Delays: Do It or Lose It. Psychology and Aging, 18(4), 823–835.

[7] Chesnais, JC. (1996) Fertility, Family, and Social Policy in Contemporary Western Europe, Population and Development Review , Vol. 22, No. 4 (Dec., 1996), pp. 729-739

[8] World Health Organization (2012), Global Health Observatory Data Repository, http://apps.who.int, accessed on 31 July 2012

[9] Mauro Migliardi, Marco Gaudina, A mobile platform for the improvement of personal efficiency, Proc. of the IEEE Conference on Intelligent Systems, Opatija (HR), 23-27 May 2011

[10] Mauro Migliardi, Marco Gaudina, Active Personal Information Manager: a System for Human Memory

Support, Proc. of the 4th International Conference on Intelligent Interfaces for Human-Computer Interaction, Korean Bible University, Seoul (S. Korea), June 30th - July 2nd, 2011.

[11] Mauro Migliardi, Marco Gaudina, Andrea Brogni, Enhancing personal efficiency with pervasive services, Proc. of the Sixth International Conference on Broadband and Wireless Computing, Communication and Applications, October 26-28, 2011, Technical University of Catalonia, Barcelona, Spain.

[12] Mauro Migliardi, Marco Gaudina, Andrea Brogni, Pervasive Services and Mobile Devices May Support Human Memory and Enhance Daily Efficiency, International Journal of Space-Based and Situated Computing (IJSSC), Vol. 2, No. 3, 2012.

[13] Mauro Migliardi, Marco Gaudina, The 4W (What-Where-When-Who) Project Goes Social, Proc. of The Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, IMIS 2012, Palermo (IT), July 4-6 2012.

[14] A. Gentile, A. Santangelo, S. Sorce, S. Vitabile, Novel Human-to-Human Interactions from the Evolution of HCI, Proc. Of the 2011 International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), Seoul, (SKR) July 2011.

[15] M. Bates, Models of natural language understanding, Proceedings of the National Academy of Sciences of the United States of America, Vol. 92, No. 22 (Oct. 24, 1995), pp. 9977–9982.

[16] J. DiCarlo, Retrospective: What we learned from ubiquity - blog of mozilla labs developer. URL http://jonoscript.wordpress.com/2010/01/20/retrospective-what-we-learned-from-ubiquity/, January 2010

[17] G. Pilato, A. Augello, G. Vassallo, S. Gaglio, "Sub-Symbolic Semantic Layer in Cyc for Intuitive Chat-Bots" Proc. of International Conference on Semantic Computing, 2007. ICSC 2007. 17-19 Sept. 2007 pp:121 – 128, Irvine, California, USA

[18] G. Berio and M. Harzallah, Towards an integrating architecture for competence management, Special Issue: Competence Management in Industrial Processes, guest editors X. Boucher, E. Bonjour, N. Matta, Computers in Industry, vol. V58, issue 2, 2007.

[19] E. Ermilova and H. Afsarmanesh, Modeling and management of profiles and competences in vbes, Journal of Intelligent Manufacturing, vol. V18, pp. 561–586, 2007.

[20] J. Hodik, J. Vokrinek, J. Biba, and P. Becvar, Competences and profiles management for virtual organizations creation. CEEMAS, pp. 93–102, 2007.

[21] K. Neshatian and Hejazi, Text categorization and classification in terms of multi-attribute concepts for enriching existing ontologies, Proceedings of the 2nd Workshop on Information Technology and its Disciplines, pp. 43–48, 2004.

[22] A. F. A and R. Steinmetz, Ontology enrichment with texts from the www, Proceedings of the 2nd Semantic Web Mining Workshop at ECMLI/PKDD, WS'02, 2006.

[23] G. Stumme and A. H. B. Berendt, Semantic web mining: State of the art and future directions, Web Semantics : Science, Services and Agents on the World Wide Web, vol. 4, n. 2, pp. 124–143, 2006.

[24] D. Brabham, Crowdsourcing as a Model for Problem Solving: An Introduction and Cases, Convergence: The International Journal of Research into New Media Technologies 14 (1): 75–90, 2008

[25] J. Surowiecki, The Wisdom of Crowds - why the many are smarter than the few, New York: First Anchor Books Edition, 2005.

[26] W. Ebner, M. Leimeister, U. Bretschneider, H. Krcmar, Leveraging the Wisdom of Crowds: Designing an IT-Supported Ideas Competition for an ERP Software Company, Hawaii International Conference on System Sciences, Proceedings of the 41st Annual , vol., no., pp.417, 7-10 Jan. 2008, doi: 10.1109/HICSS.2008.233

[27] D.C. Montgomery (2001) Introduction to Statistical Quality Control, 4th Edition. John Wiley, New York.

[28] H. Tobi, P. B. van den Berg and L.TW de Jong-van den Berg Small proportions: what to report for confidence intervals? Pharmacoepidemiology and Drug Safety 2005; 14: 239–247.

[29] L.D. Brown, T.T. Cai, and A. Dasgupta. Confidence intervals for a binomial proportion and asymptotic expansions. The Annals of Statistics Vol. 30, No. 1, 160–201,2002.

[30] A. Sukasih and D. Jang, 2005. An Application of Confidence Interval Methods for Small Proportions in the Health Care Survey of DoD Beneficiaries. American Statistical Association Section on Survey Research Methods, Mathematica Policy Research Reports 6583, Mathematica Policy Research. ASA Section on Survey Research Methods DA

[31] A. Agresti and B.A. Coull (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *Amer. Statist.* **52** 119–126.

[32] T.T. Cai (2005). One-sided confidence intervals in discrete distributions. J. Statist. Plann. Inference 131, 63-88.

[33] A. Jøsang, R. Ismail and C. Boyd (2007). A survey of trust and reputation systems for online service provision. Decision Support Systems 43, 618– 644.

Bayesian Posterior Probability Classification of Colorectal Cancer Probed with Affymetrix Microarray Technology

Monika Simjanoska, Ana Madevska Bogdanova and Zaneta Popeska
Ss. Cyril and Methodius University,
Faculty of Information Sciences and Computer Engineering,
Skopje, Macedonia
Email: m.simjanoska@gmail.com, {ana.madevska.bogdanova, zaneta.popeska}@finki.ukim.mk

***Abstract*—Colorectal cancer is one of the most common types of cancer worldwide. Assuming increased or decreased gene expression is the reason for abnormal cells work and processes interference in the colorectal region, in our previous work we used data from Illumina microarray technology to analyse gene expression values. Once we have unveiled biomarker genes and developed methodology for Bayesian posterior probability classification, we proceeded with implementing the same methodology on data obtained from Affymetrix microarray technology. However, our research results showed that different microarray technologies require different statistical approach for classification analyses. In this paper we use colorectal data probed with Affymetrix microarray technology, and propose a new methodology that intends to eliminate the noise and produce more robust preprocessed data appropriate for prior distribution modelling. This allows us to construct an efficient Bayesian a posteriori classificator. In order to test the procedure reliability we used different set of carcinogenic and healthy patients.**

***Index Terms*—Colorectal Cancer, Bayesian Classification, Affymetrix, Illumina, Microarray Technology, Machine Learning**

I. Introduction

The World Health Organization provided a research of the cancer incidence, mortality and prevalence worldwide in 2008. According to the GLOBOCAN project results, colorectal cancer is the third most common cancer in men, and the second in women with a total incidence of 1,234,000 cases, of which, 608,000 deaths. The mortality results make this type of cancer to be the fourth most common cause of death from cancer. Prevalence results showed that almost 60% of the cases occur in developed regions [1].

In this paper we consider the colorectal cancer problem as tightly connected to the gene expression phenomena. We believe that the reason for its occurrence lies in the increased or decreased level of expression of particular genes which intent to disrupt the biological processes they are associated with. Gene expression profiling by microarrays should advance the progress of personalized cancer treatment based on the molecular classification of subtypes [2]. Therefore, in our previous work [3] we showed that using the Illumina microarray technology data in combination with appropriate statistical analysis can lead to unveiling a set of particular genes, referred to as biomarkers, that can be used in building an accurate diagnostic system based on Bayesian a posteriori probability computation.

Assuming platform independence as presented in [4], [5], we proceeded our research in [6] using the same methodology as in [3] on colorectal data probed with the Affymetrix Human Genome U133 Plus 2.0 Array. However, the results showed poor distinctive capability of the biomarker genes. In this paper, we embraced another conclusion derived in [7] which states that each platform requires different statistical treatment.

Hence, we confronted the challenge of inventing a new methodology that leads to unveiling the Affymetrix colorectal cancer biomarkers, eliminate the noise and produce more robust data appropriate for prior distribution modelling that can be used, and therefore, is appropriate for as a part of Bayesian a posteriori classification. In order to obtain reliable results, we used data from patients with different geographical distribution, probed with Affymetrix Human Genome U133 Plus 2.0 Array. Data sets are retrieved from the Gene Expression Omnibus biological database, [8], [9], [10], [11].

The rest of the paper is organized as follows. In Section II we give a brief preview of the related work and the latest results. The methodology used in this paper is described in Section III. In Section IV we exhibit the experiments and the results according to which we derive conclusion in Section V.

II. Related Work

In this section we present some of the research related to the problem of statistical preprocessing methods and classification.

Considering the preprocessing is crucial in pointing out statistical and biological significance, the authors in [12] present a comprehensive study of the effect that normalization, gene selection, the number of selected genes and machine learning method have on the predictive performance of resulting models. The best machine learning methods in this study were Support Vector Machines with the three basic kernel configurations in comparison to Artificial Neural Networks and Decision Trees.

The authors in [13] state that at the moment there is no commonly agreed gold standard pre-processing method and

each researcher has the responsibility to choose one method, incurring the risk of false positive and false negative features arising from the particular method chosen. Therefore, they aim at providing a method of analysis of a gene-expression experiment that combines and synthesises the information from several pre-processing methods, to obtain a better calibrated estimate of differential expression. We agree with the necessity of combining several methods to confirm different gene expression, but we additionally evaluate the efficiency of the methods by including the differently expressed genes in the process of patients classification.

Since we are interested in Affymetrix microarray experiment noise, we present two studies of this kind. The authors in [14] assume that a typical microarray experiment has many sources of variation which can be attributed to biological and technical causes. Their analysis showed that the greatest source of variation at Affymetrix is biological variation, and the variation due to labelling. Similarly, a research on the of systematic noise in Affymetrix and Illumina gene-expression microarray experiments is presented in [15]. The authors suggest that it is not recommended to analyse individual test samples (e.g. to try and classify), but instead to run several experiments at the same time to get a better estimate of the experimental variation.

Gene expression data sets used in this research have also been used in other scientific researches. The authors in [16] aimed to find a metastasis-prone signature for early stage mismatch-repair proficient sporadic colorectal cancer (CRC) patients for better prognosis and informed use of adjuvant chemotherapy. A transcriptome profile of human colorectal adenomas is given in [17] where they characterize the molecular processes underlying the transformation of normal colonic epithelium. One of the data sets has been used in [18] to clarify the difference between MSI and microsatellite stability (MSS) cancers and, furthermore, to determine distinct characteristics of proximal and distal MSI cancers. A similar research is presented in [19] where the scientists showed cross-study consistency of MSI-associated gene expression changes in colorectal cancers.

III. The Methodology

In this section we present the original methodology which includes several steps to prepare the data for the classification process.

A. Preprocessing

We used gene expression profiling of 32 colorectal tumors and matched adjacent 32 non-tumor (healthy) colorectal tissues probed with Affymetrix Human Genome U133 Plus 2.0 Array, which contains 54675 probes, but the unique genes observed are 21050. All the statistical analysis presented below aim to reduce the number of genes in order to distinguish the genes with most biological information, the colorectal cancer biomarkers.

1) Normalization: Since our aim is to unveil the difference in gene expression levels between the carcinogenic and healthy tissues, choosing the appropriate normalization method is essential. Our assumption states that only a small set of genes are differently expressed compared to the biomarker genes, i.e., most of the genes are not correlated to the colorectal cancer. In such cases Quantile normalization (QN) is a suitable normalization method, because it makes the distribution of the gene expressions as similar as possible among each other across all samples [20].

2) Filtering methods: Some genes may not be well distributed over their range of expression values, i.e. low expression values can be seen in all samples except one [21]. This can lead to incorrect conclusion about gene behaviour. To remove the genes of this kind, we used an entropy filter. Entropy measures the amount of information (disorder) about the variable. Higher entropy for a gene means that its expression levels are more randomly distributed [22], while low entropy for a gene means that there is a low variability in its expression levels across the samples [23]. Therefore, we used low entropy filter to remove the genes with almost ordered expression levels.

3) Paired-sample T-test: Knowing the facts that both the carcinogenic and healthy tissues are taken from the same patients, and that the whole-genome gene expression follows normal distribution [24], we used a paired-sample t-test. Assuming that the most of the genes do not have different expressions, the null hypothesis states that there is no statistical difference between the carcinogenic and the healthy samples. The rejection of the null hypothesis depends on the significance level which we determine. In this paper we consider the genes as statistically significant for a p-value less than 0.01, which means that the chances of wrong rejection of the null hypothesis is less than 1 in 100.

4) False Discovery Rate: False Discovery Rate (FDR) is a reduction method that usually follows the t-test. FDR solves the problem of false positives, i.e., the genes which are considered statistically significant when in reality there is not any difference in their expression levels. For a threshold of 0.01 we expect 10 genes to be false positive in a set of 1000 positive genes. The significance in terms of FDR is measured as a q-value. It is described as a proportion of significant genes that turn out to be false positives [25].

5) Volcano Plot: Both the t-test and the FDR method identify different expressions in accordance with statistical significance values, and do not consider biological significance. The biological significance is measured as a fold change [26] which describes how much the expression level changed starting from the initial value. Fold change is measured as ratio between the two expression intensities and does not take into account the variance of the expression levels. In order to display both statistically and biologically significant genes we used volcano plot visual tool. The genes that lie in the area cut off by the horizontal threshold, i.e. the p-value of 0.01 which implicates statistical significance, and the vertical thresholds, i.e. the fold change of 1.68 which

implicate biological significance, are the genes that are up or down regulated depending on the right and the left corner of the plot respectively. When plotting the genes a better transformation procedure is to take the logarithm base 2 value of the expression. This has the major advantage that it treats different up-regulation and down-regulation equally, and also has a continuous mapping space [27].

B. Modelling the a Priori Distribution

Once we have unveiled biomarker genes that discriminate carcinogenic tissue from the healthy one, we proceeded with another series of methods to model the a priori distribution of these biomarkers. In our previous research [6] we showed that the distribution of biomarker genes highly overlaps at both tissues. Therefore, in this paper we used the modified methods in III-A, and obtained another set of biomarker genes that showed different statistical distribution, even visually. However, few more steps are needed to generate more reliable prior distributions.

1) Round-up threshold method: When observing gene expression values, we notice that a large percentage of the data set values are negative. The authors in [28] explain this phenomena within a few processing steps. At the beginning, the background signal is estimated and then it is subtracted from all expression values. The algorithms used in Affymetrix, calculate gene expression values by comparing the signals obtained with perfect-match and one-base-mismatch hybridization oligonucleotides on the microarrays. Sometimes the one-base-mismatch oligonucleotides hybridize to other mRNAs, which does not always gives a good representation for non-specific background signal, thus, negative expression values for genes can result. Negative values might also be results because of noise in the data. One way to remove these genes is to transform all gene expression values below some threshold cut-off value to that threshold value [29]. This method is known as Round-up threshold method. In this paper we used 0 as threshold cut-off value, and any expression value below this threshold is mapped into the interval [0,2]. In order to sustain the prior distribution shape, we want to avoid eventual gene accumulation at one point. Thus, we chose a whole interval instead of particular value.

2) Appropriate tissue selection: Instead of using the cross-validation method for determining the training and the testing set as we did in our previous work [3], [6], here we propose different approach for choosing the training set. Considering the visual representation of the distribution of both tissues, we realized that the data have differently skewed distributions. Therefore, in order to eliminate the possibility of randomly picking up the tissues whose distributions overlap at some genes, we choose the training set according to the skewness factor. If the skewness factors are with opposite signs, then these tissues are involved into the training process.

3) Hypothesis testing: Our generative model fits four types of distributions: Normal, Lognormal, Gamma, and Extreme Value. At first, the two sets of tissues are confirmed to be differently distributed using the Kolmogorov-Smirnov test. After we have fulfilled our assumption for different distribution, we proceed with appropriate hypothesis testing for all the distribution mentioned above. The distribution parameters are estimated using the Maximum Likelihood Estimation (MLE) method, with a confidence level of $\alpha = 0.01$. Then we perform the Chi-square goodness-of-fit test of the default null hypothesis that the data in the tissue (vector) comes from the particular distribution with mean and variance estimated from the MLE method, again with significance level of 0.01. Once we have obtained the probabilities from the testing for each gene distinctively, we choose the distribution whose probability is the highest and we assign it to the particular gene.

Once we modelled the prior distributions, we can use them to calculate the Bayesian a posteriori probability and correctly classify the tissues.

C. Bayesian a Posteriori Classification

Using all the methods explained above, we eventually reached our purpose of building an accurate Bayesian classifier based on posterior probability computation. Modelling the prior distributions, we are now able to compute the class-conditional densities, $p(\vec{x}|C_i)$, which we calculate as the product of the continuous probability distributions of each gene distinctively:

$$p(\vec{x}|C_i) = \prod f_1 f_2 ... f_n \tag{1}$$

For the prior probabilities $P(C_i)$, we defined two test cases:

- Test Case 1: Since we have equal number of tissues into both of the classes, the prior probabilities are also equal $P(C_1) = P(C_2) = 0.5$;
- Test Case 2: The prior probabilities are estimated according to the statistics in [1], and $P(C_1) = 0.0002$ and $P(C_2) = 0.9998$, where C_1 denotes carcinogenic class, and C_2 denotes healthy class.

Therefore, we calculate the posterior probability $P(C_i|\vec{x})$, as:

$$p(C_i|\vec{x}) = \frac{p(\vec{x}|C_i) * P(C_i)}{\sum_{1}^{2} p(\vec{x}|C_i) * P(C_i)} \tag{2}$$

The tissue $\vec{x}$ is classified according to the rule of maximizing the a posteriori probability (MAP):

$$C_i = \max p(C_i|\vec{x}) \tag{3}$$

IV. Experiments and Results

In this section we present the experiments and the obtained results.

Our main goal is to construct an efficient Bayesian classifier that can distinguish healthy from carcinogenic tissue when gene expression levels from Affymetrix DNA chip are used. For our research purpose we retrieved data sets from GEO functional genomics data repository. The 32 colorectal tumors and matched adjacent 32 non-tumor colorectal tissues have

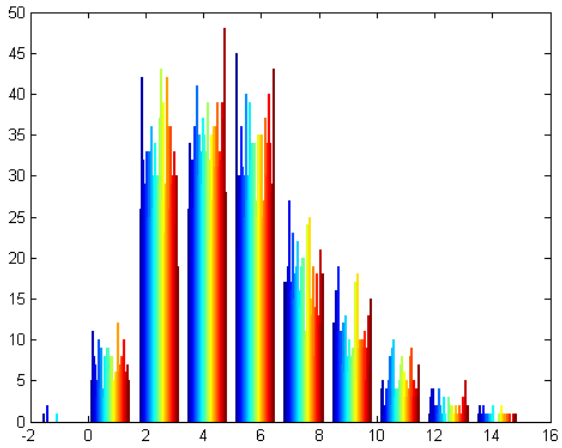


Fig. 1. Biomarkers distribution at carcinogenic tissues

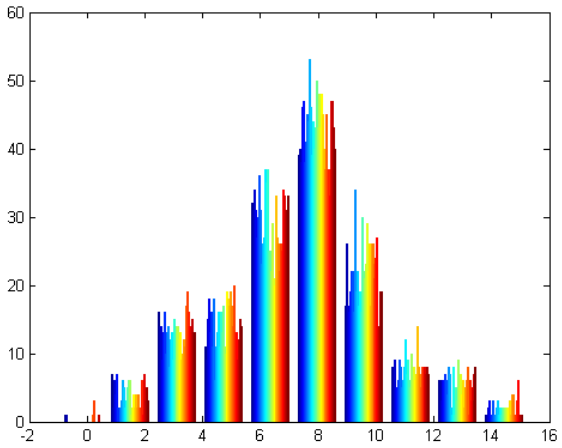


Fig. 2. Biomarkers distribution at healthy tissues

been probed with Affymetrix Human Genome U133 Plus 2.0 Array. This type of technology allows 54675 probes.

We intent to use as biomarkers only the genes which are considered to be statistically and biologically significant. The developed methodology presented in III-A will preprocess the data in order to make the best model of the prior distributions.

After normalizing the gene expression values, we used the low-entropy filter which reduced the number of genes to 49,206. However, this number of gene is still inappropriate to model the carcinogenic and the healthy distribution. Furthermore, we proceed with the paired-sample t-test, the FDR method, and eventually we applied the volcano plot. The results are given in Table I. The final number of revealed biomarkers is 138.

TABLE I
BIOMARKERS REVEALING

Statistical methods	Up expression	Down expression	Sum
Paired-sample t-test	7630	10691	18321
FDR	7581	10539	18120
Volcano plot	29	109	138

The number of genes left to be biomarkers is appropriate for modelling the prior distribution. Using a histogram tool we presented the frequency of the genes within particular range for each of the carcinogenic and healthy tissues distinctively. Figure 1 and Figure 2 depict the prior distributions of the carcinogenic and healthy tissues, respectively. Obviously, the distributions differ from each other. Before we proceed with hypothesis testing, we must cut-off the negative values, since we assume they occurred as a result of a noise.

After the process of rounding up the negative values using a predefined threshold value, both of the distributions depicted in Figure 3 and Figure 4, suffered minor reshape.

Observing the two distributions, we can easily assume that both of them are with opposite skewness. Therefore, to avoid

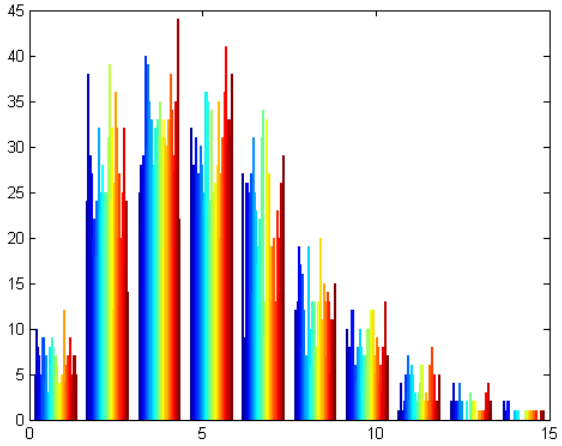


Fig. 3. Biomarkers distribution at carcinogenic tissues after Round-up

the possible selection of tissues with the same skewness by coincidence, we used the method of appropriate tissues selection. This method outcome determined 15 appropriate tissues to be involved in the training process. The distribution of those tissues is depicted in Figure 5 and Figure 6, respectively.

The testing of the biomarkers extracted from the carcinogenic and healthy tissues respectively for equality in their probability distribution, is done by the Kolmogorov-Smirnov test. It confirmed that 135 out of 138 has declined the null hypothesis ”belong to the same distribution”. Hereupon, we performed generative probability distribution fitting by using appropriate hypothesis testing over the Normal, Lognormal, Gamma and Extreme Value distribution. Once we assigned the distribution which showed to be the most probable to each gene, we proceeded to calculate the Bayesian posterior probability as defined in III-C.

Moreover, to confirm the procedure reliability, we addition-

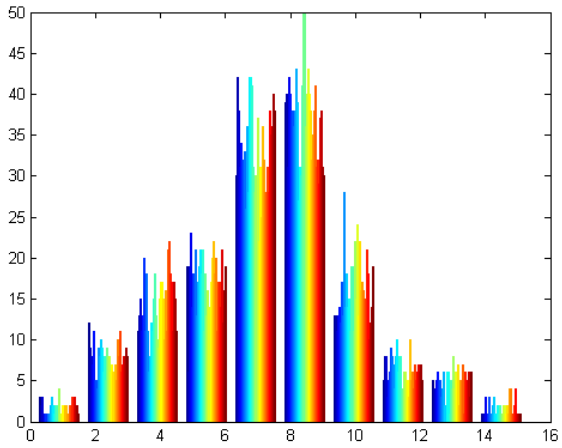


Fig. 4. Biomarkers distribution at healthy tissues after Round-up

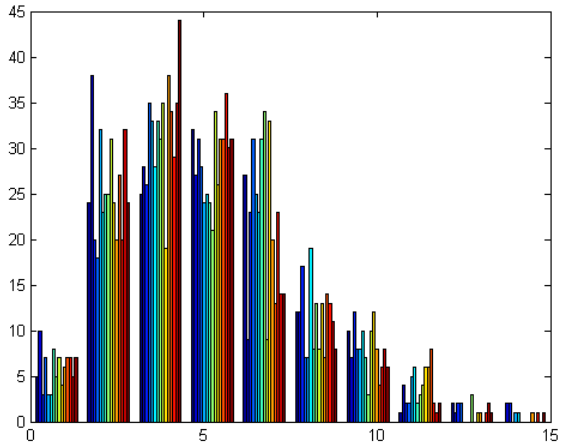


Fig. 5. Biomarkers distribution of selected tissues at carcinogenic case

ally tested our trained classificator with new 239 carcinogenic patients, as well as, with new 12 healthy patients. In order to avoid any negative values, the new patients are also processed with the Round-up threshold method.

The prior probabilities we used are those we defined earlier as the two test cases in III-C. The results presented in Table II are derived according to all three equations specified in III-C.

TABLE II
BAYESIAN A POSTERIORI CLASSIFICATION

Bayesian classification	Test Case 1	Test Case 2
32 carcinogenic tissues	100%	93.75%
32 healthy tissues	84.38%	100%
239 carcinogenic patients	97.5%	89.5%
12 healthy patients	91 %	100%

We evaluated the classificator performance through relative trade-off between the true positives and the false positives.

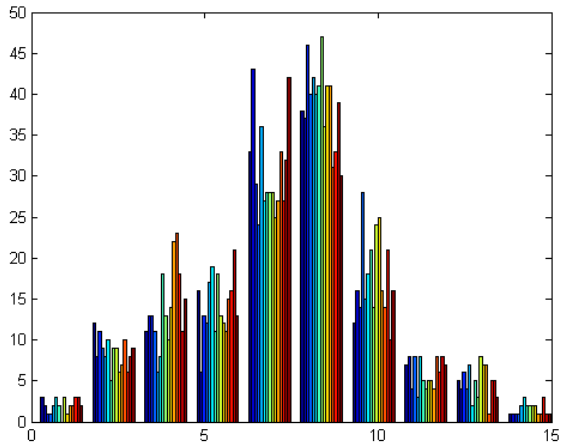


Fig. 6. Biomarkers distribution of selected tissues at healthy case

True positive rate (TPR), known as recall or sensitivity, is a term that refers to the ability of the classificator to correctly classify carcinogenic tissues. The ability of the classificator to correctly classify healthy tissues is measured as specificity. The results given in Table III show the ability of the classificator to correctly classify carcinogenic and healthy conditions with high accuracy at both tissues and patients which we used as a test sets.

TABLE III
SENSITIVITY AND SPECIFICITY

Performance	Sensitivity	Specificity	Test Cases
Tissues	1	0.84	Test case 1
	0.94	1	Test case 2
Patients	0.98	0.91	Test case 1
	0.90	1	Test case 2

V. CONCLUSION

In this paper we showed how to construct an efficient Bayesian classifier that distinguishes healthy from carcinogenic tissue when gene expression levels from Affymetrix DNA chip are used.

In order to make the best model of the prior distributions, which is essential for proper Bayesian classification, we developed an original methodology presented in III-A that preprocess the input data.

We focused on the genes whose expression clearly differs in the carcinogenic opposed to the healthy tissue. As we discovered a set of biomarkers, in the next few steps we provided series of statistical analysis in order to produce a well shaped distribution for both classes of tissues. Once we applied our generative approach, we classified the tissues calculating the Bayesian a posteriori probability.

This research follows two previous papers where we examined colorectal cancer using gene expression values obtained

from Illumina microarray BeadChip [3] and performed platform comparison between Illumina and Affymetrix [6]. Comparing the two platforms we realized that gene expressions obtained from Affymetrix require different statistical approach.

The obtained results in this paper, confirmed that the generative Bayesian approach gives excellent performance when appropriate preprocessing methodology is used. In our future research we will focus on distinguishing the different colorectal cancer stages in the carcinogenic tissues.

REFERENCES

[1] GLOBOCAN, 2008. [Online]. Available: http://globocan.iarc.fr/factsheets/cancers/colorectal.asp

[2] K. Jain, "Applications of biochips: From diagnostics to personalized medicine." *Current opinion in drug discovery & development*, vol. 7, no. 3, pp. 285–289, 2004.

[3] M. Simjanoska, A. M. Bogdanova, and Z. Popeska, "Recognition of colorectal carcinogenic tissue with gene expression analysis using bayesian probability," in *to be published in ICT Innovations 2012*. Springer Berlin / Heidelberg, 2012.

[4] M. Barnes, J. Freudenberg, S. Thompson, B. Aronow, and P. Pavlidis, "Experimental comparison and cross-validation of the affymetrix and illumina gene expression analysis platforms," *Nucleic acids research*, vol. 33, no. 18, pp. 5914–5923, 2005.

[5] L. Shi, L. Reid, W. Jones, R. Shippy, J. Warrington, S. Baker, P. Collins, F. de Longueville, E. Kawasaki, K. Lee *et al.*, "The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements," *Nature biotechnology*, vol. 24, no. 9, pp. 1151–1161, 2006.

[6] A. M. Bogdanova, M. Simjanoska, and Z. Popeska, "Classification of colorectal carcinogenic tissue with different dna chip technologies," Tech. Rep., 2013.

[7] W. Wong, M. Loh, and F. Eisenhaber, "On the necessity of different statistical treatment for illumina beadchip and affymetrix genechip data and its significance for biological interpretation," *Biology direct*, vol. 3, no. 1, p. 23, 2008.

[8] G. Marra, J. Sabates, and H. Rehrauer, "Transcriptome profile of human colorectal adenomas," 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8671

[9] Y. Hong, "Expression data from healthy controls and early stage crc patient's tumor," 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9348

[10] T. Orntoft and C. Andersen, "Expression data from primary colorectal cancers," 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13294

[11] T. Watanabe, T. Kobunai, E. Toda, Y. Okayama, Y. Sugimoto, and T. Oka, "Gene expression signature of colorectal cancer with microsatellite instability," 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4554

[12] J. Önskog, E. Freyhult, M. Landfors, P. Rydén, and T. Hvidsten, "Classification of microarrays; synergistic effects between normalization, gene selection and machine learning," *BMC bioinformatics*, vol. 12, no. 1, p. 390, 2011.

[13] M. Blangiardo and S. Richardson, "A bayesian calibration model for combining different pre-processing methods in affymetrix chips," *BMC bioinformatics*, vol. 9, no. 1, p. 512, 2008.

[14] S. Zakharkin, K. Kim, T. Mehta, L. Chen, S. Barnes, K. Scheirer, R. Parrish, D. Allison, and G. Page, "Sources of variation in affymetrix microarray experiments," *BMC bioinformatics*, vol. 6, no. 1, p. 214, 2005.

[15] R. Kitchen, V. Sabine, A. Simen, J. Dixon, J. Bartlett, and A. Sims, "Relative impact of key sources of systematic noise in affymetrix and illumina gene-expression microarray experiments," *BMC genomics*, vol. 12, no. 1, p. 589, 2011.

[16] Y. Hong, T. Downey, K. Eu, P. Koh, and P. Cheah, "A metastasis-pronesignature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics," *Clinical and Experimental Metastasis*, vol. 27, no. 2, pp. 83–90, 2010.

[17] J. Sabates-Bellver, L. Van der Flier, M. de Palo, E. Cattaneo, C. Maake, H. Rehrauer, E. Laczko, M. Kurowski, J. Bujnicki, M. Menigatti *et al.*, "Transcriptome profile of human colorectal adenomas," *Molecular Cancer Research*, vol. 5, no. 12, pp. 1263–1275, 2007.

[18] T. Watanabe, T. Kobunai, E. Toda, Y. Yamamoto, T. Kanazawa, Y. Kazama, J. Tanaka, T. Tanaka, T. Konishi, Y. Okayama *et al.*, "Distal colorectal cancers with microsatellite instability (msi) display distinct gene expression profiles that are different from proximal msi cancers," *Cancer research*, vol. 66, no. 20, pp. 9804–9808, 2006.

[19] R. Jorissen, L. Lipton, P. Gibbs, M. Chapman, J. Desai, I. Jones, T. Yeatman, P. East, I. Tomlinson, H. Verspaget *et al.*, "Dna copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers," *Clinical Cancer Research*, vol. 14, no. 24, pp. 8061–8069, 2008.

[20] Z. Wu and M. Aryee, "Subset quantile normalization using negative control features," *Journal of Computational Biology*, vol. 17, no. 10, pp. 1385–1395, 2010.

[21] I. Kohane, A. Butte, and A. Kho, *Microarrays for an integrative genomics*. MIT press, 2002.

[22] A. Butte and I. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pac Symp Biocomput*, vol. 5, 2000, pp. 418–429.

[23] C. Needham, I. Manfield, A. Bulpitt, P. Gilmartin, and D. Westhead, "From gene expression to gene regulatory networks in arabidopsis thaliana," *BMC systems biology*, vol. 3, no. 1, p. 85, 2009.

[24] Y. Hui, T. Kang, L. Xie, and L. Yuan-Yuan, "Digout: Viewing differential expression genes as outliers," *Journal of Bioinformatics and Computational Biology*, vol. 8, no. supp01, pp. 161–175, 2010.

[25] J. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences*, vol. 100, no. 16, pp. 9440–9445, 2003.

[26] K. HC and S. AA, "Gene expression profile analysis by dna microarrays: Promise and pitfalls," *JAMA*, vol. 286, no. 18, pp. 2280–2288, 2001.

[27] M. B. M., "An introduction to microarray data analysis." [Online]. Available: http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/chapter-final.pdf

[28] M. Weir and M. Rice, "Microarray clustering analysis." [Online]. Available: https://wesfiles.wesleyan.edu/courses/biol265/microarray_lab.htm

[29] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences*, vol. 96, no. 6, pp. 2907–2912, 1999.

Feature extraction from electroencephalographic records using EEGFrame framework

Alan Jović*, Lea Suć*, and Nikola Bogunović*
* Faculty of Electrical Engineering and Computing, University of Zagreb / Department of Electronics, Microelectronics, Computer and Intelligent Systems, Unska 3, 10 000 Zagreb, Croatia
{alan.jovic, lea.suc, nikola.bogunovic}@fer.hr

Abstract - Analysis of electroencephalographic (EEG) signals usually includes visual inspection of the signal, feature extraction, and model generation. Computer-aided nonlinear feature extraction from EEG in particular has already led to improved descriptive and prognostic models of brain states and disorders. However, in this field, there is a lack of freely available powerful tools for scientific exploration of EEG that would help researchers to compare the results of their work with others. Especially, because of the great diversity of the proposed methods for EEG analysis, there exists a need for a joint framework for inspection, extraction and visualization performed on the EEG records. The aim of this paper is to introduce such a framework, called EEGFrame, with its implementation in Java. The framework currently supports the analysis of standard EDF records via signal inspection, feature extraction, and feature vectors storage for knowledge discovery. EEGFrame is the result of refactoring and extension of the HRVFrame framework for heart rate variability analysis, with added methods for EEG analysis. This paper describes the properties and capabilities of the framework and discusses its relevance with respect to similar work. The main advantage of EEGFrame is its support for numerous linear and nonlinear methods described in literature.

I. INTRODUCTION

Computer models of brain functioning have become increasingly complex in recent years. The extraction of nonlinear features of neurological time-series, in particular, electroencephalography (EEG) and magnetoencephalography (MEG) enables ever more accurate analysis of diverse brain disorders and states. The analysis of EEG/MEG using nonlinear features aids medical personnel in better understanding of the underlying physiological aspects of the disorder as well as for classification of patients and prediction of disorder onset [1]. Typical applications include modeling of disorders such as epilepsy, coma, schizophrenia, Alzheimer's disease, Parkinson's disease, etc., while brain functioning during normal or altered conscious states include normal wake state, sleep, relaxation and meditation [1,2]. Complexity of the time-series is particularly pronounced under epileptic seizures, and this area of research usually involves creating accurate models for early detection of possible seizure [2].

Most of the features used for EEG/MEG analysis are nonlinear due to the inherent nonlinear and non-stationary behavior of the time-series. There are a few exceptions, mostly covering basic statistical properties of the series. Due to a large number of methods available for studying variability of biomedical time-series [3], computer-aided analysis of EEG/MEG relies on availability of tools for feature extraction and data mining. While there already exist several tools for such analyses, there are very few open-source frameworks that allow researchers to experiment with novel models of disorders. Currently, there are a few initiatives that try to standardize data inputs and outputs. Most of them are based on Matlab and/or C/C++ [4,5]. A framework in Python (PyEEG) was recently developed that tries to bridge the gap between EEG data input and data mining [6]. To our knowledge, there is no extensive stand-alone open-source framework that would cover the majority of features employed in EEG analysis, while at the same time enabling data input, feature extraction, EEG visualization, and storing feature vectors in a format suitable for data mining.

The aim of this paper is to introduce and describe such a framework, which we call EEGFrame. The framework is written in Java as a stand-alone application, but with embedding possibility, and is intended for use by both medical practitioners as well as researchers in the field of biomedical engineering. EEGFrame has been developed as an offspring of an already existing framework, HRVFrame that is used for heart rate variability analysis [7]. The extension includes visualization of the EEG signals loaded from a standard EDF (European Data Format) record, extraction of a number of nonlinear features applicable to EEG, and adaptation of frequency domain features to cover typical EEG frequency bands.

The paper is organized as follows. In Section 2, we show an overview of the framework, with description of input and output data formats. Section 3 describes the features that are implemented in the framework. Section 4 describes the visualization of the EEG records. Comparison to similar work and discussion are shown in section 5. Conclusion is given in section 6.

II. OVERVIEW OF THE FRAMEWORK

A. System overview, input and output

The main purpose of the EEGFrame framework is to extract feature vectors from EEG records and store them for further knowledge discovery. The framework currently

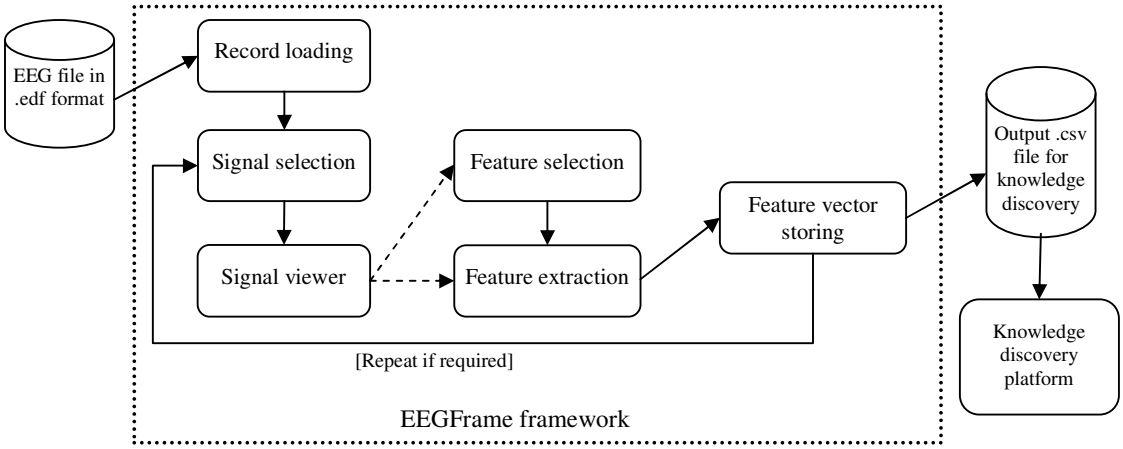


Fig. 1. Overview of the EEGFrame framework

supports extraction from standard EDF records with plans to support EDF+ standard in the future. Additionally, the framework enables visual inspection of each EEG record with numerous display options to suit the researcher's needs. Transformation of signal data from .edf file to a file in textual format is also possible through the framework. The overview of the system is shown in Fig. 1.

The framework assumes that the data in EDF format is available, and that it has been filtered (the framework does not include EEG preprocessing methods). It then enables visual inspection of a single record and feature extraction from the record, by specifying the extraction parameters. The output file is recorded in .csv format that can be read by most of the open-source knowledge discovery platforms such as Weka [8] or Rapidminer [9]. The output file contains feature vectors that contain values for all the features that were selected for the analysis. Each new feature vector is appended as a row to the end of the file if the analysis involves multiple segments. The user can always create a new output file. The header of the output file contains a list of all the possible features implemented in the framework. For those features that are not extracted as a part of the feature vector, the symbol "?" is added.

B. *Framework's internal structure*

EEGFrame consists of several Java packages, Fig. 2. Basically, there are two types of packages. There are the packages that contain all of the classes intended for data transformation and feature extraction. These packages start with the label "features". There are also other auxiliary packages that consist of classes used for constructing graphical user interface and record visualization, data input, data output, and testing of the framework. The package "statisticMeasure" is essential for functioning of some of the feature extraction classes. Likewise, the package "features.linear.frequencyDomain. operations" contains several classes needed for frequency domain analysis of EEG records (e.g. complex numbers, fast Fourier transform).

The packages consist of public classes with well-documented and thoroughly tested methods. All of the implemented classes with feature extraction methods provide reference about the source document(s) that describe the mathematical background and rationale for the implemented methods.

III. FEATURE EXTRACTION METHODS

The framework currently supports extraction of roughly 50 time-domain, frequency domain, time-frequency and nonlinear features. In Table I, a summary report of the implemented methods is provided. The reference next to the name of the method refers to the literature from which the method was implemented. If a method allows extraction of several features, a list of the features is also included in Table I. If a feature is parametric, the parameters that need to be provided are shown. Some of the implemented nonlinear methods are used for measuring a common or combined property of two or more time-series (e.g. mutual dimension, synchronization likelihood). We have included these methods in a separate package, "features.nonlinear.multiSeries", as to distinguish them from the other nonlinear methods that can be applied to only a single time-series (i.e. single EEG electrode signal).

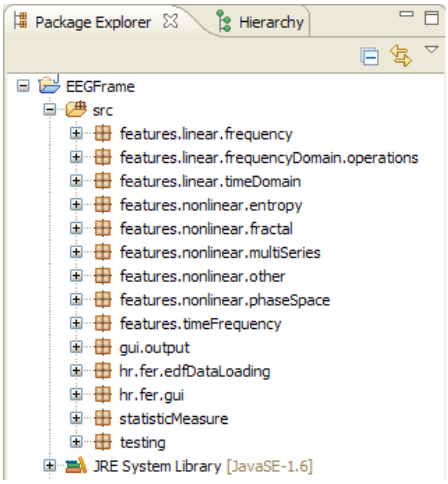


Fig. 2. The structure of the framework

TABLE I. METHODS AND FEATURES IMPLEMENTED IN THE FRAMEWORK

Method [reference]	List of features / parameters	Package	Class
Mean [10]	Single feature	features.linear.timeDomain	Mean
Standard deviation [10]	Single feature	features.linear.timeDomain	StandardDeviation
Fano factor [11]	Single feature	features.linear.timeDomain	FanoFactor
Autocorrelation coefficient [11]	Single feature	features.linear.timeDomain	Autocorrelation-Coefficient
Mean of absolute values of first differences [10]	Single feature	features.linear.timeDomain	MeanOfAbsoluteValues OfFirstDiff
Mean of absolute values of second differences [10]	Single feature	features.linear.timeDomain	MeanOfAbsoluteValues OfSecondDiff
Mean of absolute values of first differences normalized [10]	Single feature	features.linear.timeDomain	MeanOfAbsoluteValues OfFirstDiffNormalized
Mean of absolute values of second differences normalized [10]	Single feature	features.linear.timeDomain	MeanOfAbsoluteValues OfSecondDiffNormalize d
Spectral analysis [10]	AlfaPSD, BetaPSD, GamaPSD, DeltaPSD, ThetaPSD (5 feat.) / sample frequency, FFT PSD (window) or Burg PSD estimate (AR model order)	features.linear.frequency	SpectralAnalysis
Spectral entropy [12]	Single feature / lower and upper frequency limit, FFT PSD or Burg PSD estimate	features.linear.frequency	SpectralEntropy
Approximate entropy [13]	Single feature / m factor, r	features.nonlinear.entropy	ApEn
Maximum approximate entropy [13]	MaxApEn, r for MaxApEn / *m* factor	features.nonlinear.entropy	ApEn
Carnap entropy 1D [14]	Single feature	features.nonlinear.entropy	CarnapEntropy1D
Corrected conditional Shannon entropy [15]	Single feature / dimension, no. of bins	features.nonlinear.entropy	CorrectedConditional-SnannonEntropy
Rényi entropy [16]	Single feature / order	features.nonlinear.entropy	RenyiEntropy
Sample entropy [17]	Single feature / m factor, r	features.nonlinear.entropy	SampEn
Maximum sample entropy [17]	MaxSampEn, r for MaxSampEn / *m* factor	features.nonlinear.entropy	SampEn
Detrended fluctuation analysis [18]	DFAAlphaS, DFAAlphaL (2 feat.) / minimum analyzed segment length, bound for AlphaL, long range calculation flag	features.nonlinear.fractal	DFA
Higuchi's fractal dimension [19]	Single feature / *kmax*	features.nonlinear.fractal	HiguchiDimension
Hurst exponent [11]	Single feature	features.nonlinear.fractal	HurstExponent
Cross recurrence [20]	CRPRecurrence rate, CRPLmean, CRPDET, CRPSh. ent. rec., laminarity (5 feat.) / probe, dimension,lag, r	features.nonlinear. multiSeries	CrossRecurrence
Mutual dimension [21]	Single feature / dimension 1, dimension 2, lag 1, lag 2, no. of bins	features.nonlinear. multiSeries	MutualDimension
Synchronization likelihood [22]	Single feature / signal indices, dimension, lag, no. of bins, recurrence number, ro	features.nonlinear. multiSeries	SynchronizationLikeliho od
Allan factor [11]	Single feature / observational window	features.nonlinear.other	AllanFactor
Lempel-Ziv complexity [23]	Single feature	features.nonlinear.other	LempelZivComplexity
Nonlinear forecasting [24]	Single feature / dimension, lag	features.nonliner.other	NonlinearForecasting
Correlation dimension [25]	Single feature / dimension, lag, no. of bins	features.nonlinear. phaseSpace	CorrelationDimension
Central tendency measure [26]	Single feature / dimension, lag, r	features.nonlinear. phaseSpace	CTM
Largest Lyapunov exponent [27]	Single feature / dimension, trajectory length	features.nonlinear. phaseSpace	LyapunovExponent
Recurrence plot [28]	AVG number of neighbors, recurrence rate, Lmean, DET, Sh. ent. rec., laminarity (6 feat.) / dimension, lag, r	features.nonlinear. phaseSpace	RecurrencePlot
Spatial filling index [29]	Single feature / dimension, lag, no. of bins	features.nonlinear. phaseSpace	SpatialFillingIndex
Standard deviations ratio [30]	Single feature	features.nonlinear. phaseSpace	StandardDeviationsRatio
Haar wavelet standard deviation [11]	Single feature / scale	features.timeFrequency	HaarWaveletStandard Deviation
HilbertHuangTransform [31]	Instantaneous frequencies (IF), amplitudes, intrinsic mode functions, max IF, amplitudes for max IF (multiple features) / sampling period	features.timeFrequency	HilbertHuangTransform

IV. EEG SIGNAL VISUALIZATION

It is important for researchers to have the possibility of visual inspection of EEG records. The standard EDF format supports the International 10-20 system of electrode placements [32]. In this EEG system, it is possible to have 21 or more electrodes that measure the voltages of the brain's micro-currents. The visualization of the signals recorded by such a system can be performed in many ways. The standard display is a rectangular grid with each rectangle covering exactly 0.2 s times 50 μV. Since the number of signals can vary, and not all of them are always necessary to include for certain disorders, the visualization part includes the options to display only specific electrode signals as well as all signals. Additionally, several time and amplitude scales can be selected, and the user can easily browse through the entire record. An example of EEG visualization is shown in Fig. 3.

The graphical user interface also allows users to display the properties of the records (e.g. date taken, start time, length in seconds, number of signals, header length in bytes), that are available from the .edf file header.

V. COMPARISON TO SIMILAR WORK AND DISCUSSION

In Table II, we compare the currently most popular open-source software in EEG analysis community with EEGFrame. There are several open-source EEG analysis software tools available today. Most of the software is based on Matlab/Octave or C/C++, and only a few of them are specialized for EEG analysis. One of the problems with Matlab based software is that its use depends on whether the user has a Matlab license. This can be solved to an extent by using Octave instead of Matlab (e.g. project BioSig [5]). The other problem is that the software is not monolithic or cohesive – i.e. not all of the required methods can be easily found and used. This severely limits the usefulness and user-friendliness of the large Matlab-based software toolboxes such as EEGLab [4]. The main advantage is the large community support and extensions that Matlab provides. This includes signal analysis and transformations, feature extraction, classification, as well as visualization capabilities.

While no software or framework today contains all of the methods that the researchers used in EEG analysis, it can be safely concluded that two of the frameworks: PyEEG, written in Python, and EEGFrame, written in Java, provide the largest number of individual feature extraction methods. Both of the frameworks focus on nonlinear features. The advantage of EEGFrame, when compared with PyEEG lies both in the larger number of currently implemented individual features, as well as in visualization and data input/output capability. PyEEG is a framework that contains only the feature extraction methods; it relies heavily on other software (e.g. BioSig) for particular application [6].

The principle goal of the researchers in EEG analysis is to acquire the best model for a particular brain state or disorder. In order to claim that their model is the best possible one within the limits of current human knowledge, the researchers need to be able to efficiently compare their results, particularly, their features or features combinations [33], with the work of others. EEGFrame facilitates this comparison because it provides the researcher with a large number of features' implementations that other researchers applied in EEG

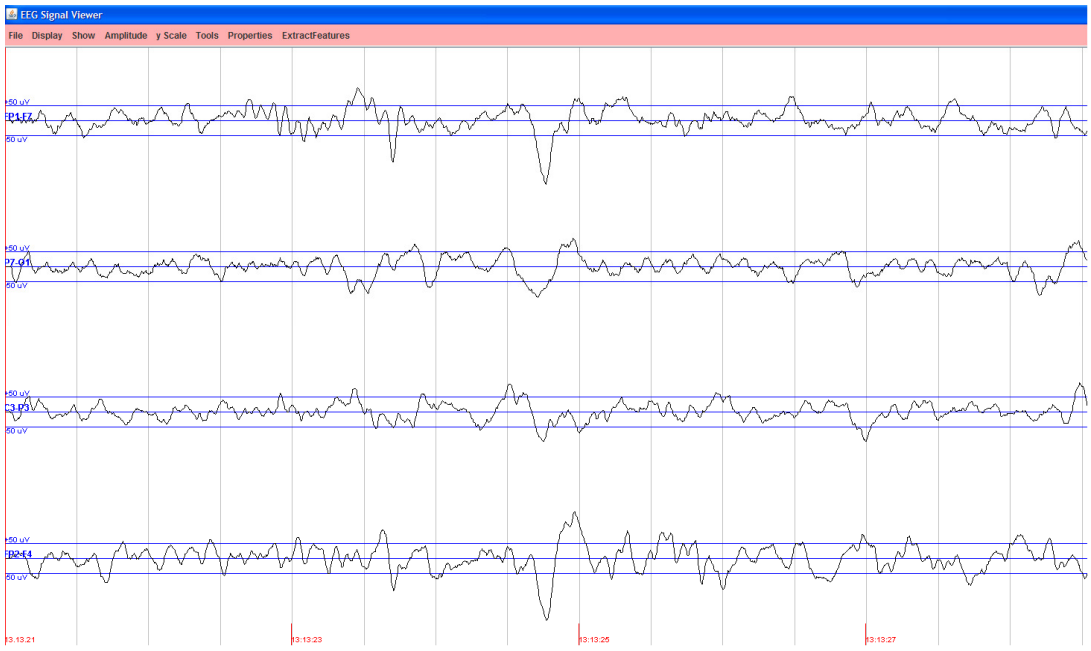


Fig. 3. Visualization of the EEG record chb01_27.edf from Physionet CHB-MIT database for electrodes FP1-F7, P7-O1, C3-P3, and FP2-F4. The user can select any number of signals present in the record for display

TABLE II. OPEN-SOURCE EEG ANALYSIS SOFTWARE COMPARISON

Software	Purpose	Implementation language	Type	Implemented features
EEGLab [4]	Extensive Matlab toolbox for EEG analysis: visualization, 3D brain modelling, feature extraction, several plugins (NFT, ERICA, BCILAB...)	Matlab	Embedded	Time/frequency/time-frequency/independent component transformations and features / unknown total number of features
BioSig [5]	Reading and writing routines for many biomedical time-series data formats; EEG preprocessing, visualization, feature extraction (multivariate autoregressive modeling) and classification (via Matlab/Octave)	C/C++, Matlab (or Octave)	Some functions standalone, mostly embedded	Time/frequency/time-frequency transformations and features, unknown total number of features
PyEEG [6]	Feature extraction framework, feature vector output for data mining	Python	Embedded	Frequency/nonlinear features, currently 21 features in total
EEGFrame	Signal inspection, feature extraction framework, handles .EDF input, feature vector output for data mining	Java	Stand-alone or embedded	Time/frequency/time-frequency/nonlinear features, currently 49 features in total

analysis. Hence, with the assumption of using the same dataset, a researcher can claim that his feature or features' combination is superior to the features employed by others for a particular problem.

Batch feature extraction from multiple EEG records is currently not possible through the EEGFrame graphical user interface. However, adapting the framework for batch extraction from several records is not difficult and is planned for future versions. Adding more methods to EEGFrame, especially other nonlinear ones (e.g. phase synchronization, Hjorth mobility and complexity, SVD entropy, Fisher information, etc.) is also planned.

EEGFrame can be employed both as a stand-alone product, and as a part of a larger system. Herein, only the feature packages would be integrated with other software. The framework is written entirely in Java, which alleviates the problems with platform dependencies. Thus, EEGFrame can be used on any operating system, provided that the Java virtual machine (version 5.0+) is installed on the system. EEGFrame is available as open-source, GPL 2 licensed software for non-commercial biomedical time-series analysis applications from the following web site: http://www.zemris.fer.hr/~ajovic/eegframe/eegframe.html.

VI. CONCLUSION

This paper presented a novel feature extraction framework from EEG that is implemented in Java. The framework currently implements many features used by researchers in the domain of EEG analysis and thus enables easier comparison of academic work. It can be used both as a part of another, larger system, or as a stand-alone EEG visualization and analysis application.

For future work, the framework is planned to be updated with additional nonlinear features. Also, visualization of phase space features as well as batch analysis of EEG records is in order.

REFERENCES

[1] C. J. Stam, "Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field," Clin. Neurophysiol., vol. 116, pp. 2266–2301, Oct. 2005.

[2] K. Lehnertz, "Epilepsy and Nonlinear Dynamics," J. Biol. Phys., vol. 34, pp. 253–266, Aug. 2008.

[3] A. Bravi, A. Longtin, and A. J. E. Seely, "Review and classification of variability analysis techniques with clinical applications," BioMed. Eng. OnLine, vol. 10, p. 90, Oct. 2011.

[4] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics", J. Neurosci. Methods, vol. 134, pp. 9–21, Mar. 2004.

[5] C. Vidaurre, T. H. Sander, and A. Schlögl, "BioSig: The Free and Open Source Software Library for Biomedical Signal Processing," Comput. Intell. Neurosci., p. 935364, 2011.

[6] F. S. Bao, X. Liu, and C. Zhang, "PyEEG: An open source Python module for EEG/MEG feature extraction," Comput. Intell. Neurosci., p. 406391, 2011.

[7] A. Jovic and N. Bogunovic, "HRVFrame: Java-Based Framework for Feature Extraction from Cardiac Rhythm," Lecture Notes in Artificial Intelligence, vol. 6747, pp. 96–100, Jul. 2011.

[8] I. H. Witten and E. Frank, Data mining: Practical machine learning tools and techniques, 3rd ed., San Francisco: Morgan Kaufmann, 2011.

[9] (2012) RapidMiner. [Online]. Available: http://rapid-i.com/content/view/181/190/

[10] X.-W. Wang, D. Nie, and B.-L. Lu, "EEG-Based Emotion Recognition Using Frequency Domain Features and Support Vector Machines," Lecture Notes in Computer Science, vol. 7062, pp. 734–743, 2011.

[11] M. C. Teich, S. B. Lowen, B. M. Jost, K. Vibe-Rheymer, and C. Heneghan, "Heart-Rate Variability: Measures and Models," in Dynamic Analysis and Modeling, ser. Nonlinear Biomedical Signal Processing, M. Akay, Ed. New York: IEEE Press, 2001, vol. II, ch. 6, pp. 159–213.

[12] I. A. Rezek and S. J. Roberts, "Stochastic Complexity Measures for Physiological Signal Analysis," IEEE Trans. Biomed. Eng., vol. 45, no. 9, pp. 1186–1191, Sep. 1998.

[13] S. M. Pincus and A. L. Goldberger, "Physiological time-series analysis: what does regularity quantify?" Am. J. Physiol., vol. 266, no. 4, (Heart Circ. Physiol., vol. 35), pp. H1643–H1656, Apr. 1994.

[14] F. Jovic, D. Krmpotic, and A. Jovic, "Process entropy and informational macrodynamics in a ceramic tile plant," in Proc. 32nd Croatian Society for Information and Communication Technology, Electronics and Microelectronics - MIPRO, vol. III, pp. 50–53, 2009.

[15] A. Porta, G. Baselli, D. Liberati, N. Montano, C. Cogliati, T. Gnecchi-Ruscone, et al., "Measuring regularity by means of a corrected conditional entropy in sympathetic outflow," Biol. Cybern., vol. 78, no. 1, pp. 71–78, 1998.

[16] K. Waheed and F. M. Salam, "A Data-Derived Quadratic Independence Measure for Adaptive Blind Source Recovery in Practical Applications," in Proc. 45th IEEE Int. Midwest Symposium on Circuits and Systems, pp. 473–476, 2002.

[17] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," Am. J.

Physiol. (Heart Circ. Physiol.), vol. 278, no. 6, pp. 2039–2049, Jun 2000.

[18] C.-K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger, "Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series," Chaos, Solitons, & Fractals, vol. 5, no. 1, pp. 82–87, Jan. 1995.

[19] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory", Physica D, vol. 31, issue 2, pp. 277--283, Jun. 1988.

[20] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," Physics Reports, vol. 438, issues 5–6, pp. 237–329, Jan. 2007.

[21] C. J. Stam, T. C. van Woerkom, W. S. Pritchard, "Use of non-linear EEG measures to characterize EEG changes during mental activity", Electroencephalogr. Clin. Neurophysiol., vol. 99, no. 3, pp. 214–224, Sep. 1996.

[22] C. J. Stam, B. W. van Dijk, "Synchronization likelihood: an unbiased measure of generalized synchronization in multivariate data sets", Physica D: Nonlinear Phenomena, vol. 163, issues 3–4, pp. 236–251, Mar. 2002.

[23] X.-S. Zhang, Y.-S. Zhu, N. V. Thakor, and Z.-Z. Wang, "Detecting Ventricular tachycardia and fibrillation by complexity measure," IEEE Trans. Biomed. Eng., vol. 46, no. 5, pp. 548–555, May 1999.

[24] G. Sugihara and R. M. May, "Non-linear forecasting as a way of distinguishing chaos from measurement error in time series," Nature, vol. 344, pp. 734–740, 1990.

[25] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," Physica D: Nonlinear Phenomena, vol. 9, no. 1–2, pp. 189–208, Oct. 1983.

[26] M. E. Cohen, D. L. Hudson, and P. C. Deedwania, "Applying continuous chaotic modeling to cardiac signal analysis," IEEE Eng. Med. Biol. Mag., vol. 15, no. 5, pp. 97–102, Sep./Oct. 1996.

[27] M. Rosenstien, J. J. Colins, and C. J. de Luca, "A practical method for calculating largest Lyapunov exponents from small data sets," Physica D: Nonlinear Phenomena, vol. 65, no. 1–2, pp. 117–134, May 1993.

[28] J. P. Zbilut, N. Thomasson, and C. L. Webber, "Recurrence quantification analysis as a tool for nonlinear exploration of nonstationary cardiac signals," Med. Eng. & Phys., vol. 24, no. 1, pp. 53–60, Jan. 2002.

[29] O. Faust, R. U. Acharya, S. M. Krishnan, and L. C. Min, "Analysis of cardiac signals using spatial filling index and time-frequency domain," BioMed. Eng. OnLine, vol. 3, p. 30, Sep. 2004.

[30] C.-W. Lin, J.-S. Wang, and P.-C. Chung, "Mining Physiological Conditions from Heart Rate Variability Analysis," IEEE Computat. Intell. Mag., vol. 5, no. 1, pp. 50–58, Feb. 2010.

[31] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, et al., "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis," in Proc. of the Royal Society A, vol. 454, no. 1971, pp. 903–995, Mar. 1998.

[32] B. Kemp, A. Värri, A. C. Rosa, K. D. Nielsen, and J. Gade, "A simple format for exchange of digitized polygraphic recordings," Electroencephalogr. Clin. Neurophysiol., vol. 82, no. 5, pp. 391–393, May 1992.

[33] S.-F. Liang, H.-C. Wang, and W.-L. Chang, "Combination of EEG Complexity and Spectral Analysis for Epilepsy Diagnosis and Seizure Detection", EURASIP Journal on Advances in Signal Processing, p. 853434, 2010.

Technique for indirect analysis of domestic power consumers based on power pattern recognition for smart energy metering

N. Nikolaev*, Y. Rangelov*, V. Valchev** and A. Marinov**
* Technical University of Varna/Department "Electric Power Engineering", Varna, Bulgaria
** Technical University of Varna/Department "Electronics and Microelectronics", Varna, Bulgaria
n.nikolaev@tu-varna.bg; y.rangelov@tu-varna.bg; vencivalchev@hotmail.com; a.marinov@tu-varna.bg

Abstract - The current paper presents an approach for measuring energy behavior of domestic power consumers based on power pattern recognition. This approach can be implemented into a central smart energy meter that will be able to identify the energy consumed by each particular power appliance. The suggested approach is based on recognizing the transient power change that occurs during the turning 'on' and 'off' electric devices. This transient power change is then compared against pre-set patterns, where the different devices are recognized using fuzzy logic. The approach is tested in simulation and proves functional for most electric appliances. Implementing it together with a smart energy meter forms an intelligent system that allows users and operators to estimate the contribution of each electric device to the total energy consumption. The advantage of so formed system is that only a single central measuring is required, omitting the necessity of installing a meter to each device.

I. INTRODUCTION

Over the last decade, increasing attention has been paid to energy efficiency, not only in industry but also in households. The fast growing economies demand higher amounts of energy. Still fossil fuels are dominant source in the electric power production. However the EU has set frames to ensure the displacement of these non eco fuels by technologies based on renewable energy sources.

The roadmap for the energy sector development forecasts that big share of the energy needed to supply the households will be produced by their own photovoltaic, wind or other renewable energy installations. Along with the transition to local (distributed) power generation, the energy consumption efficiency of the households has to be increased. A fundamental problem for the domestic consumers is to identify the energy consumption of each electric appliance. The availability of this information could allow the users to improve the efficiency. The recent years' research is focused on the development of a central measurement device, which is able to recognize the devices plugged in the mains electricity.

Reference [1] presents an algorithm for indirect recognition of the devices plugged into the domestic electrical installation. The algorithm is based on the Matrix Pencil Method. A comparison of three controlling methods – k Nearest Neighbor, Naive Bayes and Support Vector Machine is presented.

Bulgarian scientific fund

Similar algorithm able to adapt itself to devices for which the system has not been trained is proposed in [2]. In [3] an algorithm for indirect device recognition, based on neural networks and electrical current harmonics is presented. Other methods for nonintrusive domestic load monitoring are available in [4-6].

The current contribution presents a new algorithm for nonintrusive indirect recognition of the appliances powered by the domestic mains electricity. The algorithm detects deviations in the real power flowing through the main cable which supplies the household. Then based on particular power signatures and a fuzzy logic controller the switched appliance is determined. The presented simulation results prove its stability in recognition of most of the "passive" appliances which form the biggest part of the monthly electricity bill.

The paper continues with a brief description of the simulation circuit, which is subject to test of the algorithm. The proposed algorithm along with the simulation results are presented in section III. The conclusions are drawn in section IV.

II. METHODOLOGY

The simulated circuit is shown on Fig. 1. The voltage source supplying the house hold along with its internal impedance are designated as V_S и Z_S. The impedance reflects the voltage drop due to the loading of the electrical circuit.

The measurement acquisition system includes an instantaneous current and voltage recorders. They are connected next to the voltage source in order to measure

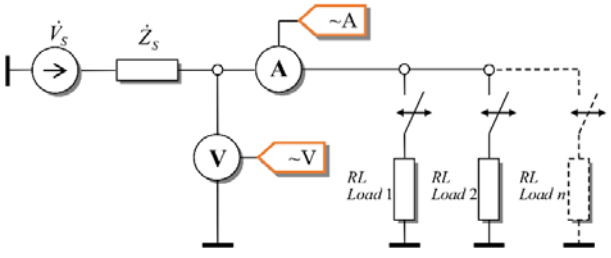


Figure 1. Simulated circuit

the total current of the household. The measured current and voltage instantaneous values are recorded and fed to the recognition algorithm.

During the simulation the appliances designated as RL Load are switched ‘on’ and ‘off’. Fig. 2 shows a simulated scenario with consecutively switched appliances with three different power ratings - 865 W, 414 W and 112 W respectively.

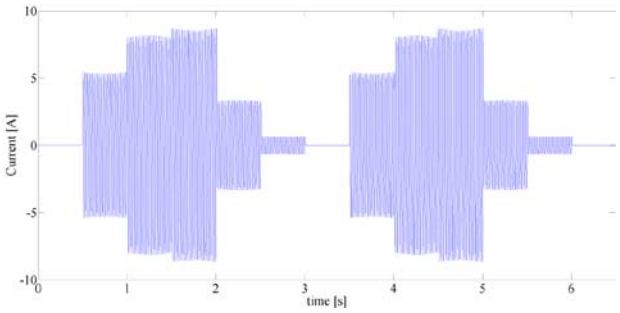


Figure 2. Instantaneous current simulating the switching of three appliances

III. Algorithm for Indirect Recognition of the Switched Domestic Allpliances

The structure diagram of the algorithm for indirect recognition of electrical appliances is depicted in Fig. 3. The required input data is obtained from the data acquisition system which measures the instantaneous current and voltage values. After a dedicated signal processing, the data (power signature) is fed to the device recognition block, which determines the switched appliance.

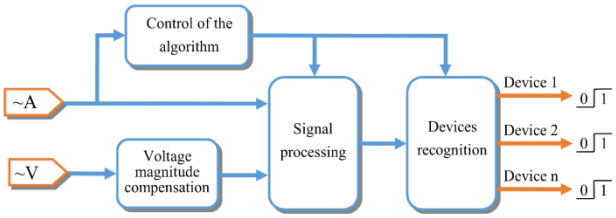


Figure 3. General structure diagram of the recognition algorithm

A. Control Block

The control block detects the events of switching the electrical appliances. Its structure diagram is shown in Fig. 4. The input signal is the measured instantaneous current. Then its mean value is calculated. During a normal operation (no switching of devices) the current mean value deviates around zero. However, a quickly decaying impulse appears at a moment of device switching. This impulse is differentiated and filtrated by the next sub-block. Then its absolute value is determined in order to make the control sensitive to both ‘on’ and ‘off’ events. When an event occurs the signal level increases above the threshold of the relay block. The raising pulse on the output of the relay triggers the next block, which creates a single rectangular output signal with a length of exactly 0.2 sec.

The interested reader can find more information about modelling of differentiating and filtering functions by Laplace transformation in [7], or in any other book dedicated to control theory.

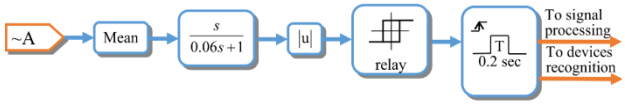


Figure 4. Structure diagram of the control block

B. Block for Voltage Magnitude Compensation

Fig. 5 shows the diagram of the voltage magnitude compensation block. Its purpose is to correct the measured voltage when it is different from the nominal. This is imposed by the fact that the heating appliances have a quadratic power-voltage dependence:

$$P = U^2 / R \qquad (1)$$

Therefore, to prevent the influence of the power deviation on the signal processing, the following compensation is done:

$$V_{corr} = \sim V \cdot \left(\frac{V_{nom}}{V_{rms}} \right)^2 \qquad (2)$$

In other words each instantaneous value of the voltage is multiplied by the squared ratio of the nominal and the measured rms values of the voltage. The corrected instantaneous voltage value is then fed to the signal processing block. This compensation method was originally proposed in [8].

The voltage waveform of the source used in the simulation is shown in Fig. 6. The algorithm is tested with a voltage magnitude deviation of 13 V and a 1 sec period. Moreover, 5 % of 3rd and 5th harmonics are added.

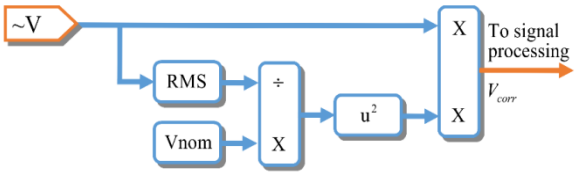


Figure 5. Voltage magnitude compensation block

C. Signal Processing Block

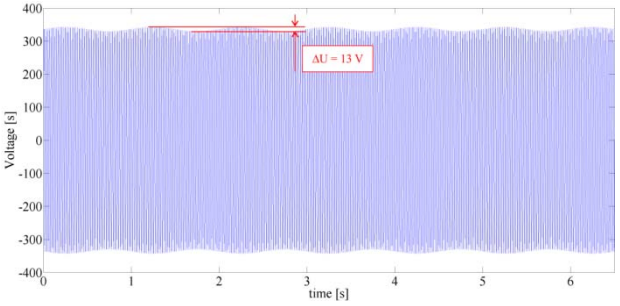


Figure 6. Waveform of the simulated voltage source

The structure diagram of the signal processing block is depicted in Fig. 7. The input signals are the instantaneous current and the corrected instantaneous voltage. The normalized instantaneous value of the power is obtained by the product of this two signals. After that, the mean value (the averaging period is $1/f_{Fund}$) gives the value of

the normalized real power. The normalized real power for the simulated circuit is shown in Fig. 8 (a).

In order to distinguish the moment of device switching and the steady-state of the circuit the active power is fed through the transfer function TF. This function differentiates and filtrates the input signal. The output of the TF is presented in Fig. 8 (b). This plot shows that when the circuit is in a steady-state the output signal is nearly zero. Further on, the filter smoothens the fast signal deviations originating from the differentiation.

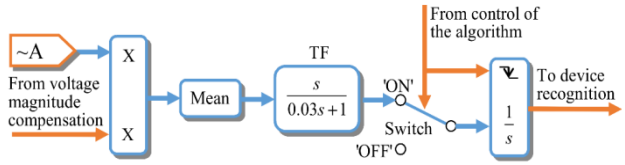


Figure 7. Signal Processing Block

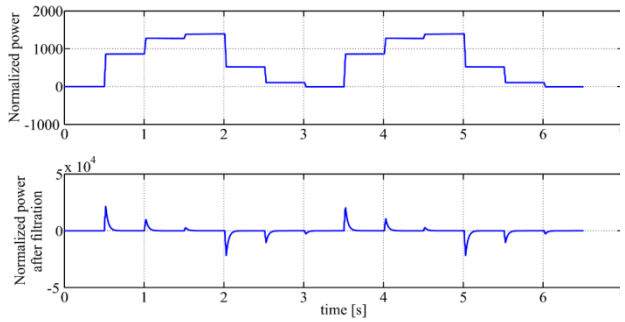


Figure 8. Normalized active power and (b) the signal obtained after its differentiation and filtration

When the rectangular 0.2 sec long impulse appears, the logical Switch turns into 'ON' position and feeds the output signal of the TF to the integrator. After 0.2 sec, when the impulse turns from logical true to false, the integrator is restarted and the logic Switch is turned back to 'OFF' position. As a result an impulse with particular peak value appears on the output of the integrator. This peak value is the power signature which is unique for each appliance. The integrator output signal from the simulation is presented in Fig. 9. This plot shows that despite the voltage magnitude deviation, the peak value for the same device remains unchanged. However, if there is no voltage magnitude compensation the power signature will be different every time.

D. Device Recognition Block

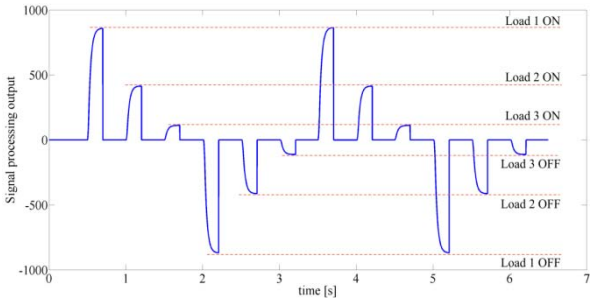


Figure 9. Output of the signal processing block (power signatures)

The device recognition block is depicted in Fig. 10. This block uses the power signature provided by the signal processing block. The signature value is fed in the *Sample/Hold* block which is synchronized by the control block to provide as an output only the peak value of the power signature. This is important in order to ensure correct operation of the *Fuzzy logic controller*. Otherwise, if all values of the signal processing output are fed into the *Fuzzy logic controller* the output will provide fake results.

Two input membership functions are defined for each device, respectively for 'on' and 'off' state. Fig. 11 shows the input membership functions for the presented simulation with three devices. The probability distribution functions are defined as triangular. The tip of the triangle defines the value most likely to be observed when a device is switched. The signal values spanned by the triangle are set as ±10 % from the tip value.

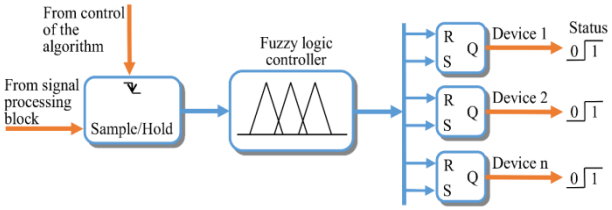


Figure 10. Device recognition block

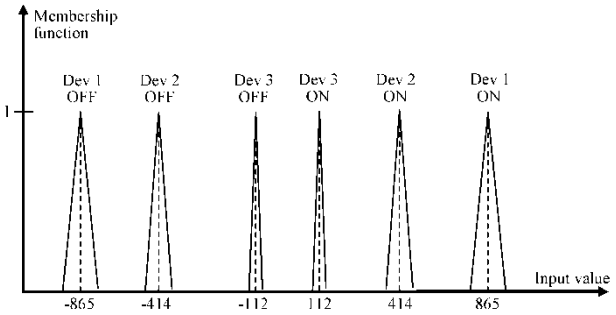


Figure 11. Input membership functions

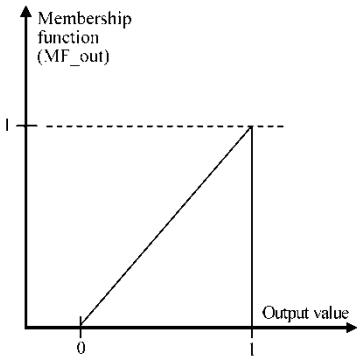


Figure 12. Outputs membership functions

A membership function is defined for each output. The *Fuzzy logic controller* implemented in the simulation has two outputs for each device – one output for turning the device 'on' and another one for turning the device 'off'. For all outputs the membership function is the same (see Fig. 12).

The rules defining the logical relation between the inputs and the outputs of the *Fuzzy logic controller* are the following:

- **If Input is** Dev_1_ON **then** (Dev_1_ON is MF_out) **and** (Dev_1_OFF is not MF_out) **and** (Dev_2_ON is not MF_out) **and** (Dev_2_OFF is not MF_out) **and** (Dev_3_ON is not MF_out) **and** (Dev_3_OFF is not MF_out)
- **If Input is** Dev_1_OFF **then** (Dev_1_ON is not MF_out) **and** (Dev 1 OFF is MF_out) **and** (Dev_2_ON is not MF_out) **and** (Dev_2_OFF is not MF_out) **and** (Dev_3_ON is not MF_out) **and** (Dev_3_OFF is not MF_out)
- **If Input is** Dev_2_ON **then** (Dev_1_ON is not MF_out) **and** (Dev_1_OFF is not MF_out) **and** (Dev_2_ON is MF_out) **and** (Dev_2_OFF is not MF_out) **and** (Dev_3_ON is not MF_out) **and** (Dev_3_OFF is not MF_out)
- **If Input is** Dev_2_OFF **then** (Dev_1_ON is not MF_out) **and** (Dev_1_OFF is not MF_out) **and** (Dev_2_ON is not MF_out) **and** (Dev_2_OFF is MF_out) **and** (Dev_3_ON is not MF_out) **and** (Dev_3_OFF is not MF_out)
- **If Input is** Dev_3_ON **then** (Dev_1_ON is not MF_out) **and** (Dev_1_OFF is not MF_out) **and** (Dev_2_ON is not MF_out) **and** (Dev_2_OFF is not MF_out) **and** (Dev_3_ON is MF_out) **and** (Dev_3_OFF is not MF_out)
- **If Input is** Dev_3_OFF **then** (Dev_1_ON is not MF_out) **and** (Dev_1_OFF is not MF_out) **and** (Dev_2_ON is not MF_out) **and** (Dev_2_OFF is not MF_out) **and** (Dev_3_ON is not MF_out) **and** (Dev_3_OFF is MF_out)

The settings of the *Fuzzy logic controller* are the following:

- Mamdani system:
- And method – min
- Or method – probor
- Implication – prod
- Aggregation – max
- Defuzzification – lom

Depending on the input value, the *Fuzzy logic controller* sets a logical true to only one of the outputs at a time. As already mentioned there are two outputs for each device which are connected to a RS flip-flop (one per device). The output of the flip-flop is logical true if the device is recognized as turned-on and is logical false if recognized as turned-off. The outputs for the states of the three devices from the simulation are shown in Fig. 13. These output values are then used to evaluate the individual consumption of the devices.

E. Individual Consumption Data Records

After the information for each device state ('on' or 'off', see Fig. 13) is available an information for their individual consumption is collected. For a given period of time the individual consumption is evaluated by the following integral function:

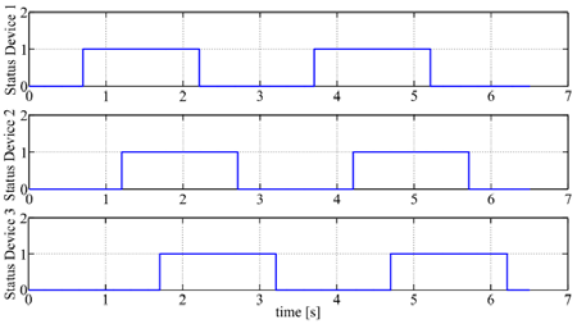


Figure 13. Output of the device recognition block

$$E_{device} = \int_{0}^{\text{end of period}} \left[state \cdot \left(\frac{V_{rms}}{V_{nom}} \right)^2 \cdot P_{nom} \right] \cdot dt \qquad (3)$$

As explained in section III.B. the power-voltage dependence is quadratic in general. Therefore the nominal power of the device is corrected with the squared ratio between the measured rms and the nominal voltage value.

IV. CONCLUSIONS

The developed algorithm for indirect recognition of the switched domestic appliances is stable despite of the disturbances of the power supply (voltage magnitude deviation and presence of harmonics). With its help most of the "passive" appliances, such as boilers, electrical heaters, cookers, lighting bulbs and others are identified. Based on that a database with their individual consumption is collected, providing the costumers with useful information for their energy expenses on each device.

ACKNOWLEDGMENT

This paper is prepared in the frames of Project MU03/164 - "Intelligent systems for energy management and control of consumer expenses", Ministry of

Education Youth and Science, Bulgarian National Science Fund.

REFERENCES

[1] K. Chahine, K. Drissi, C. Pasquier, K. Kerroum, C. Faure, T. Jouannet, and M. Michou, "Electric Load Disaggregation in Smart Metering Using a Novel Feature Extraction Method and Supervised Classification," *Energy Procedia*, vol. 6, pp. 627-632, 2011.

[2] M.-S. Tsai and Y.-H. Lin, "Modern development of an Adaptive Non-Intrusive Appliance Load Monitoring system in electricity energy conservation," *Applied Energy*, vol. 96, pp. 55–73, 2012.

[3] D. Srinivasan, W. Ng, and A. Liew, "Neural-network-based signature recognition for harmonic source identification," *Power Delivery, IEEE Transactions on*, vol. 21, pp. 398-405, 2006.

[4] G. Hoogsteen, J. Krist, V. Bakker, and G. Smit, "Non-Intrusive Appliance Recognition," in *3rd IEEE PES Innovative Smart Grid Technologies Europe*, Berlin, 2012.

[5] S. Drenker and A. Kader, "Nonintrusive monitoring of electric loads," *Computer Applications in Power, IEEE*, vol. 12, pp. 47-51, 1999.

[6] A. Lima, A. Silva, and D. Nascimento, "Noninvasive Monitoring of Residential Loads," in *3rd IEEE PES Innovative Smart Grid Technologies Europe*, Berlin, 2012, pp. 1-5.

[7] W. Levine, *The control handbook*, 2 ed.: CRC Press, 2010.

[8] G. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, pp. 1870-1891, 1992.

Testing of several overlapping optimization methods for bin-packing problem

T. Rolich*, D. Domović*, D. Grundler*
* University of Zagreb, Faculty of Textile Technology,
Department of Basic Natural and Technical Sciences, Zagreb, Croatia
{tomislav.rolich, daniel.domovic, darko.grundler}@ttf.hr

Abstract - In this paper several optimization methods for NP-hard bin-packing optimization problem are considered. Bin-packing problem consists of a number of items that need to be packed into a limited number of containers while satisfying the following constraints: items need to lie entirely within a container without overlapping. Bin-packing methods are also used in textile industry to optimize the material usage while cutting regular and irregular items from a fixed-width material strip. Giving a brief overview of the problem, several computational methods are described in this paper such as items and container data representation, overlapping detection and avoidance methods (constraints graphs, no-fit polygons) and optimization methods: direct search, genetic algorithm and nonlinear programming. Experiments with different number of rectangles, circles, triangles, ellipses and hexagons were performed using three methods - direct search, genetic algorithm and nonlinear programming to compare the overlapping area, total cover area and total overlapping area with nonlinear programming giving the best results.

I. INTRODUCTION

Bin-packing problem is a problem where a number of regular or irregular shapes (items) must be placed within a container (rectangle, circle, irregular shape) such that no two items overlap and all items are contained within the boundaries of a container [1].

Since the problem is applicable to a wide range of problems, bin-packing problem is not equally named nor defined in literature. In textile industry it is often referenced as a trim-loss problem, in computer science as packing problem, in logistics as knapsack problem, in industrial engineering as cutting stock problem [2].

Majority of researches have been focused on problems with packing regular geometric shapes such as rectangles and circles in larger container of regular shape such as rectangles.

In most literature, items shaped as rectangles aligned to the axes are considered. Rectangular shapes are often used for practical reasons because they simplify overlapping restrictions (compared to irregular polygons) making the algorithms faster. Polygons can be represented using a raster model. Polyhedrons can be represented using a triangular web with each triangle knowing its neighbors [1].

Besides the successful data representation, a good overlapping detection method and algorithm optimization method needs to be used.

In section II overview of overlapping detection methods is presented. In section III optimization methods are presented. In section IV we present results while in section V we conclude our work.

II. OVERLAPPING DETECTION

One of the constraints of the bin-packing problem is that items should not overlap when placed in a container.

We define I as the set of items, and C as a container. Formally, if $s_i \subset \mathrm{R}^d$ is a set occupied by item $i \in I$ in a d-dimensional problem, it is required for any pair of items i and j to satisfy $\mathrm{int}(s_i) \cap \mathrm{int}(s_j) = 0$, where $\mathrm{int}(\cdot)$ is the interior of the set. Items are allowed to abut. To assure a container covers the set $C \subset \mathrm{R}^d$ we require $s_i \cap C = s_i$ for any item in I.

How to check if items overlap depends on how items and the container are represented. For circles, and hyper-spheres in general, the sum of their radii has to be larger than the distance between their centers. In the case of orthogonally placed rectangles, if w_i^k is the size of a rectangle i in dimension k, the rectangle i occupies the set $[x_i^1, x_i^1+w_i^1]\times\ldots\times[x_i^d, x_i^d+w_i^d]$. In order to avoid overlapping, rectangles have to satisfy one of the following conditions $x_j^k \geq x_i^k + w_i^k$ or $x_i^k \geq x_j^k + w_j^k$, for at least one dimension $1 \leq k \leq d$ [1].

An overview of more sophisticated methods of checking for overlap: no-fit polygon and constraint graphs is given in this section, while methods implemented in this paper are described in section III.

A. No-fit polygon

No-fit polygon is an overlap detection method that can be constructed by translating polygon B around a steady polygon A keeping in mind polygon B's sides should never intersect with polygon A's sides, but they have to abut.

Mathematically, if a polygon is represented with a vector of points that indicate its vertices, no-fit polygon vertices can be calculated using Minkowski addition.

Overlap detection is done by checking the relationship between a reference point of a moving polygon B and the

no-fit polygon. If a reference point is inside the no-fit polygon, polygons *A* and *B* overlap.

A no-fit polygon always gives a non-overlapping solution. The main disadvantage of using no-fit polygon in overlapping detection is that it is very demanding in terms of computational complexity – it has to be calculated for every polygon pair. That is the reason this method can sometimes be too expensive [1, 3].

B. Constraint graphs

Constraint graph is a set of acyclic weighted graphs that are used to describe relative position of rectangular items. If an *n*-dimensional problem is considered, *n* constraint graphs have to be created – one for each dimension. Weights in a graph represent the edge length.

In case of a two-dimensional problem, a rectangle is represented with a reference point (e.g. coordinates of a lower left corner), width and height. Using constraint graphs, position of each rectangle (*x* and *y* coordinates of reference point) in such a packing can be easily calculated by finding longest path from initial node to a node representing a rectangle whose position we are trying to find [1].

In practice, constraint graphs for two dimensions can be represented using sequence pair data structure. Sequence pair is a pair of sequences of *n* elements representing a list of *n* rectangles. The relationship between rectangles can be found as:

$$(<\ldots b_i \ldots b_j \ldots>, <\ldots b_i \ldots b_j \ldots>) \rightarrow b_i \text{ is left of } b_i, \quad (1)$$

$$(<\ldots b_j \ldots b_i \ldots>, <\ldots b_i \ldots b_j \ldots>) \rightarrow b_i \text{ is below } b_j. \quad (2)$$

Sequence pair represents a block packing whose size can be calculated using the longest common subsequence length [4].

Placement defined using constraint graphs guarantees there will be no overlapping between items as long there is an edge between the nodes of any rectangle b_i and b_j in any of the graphs.

III. Implemented Methods

A. Direct search

Direct search is an optimization method that belongs to the class of pattern search algorithms. Direct search searches a set of points (mesh) around a current point until a point is found (polled) where the objective function value is lower than the objective function value of the current point [5].

A pattern is a set of vectors $\{v_i\}$ which are used by the pattern search algorithm to determine the points to search at each iteration. The set $\{v_i\}$ is defined by the number of independent variables in the objective function, *N*, and the positive basis set. In pattern search algorithm two positive basis sets are commonly used – the maximal basis with $2N$ vectors, and the minimal basis with $N+1$ vectors.

For example, if there are three independent variables in the optimization problem, the default for a $2N$ positive basis consists of the following pattern vectors:

$v_1 = [1\ 0\ 0]$ $v_2 = [0\ 1\ 0]$ $v_3 = [0\ 0\ 1]$

$v_4 = [-1\ 0\ 0]$ $v_5 = [0\ -1\ 0]$ $v_6 = [0\ 0\ -1]$

An $N+1$ positive basis consists of the following default pattern vectors.

$v_1 = [1\ 0\ 0]$ $v_2 = [0\ 1\ 0]$ $v_3 = [0\ 0\ 1]$

$v_4 = [-1\ -1\ -1]$

A point in direct search used in this paper is a $2N$ or $3N$ real-valued vector (see section IV). A set of points form a mesh. The algorithm searches the mesh at each step in order to find a point that improves the objective function i.e. a point with a smaller objective function value then the current point. A mesh is created by multiplying a generated set of vectors $\{v_i\}$ by a scalar Δ_m that represents a mesh size and adding these values to the current point. Current point is a point with the best objective value found at the previous step. The pattern vector that produces a mesh point is called its direction.

At each step, the algorithm polls points in the current mesh by computing their objective function values. If a complete poll method is performed, the algorithm computes the objective function values at all mesh points. Afterwards, objective function values of a polled mesh point and a current point are compared. The poll is considered successful if a mesh point has smaller objective function value then the current point.

After polling, the algorithm changes the value of the mesh size Δ_m. The default in MATLAB is to multiply Δ_m by 2 after a successful poll, and by 0.5 after an unsuccessful poll.

Global Optimization Toolbox in MATLAB includes three direct search algorithms: the generalized pattern search (GPS) algorithm, the generating set search (GSS) algorithm, and the mesh adaptive search (MADS) algorithm [5]. The GPS algorithm uses fixed direction vectors. The GSS algorithm is identical to the GPS algorithm, except when there are linear constraints, and when the current point is near a linear constraint boundary. The MADS algorithm uses a random selection of vectors to define the mesh.

With GPS, the collection of vectors that form the pattern are fixed-direction vectors. With MADS, the collection of vectors that form the pattern are randomly selected by the algorithm.

B. Genetic algorithm

Genetic Algorithm (GA) is a heuristic algorithm based on the evolutionary principle of selection and genetics. GA author, John H. Holland, presented the algorithm in the 1970s motivated by Darwin's theory of evolution [6].

The basic techniques of GA are designed to mimic the processes in natural systems that implement the Darwin evolutionary rule of the survival of the fittest. GA uses evolutionary methods to search the solution space in order to find an optimization problem solution. GA is one of the algorithms based on evolutionary computation.

The algorithm starts with initializing a population i.e. a set of individuals called chromosomes. Initial population is usually generated randomly within the solution solution space to enable searching the entire solution space. Population size depends on the type of problem, but usually the size of a few hundred to a few thousand individuals is used.

The algorithm continues with the selection process where a certain number of individuals from initial populations are selected based on their fitness to enter the reproductive process (crossover and mutation). Fitness of an individual is calculated using a fitness function *f(x)*. Good individuals are considered to pass good properties to their children in the process of crossover, and improve the quality of the population in the coming generations.

Crossover is a process of exchanging genes between parents. Individuals are crossing in order to maintain the good properties of the population. The higher the fitness of an individual is, the greater the likelihood of its survival and crossover. Crossover can create one or two children. The purpose is to get children chromosomes that will be just as good as or better than their parents [6].

After crossover some individuals mutate. Upon completion of the mutation, the new individuals are evaluated and copied into the population. Poor specimens are deleted from the population. The process of selection and reproduction is repeated until the termination condition is met. Termination condition may include maximum number of iteration achieved, maximum number of fitness function evaluations, the desired fitness function value has been achieved [7].

1) Stochastic universal sampling

In a selection process in this paper a stochastic universal sampling is used [8, 9]. Individuals are mapped to contiguous line segments. Size of each individual's segment corresponds to its fitness. Selection process is conducted by placing the equally spaced pointers over the line. Number of pointers corresponds to the number of individuals that need to be selected.

Consider number of individuals that need to be selected as NPointer. The distance between pointers equals to 1/NPointer. Position of a first pointer is given by a random number in the range [0, 1/NPointer].

2) Shuffle crossover

In this paper shuffle crossover is used to produce offspring [9, 10]. Two parents are selected, and a single crossover position is chosen. Before recombination, genes in parents are randomly shuffled, equally in both parents. After recombination, genes are unshuffled in reverse.

3) Binary mutation

In this paper binary mutation is used [9]. Since individuals are represented using binary values in which each gene has only two states, mutation in this case means flipping of gene values: 0 into 1, 1 into 0. For every individual the gene to change is chosen at random.

C. Nonlinear programming

Nonlinear programming is a method to find an optimum of a nonlinear objective function by solving a system of constraints in a form of equalities and inequalities which can be linear or nonlinear over a set of unknown real variables [11, 12, 13].

Non-linear programming is used to solve constrained or unconstrained problems. Constrained nonlinear programming is the mathematical problem of finding a vector x that minimizes a nonlinear function f(x) subject to one or more constraints. Unconstrained means that there are no restrictions placed on the range of *x*.

1) Fmincon Active Set Algorithm

In this paper a minimum of constrained nonlinear multivariable function is found using function *fmincon* from MATLAB Optimization Toolbox. Function *fmincon* attempts to find a minimum of a function starting at an initial estimate. This is generally referred to as constrained nonlinear optimization or nonlinear programming.

This type of problem can be specified by:

$$\min f(x), \begin{cases} c(x) \le 0 \\ ceq(x) = 0 \\ Ax \le b \\ Aeq \cdot x = beq \\ lb \le x \le ub, \end{cases} \tag{3}$$

where *b* and *beq* are vectors, *A* and *Aeq* are matrices, *c(x)* and *ceq(x)* are functions that return vectors, *f(x)* is a function that returns a scalar, *lb* and *ub* are lower and upper bounds of variables. Functions *f(x)*, *c(x)*, and *ceq(x)* can be nonlinear functions.

Function *fmincon* has four algorithm options: 'trust-region-reflective', 'interior-point', 'sqp' and 'active-set'. For the research purpose of this paper, active-set algorithm is used.

When using constrained optimization, the main aim is to transform a problem into easier subproblems which can then be used as the basis of an iterative process. Efficient methods have focused on the solution of the Karush-Kuhn-Tucker (KKT) equations which are necessary conditions for optimality for a constrained optimization problem.

Referring to (3), KKT equations can be stated as

$$\begin{gathered} f(x) + \sum_{i=1}^{m} \lambda_i \cdot \nabla g_i(x) = 0 \\ \lambda_i \cdot g_i(x^*) = 0, i = 1,...,m_e \\ \lambda_i \ge 0, i = m+1,...,m, \end{gathered} \tag{4}$$

in addition to original constraints in (3).

KKT equations can be solved using non-linear methods, e.g. Sequential Quadratic Programming (SQP) methods, since a QP subproblem is solved at each iteration. The principal idea is the formulation of a QP subproblem based on a quadratic approximation of the Lagrangian function.

$$L(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i \cdot g_i(x) \tag{5}$$

Equation (5) simplifies (3) by assuming that bound constraints have been expressed as inequality constraints. QP subproblem is obtained by linearizing the nonlinear constraints.

$$\begin{aligned} &\min \frac{1}{2} d^T H_k d + \nabla f(x_k)^T d \\ &\nabla g_i(x_k)^T d + g_i(x_k) = 0, i = 1,...,m_e \\ &\nabla g_i(x_k)^T d + g_i(x_k) \le 0, i = m_e + 1,...,m \end{aligned} \tag{6}$$

This subproblem can be solved using any QP algorithm (5). The solution is used to form a new iterate

$$x_{k+1} = x_k + \alpha_k d_k. \tag{7}$$

The step length parameter α_k is determined by an appropriate line search procedure so that a sufficient decrease in a function is obtained. The matrix H_k is a positive definite approximation of the Hessian matrix of the Lagrangian function. H_k can be updated by any of the quasi-Newton methods.

IV. RESULTS

In this paper experiments with different number of rectangles and polygons have been made in order to detect which algorithm gives the best results in terms of overlapping reduction. Research goal was to avoid or at least to minimize the number of overlapping polygons at a given surface using the algorithms from section III. Afterwards a comparison is made in order to find which of the following methods gives the best results.

Experiments have been made using three methods: direct search, genetic algorithm and nonlinear programming.

Experiments have been made in MATLAB environment. For experiments using genetic algorithm a Genetic and Evolutionary Algorithm Toolbox (GEATbx) has been used. The GEATbx provides global optimization capabilities for MATLAB to solve problems not suitable for traditional optimization approaches [14].

Experiments using direct search and non-linear programming have been using Optimization Toolbox™ in MATLAB. Optimization toolbox provides widely used algorithms for standard and large-scale optimization. These algorithms solve constrained, unconstrained continuous and discrete problems [15].

For each version only five tests were conducted because of the long computation time for some versions. Afterwards, an average value for each parameter was calculated. These results can be seen in Tables I and II.

A. Algorithms and parameters

The parameters that were recorded in the experiments were: value of an objective function *f(x)* (OFV), number of overlapping (NOLP), cover area (CVR) and total overlapping area (TOA).

Cover area (8) is an area covered by items in relation to total area defined as

$$coverArea = (polyArea - intArea) / area * 100, \tag{8}$$

where *polyArea* is the total area of polygons, *intArea* is an area of polygon intersection, and *area* is the total area in which all polygons reside.

For direct search, a GSSPositiveBasis2N algorithm was used from Global Optimization Toolbox in MATLAB [16]. Direct search stops after 24501 evaluations of *f(x)*.

For genetic algorithm, one subpopulation has been used with 50 individuals. For selection process a stochastic universal sampling was used with pressure factor 2, and generation gap 0.98. Shuffled crossover has been used with recombination rate of 1. Binary mutation has been used with mutation rate of 2. The algorithm terminates after 500 generations which is in this case equal to 24501 function evaluations.

For non-linear programming *active-set* algorithm has been used. Active-set algorithm solves problems with any combination of constraints. The algorithm terminates after objective function *f(x)* has been evaluated 24501 times. Constraints used are only the ones that restrict the object to go outside the solution boundaries i.e. upper bound (*ub*) and lower bound (*lb*) in (3) are defined.

1) Candidate solution structure

Direct (pattern) search and non-linear programming (NLP) use real value representation for an individual in a form of

$[x_1\ x_2\ \dots\ x_n\ y_1\ y_2\ \dots\ y_n]$,

$x_1, x_2, \dots, x_n \in R, y_1, y_2, \dots, y_n \in R$,

with n being the dimensionality of a problem, and (x_1, y_1) representing the reference point of the i-th object. For rectangles it is the lower left corner, and for polygons it is center of the object.

In the case of ellipses, a rotation component is added to value individual structure:

$[x_1\ x_2\ \dots\ x_n\ y_1\ y_2\ \dots\ y_n\ \phi_1\ \phi_2 \dots\ \phi_n]$,

$x_1, x_2, \dots, x_n \in R, y_1, y_2, \dots, y_n \in R, \phi_1, \phi_2, \dots\ \phi_n \in R$

To represent genotype of an individual in a genetic algorithm in this paper, binary values are used:

$[x_1\ x_2\ \dots\ x_n\ y_1\ y_2\ \dots\ y_n]$,

$x_1, x_2, \dots, x_n \in \{0,1\}^m, y_1, y_2, \dots, y_n \in \{0,1\}^m$,

where m is the number of bites.

Binary values are transformed into real values (phenotype) when individual's fitness is calculated i.e. the objective function f(x) is a real-valued function.

2) Objective function

Objective function *f(x)* is the sum of overlapping between every two items in a set of items:

TABLE I. DATA STRUCTURE FOR RECTANGLE TESTING

Version	Number of rectangles	Rectangle size	Area
1-1	100	Random between 50x40 and 70x60	800x600
1-2	100	70x50	800x600
1-3	10	40x80	400x400
	10	80x40	
	10	80x80	
1-4	40	40x40	800x600
	25	40x80	
	25	80x40	
	10	80x80	

TABLE II. RESULTS OF RECTANGLE TESTING

Version	Parameters	Results		
		DS	***GA***	***NLP***
1-1	OFV	18560.8	1132.46	55.16
	NOLP	70	22	0
	CVR	58.12	56.46	44.84
	TOA	18519	1088.9	0
1-2	OFV	29609.3	7523.58	0
	NOLP	93	59	0
	CVR	66.75	71,35	72.92
	TOA	29609.3	7523,58	0
1-3	OFV	908.11	1932.88	0.02
	NOLP	13	17	11
	CVR	79.43	78.79	0.8
	TOA	908.11	1932.88	0.02
1-4	OFV	1673.7	1561.66	0
	NOLP	66	29	0
	CVR	56.51	59.68	60
	TOA	1637.1	1561.66	0

$$f(x) = \sum_{i,j \in S}' overlap(x_i, x_j). \quad (9)$$

Items x_i and x_j are from a set of items S, and *overlap* is a function that calculates the total intersection area between the items x_i and x_j. The goal is to minimize $f(x)$.

3) Number of evaluations

Number of generations are 500, and number of individuals are 50 which are empirically chosen by the papers authors.

Only in the first iteration all 50 individuals have to be evaluated. Because of the elitism implanted in the genetic algorithm, the best individual from the previous iteration is copied in the new generation alongside 49 newly created individuals whose fitness is evaluated.

Therefore, summing 49 evaluations needed in 499 generations (one individuals value is copied from the previous generation), and 50 evaluations from the first generation, a number of 24501 objective function evaluations is obtained.

B. Experiments with rectangles

Four different experiments with different number of rectangles have been conducted in this research and the results are shown in Table I. Rectangles are represented with reference point (coordinates of a lower left corner of a rectangle), width and height.

TABLE III. DATA STRUCTURE FOR POLYGON TESTING

Version	Number of polygons	Polygon size	Area
2-1	40 ellipses	R_x = 80, R_y=40	800x600
2-2	15 triangles	50	800x600
	15 hexagons		
	15 circles		
2-3	10 triangles	50	800x600
	20 hexagons	30	
	30 circles	20	
2-4	15 pentagons	50	800x600
	15 hexagons		
	15 octagons		
2-5	10 pentagons	50	800x600
	20 hexagons	30	
	30 octagons	20	

TABLE IV. RESULTS OF POLYGON TESTING

Version	Parameters	Results		
		DS	***GA***	***NLP***
2-1	OFV	37980.2	30648	11403.9
	NOLP	67	72	89
	CVR	75.64	77.17	81.18
	TOA	38040.2	30648	11403.9
2-2	OFV	23.53	18.78	0
	NOLP	4	9	0
	CVR	54.92	54.921	54.925
	TOA	23.53	18.78	0
2-3	OFV	0	0	0
	NOLP	0	0	0
	CVR	24.34	24.34	24.34
	TOA	0	0	250.66
2-4	OFV	286.5	442.5	0
	NOLP	17	24	0
	CVR	60.89	60.88	60.97
	TOA	386.5	442.5	0
2-5	OFV	0	0	1599.02
	NOLP	0	0	1
	CVR	29.2	29.2	28.86
	TOA	0	0	1599.02

In versions 1-1 and 1-2 experiments with 50, 60, 70, 80 and 90 randomly generated rectangles have also been conducted, but only experiments with 100 rectangles are shown in Table II because up until 90 rectangles algorithms in all versions have successfully found arrangement without overlapping.

As it can be seen in Table II, non-linear programming (NLP) has given the best results in all versions by being able to find a feasible representation without rectangle overlapping. Genetic algorithm gave worse results then non-linear programming, but better then direct search (except in version 1-3).

In version 1-3, non-linear programming found the best solution among the used algorithms but was not able to find a representation without overlapping prior to termination because of given smaller total area and relatively larger rectangle area that needed to be placed on that area. It is possible that algorithms would be able to find better solutions in more iterations. This example shows that a non-overlapping solution is harder to find if more rectangles have to fill smaller area. Sometimes a feasible solution is not even possible.

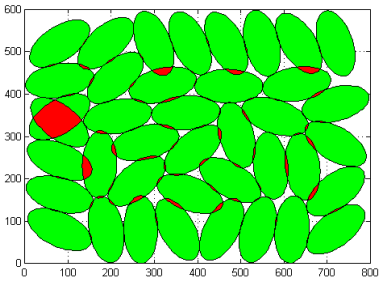


Figure 1. Results of the version 2-6 with non-linear programming

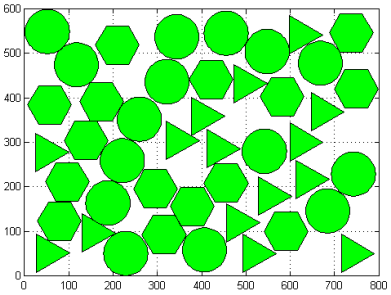


Figure 2. Results of the version 3-2 with non-linear programming

C. *Experiments with polygons*

In this part of research five experiments with different number of polygons have been conducted. Polygons are represented with center coordinates and its radius.

For version 2-1, non-linear programming gave the best results. Although the highest amount of overlaps has been found with non-linear programming (89), the algorithm still produces the greatest cover area. This means the overlapping area obtained with non-linear programming is significantly smaller than the overlapping area obtained with other algorithms regardless the number of overlaps.

In version 2-2, 2-3 and 2-4 non-linear programming has also given the best results, with genetic algorithm being the second best algorithm and direct search being the worst of the tested algorithms.

In version 2-5 direct search and genetic algorithm successfully found a feasible solution in all 5 steps, while non-linear programming was less successful, finding a solution with one overlap in average.

Only in version 2-1 none of the algorithms has been able to find a feasible non-overlapping solution prior to termination. Possible reason is that examiner allowed ellipses to rotate in order to better search the solution space, therefore search space is larger than in other cases. More iterations would have positive effect on the solution.

V. CONLUSION

In this paper four experiments with rectangles and five experiments with different number of polygons have been made. All experiments were subjected to the following methods: direct search, genetic algorithm and nonlinear programming.

The goal of the experiments was to test which of the three methods would give the best solution in terms of overlapping redundancy and minimization.

In almost all experiments non-linear programming was able to find the best feasible representation without overlapping. Depending on the problem direct search and genetic algorithm gave similar solutions, but evidently worse than non-linear programming.

REFERENCES

[1] J. Egeblad, "Heuristics for Multidimensional Packing Problems," University of Copenhagen, Ph.D. dissertation, Available http://www.diku.dk/~jegeblad/phd.pdf, 2013.

[2] H. Dyckhoff, "A typology of cutting and packing problems," European Journal of Operational Research, vol. 44, pp. 145-159., 1990.

[3] E.K. Burke, R.S.R. Hellier, G. Kendall, G. Whitwell, "Complete and robust no-fit polygon generation for the irregular stock cutting problem," European Journal of Operational Research, vol. 179, pp. 27–49, 2007.

[4] H. Murata, K. Fujiyoshi, S. Nakatake, Y. Kajitani, "VLSI Module Placement Based on Rectangle-Packing by the Sequence-Pair," IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems, vol. 15, pp. 1518-1524, December 1996.

[5] MathWorks, "Genetic algorithm and direct search toolbox 2 user's guide," Tech. rep., The MathWorks, Natick, MA, 2009.

[6] M. Golub, "Genetic algorithm, part I", "Genetski algoritam, prvi dio (skripta)". University of Zagreb, 2010., Available: http://www.zemris.fer.hr/~golub/ga/ga_skripta1.pdf, 2013. (in Croatian)

[7] D. B. Fogel, "An Introduction to Simulated Evolutionary Optimization," IEEE Transactions on Neural Networks, vol. 5, pp. 3-14, January 1994.

[8] H. Pohlheim, "Evolutionary Algorithms: Overview, Methods and Operators," Available http://www.GEATbx.com/, 2013.

[9] J. E. Baker, "Reducing bias and inefficiency in the selection algorithm," Proceedings of the Second International Conference on Genetic Algorithms on Genetic algorithms and their application, John J. Grefenstette (Ed.), L. Erlbaum Associates Inc., Hillsdale, NJ, USA, pp. 14-21, 1987.

[10] R. A. Caruana, L. J. Eshelman, J. D. Schaffer, "Representation and hidden bias II: eliminating defining length bias in genetic search via shuffle crossover," Proceedings of the 11th international joint conference on Artificial intelligence - Volume 1 (IJCAI'89), Vol. 1. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 750-755, 1989.

[11] MathWorks, "Constrained Nonlinear Optimization Algorithms," Available: http://www.mathworks.com/help/optim/ug/constrained-nonlinear-optimization-algorithms.html, 2013.

[12] W. Hock, K. Schittkowski, "A Comparative Performance Evaluation of 27 Nonlinear Programming Codes," Computing, Vol. 30, p. 335, 1983.

[13] K. Schittkowski, "NLQPL: A FORTRAN-Subroutine Solving Constrained Nonlinear Programming Problems," Annals of Operations Research, Vol. 5, pp. 485-500, 1985.

[14] GEATbx - The Genetic and Evolutionary Algorithm Toolbox for MATLAB, Available http://www.GEATbx.com/, 2013.

[15] The MathWorks - MATLAB and Simulink for Technical *Computing*, Available: http://www.mathworks.com/, 2013.

[16] T. G. Kolda , R. M. Lewis , V. Torczon, "A generating set direct search augmented Lagrangian algorithm for optimization with a combination of general and linear constraints", Sandia National Laboratories, Albuquerque, New Mexico, 2006

Investigation of coevolutionary approach in gene regulatory network inference

Danko Komlen*, Domagoj Jakobović†
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Republic of Croatia
* Email: dkomlen@gmail.com
† Email: domagoj.jakobovic@fer.hr

***Abstract*—Inference of gene regulatory networks is currently an active field of research in system biology. Evolutionary computation algorithms are lately applied for finding the optimal parameters of models. This paper presents a comparison of four evolutionary algorithms (DE, GA, PSO and the hybrid Hooke-Jeeves GA) used with a linear time-variant gene network model. The paper also investigates the efficiency of cooperative coevolution approach to cope with the increased complexity of networks with large number of genes. Experiments were performed on two artificially generated and one real microarray data set. The results are twofold: the efficiency comparison may serve as a guideline for future research, and the application of coevolution proved to be successful for most algorithms.**

I. Introduction

Each cell in a living organism has the same genome[1]. Complex mechanisms at molecular level that interpret that information are responsible for their differentiation and functionality.

A key role in such mechanisms are transcription factors that bind themselves to parts of DNA [1]. They can activate or suppress activity of a particular gene. Complex systems emerge because transcription factors are products of genes and can be regulated by other transcription factors (Fig. 1). Feedback loops are also possible.

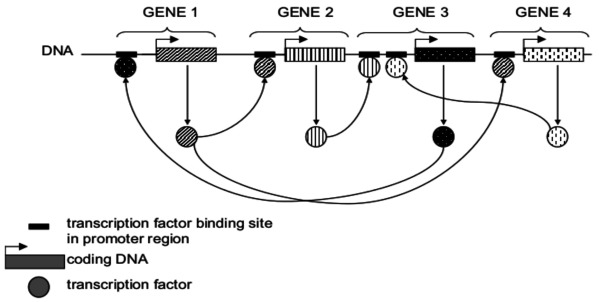


Figure 1. A simplified transcription factor network [2]

Research of these complex systems requires a way to measure the activity of individual genes within the cell. Such a procedure is called gene expression profiling and there are several different technologies available. One of the most common is DNA microarray technology [3], where levels of mRNA[2] for multiple genes are monitored in parallel.

Gene regulatory network (GRN) is an abstraction that allows understanding of these complex dynamic systems and provides an explanation for the expression of genes that are observed. Constructing a gene regulatory network involves finding a solution to a complex set of conditions, which is usually approximated by an appropriate optimization model. Descriptions of models for inferring GRN networks that are used in the paper are given in Section II.

Evolutionary computation (EC) algorithms have been successfully applied for finding the optimal or near-optimal parameters of models. This paper presents a comparison of various evolutionary algorithms used to infer the parameters of a GRN model. The comparison is based on two artificially generated data sets that were introduced in paper [4] and a real microarray data set that was also used in paper [5].

The efficiency of evolutionary algorithms may be further increased with the concept of coevolution, which is applied to all the algorithms in this work. This approach could be particularly useful for networks with a large number of genes, and consequently a large number of parameters. The results reflect the efficiency of each algorithm and the effect of applied coevolution, which may provide useful guidelines for future research.

The remainder of this paper is organized as follows: Section II outlines the GRN models, and Section III briefly covers the coevolutionary approach. Section IV details the optimization of GRN model parameters and Section V presents the results. Conclusions and future work are given in Section VI.

II. Gene Regulatory Network Models

A. S-systems

GRN networks describe biomolecular interactions that are non-linear and can be expressed by the general system of differential equations:

$$\frac{dx_i(t)}{dt} = f_i[x_1(t), ..., x_N(t)], \tag{1}$$

for i = 1,...,N, where N is number of genes, x_i gene expression level and f_i a function that describes the influence

[1] inherited information encoded in DNA

[2] messenger RNA, conveys genetic information from DNA to the ribosome

of all genes on gene i. For example, if a gene j activates gene i, then f_i increases with x_j and the other way around if a gene j inhibits gene i.

Determination of functions f_i that would define a successful model is an ill-placed problem. For this reason, we use different approximations of these functions.

S-systems are a special type of systems of differential equations where the function f_i is approximated by the following expression [6]:

$$\frac{dx_i(t)}{dt} = \alpha_i \prod_{j=1}^{N} {x_j}^{g_{i,j}} - \beta_i \prod_{j=1}^{N} {x_j}^{h_{i,j}} \tag{2}$$

Such a model is defined by a total of $2N^2+2N$ parameters.

B. Linear time-variant model

The paper [7] presents a linear time-varying model (LTV) where the following approximation is used for f_i:

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{N} W_{i,j}(t)x_j(t), \tag{3}$$

where $W_{i,j}(t)$ is a matrix of gene interactions (the control matrix). Its values depend on the current time t, which allows describing the nonlinearities in the system. Matrix values are calculated as the sum of the first two members of the Fourier series:

$$W_{i,j}(t) = \alpha_{i,j} sin(\omega_i t + \phi_{i,j}) + \beta_{i,j}. \tag{4}$$

The coefficient $W_{i,j}$ determines the strength of the influence of gene j on the regulation of gene i. A positive value indicates that the gene j activates gene i, the negative that it inhibits the gene, and zero indicates that the gene j does not affect the transcription of the gene i.

This model was also used in [8] with a slightly different approach. The authors used a discrete form of the model so that the values of genes expressions (x_i) were calculated directly by applying the sigmoid function to regulation input:

$$x_i(t+1) = \frac{1}{1+e^{(-Z_i(t))}}. \tag{5}$$

The regulation input $Z_i(t)$ denotes the influence of all genes on the gene i, and is defined as:

$$Z_i(t) = \sum_{j=1}^{N} W_{i,j}(t)x_j(t). \tag{6}$$

III. Cooperative coevolutionary algorithms

The term coevolution is most commonly used in context of biology and scientific fields devoted to the study of complex ecosystems. It is a biological change of an object that is caused by changing of another object that is interacting with it. Coevolutionary algorithms (CA) apply the concept of coevolution on metaheuristic optimization methods. Algorithms are divided into cooperative CA and competitive CA, each applicable to a specific kind of problems.

Basic classification of coevolutionary algorithms according to [9] is:

- one population competitive coevolution (1PC),
- two population competitive coevolution (2PC),
- N-population cooperative coevolution (NPC).

NPC coevolution, applied in this work, is used for problems with large solution space and where it is possible to make a decomposition into smaller subproblems. Each of the N subproblems is solved by a separate population. The fitness of the individual is calculated by the success of the entire solution, which includes itself and the best individuals from each of the other N-1 populations.

Coevolutionary algorithms can be realized sequentially and in parallel. For algorithms that operate over each population, various optimization algorithms (metaheuristics) could be used.

IV. GRN inference problem

Each GRN network model contains a set of parameters that define it. After selecting the type of model, the next step is to determine the values of parameters. This is a continuous optimization problem where the objective is to minimize the error between the data obtained by model simulation and experimental data (MAD measurements).

A. Problem description

For the modeling of GRN network, a linear time-varying model was used described in the [8], where the gene expressions were calculated by expression 5. Parameter set for the network of N genes consists of the following $3N^2 + N$ parameters:

$$\{\alpha_{i,j}, \beta_{i,j}, \phi_{i,j}, \omega_i | i,j \in \{1,...,N\}\}. \tag{7}$$

The fitness function is defined as

$$f = -\sum_{k=1}^{M}\sum_{t=1}^{T_k}\sum_{i=1}^{N}\left(\frac{X_{k,i,mod}(t) - X_{k,i,measured}(t)}{X_{k,i,measured}(t)}\right)^2, \tag{8}$$

where M is the number of data sets, t_k the number of measurements in the k-th set, and $X_{k,i,mod}(t)$ and $X_{k,i,measured}(t)$ are expression levels of the gene in time t as a result from the model and experimental results, respectively.

Each individual in EC algorithm contains a specific set of parameter values. Also, for each parameter type boundary values are predefined.

B. Application of coevolutionary algorithm

The number of parameters of the linear time-varying model grows quadratically depending on the number of genes in the data set. For this reason, evolutionary algorithms are searching a large solution space and the execution time increases. This kind of problem is appropriate for cooperative coevolution with the condition that division of the problem into smaller subproblems can be defined. Two different ways of divisions used will be described next.

The first problem division approach is the use of N populations, where individual from population i contains $3N + 1$ parameters: $\{i,j, \beta_{i,j}, \phi_{i,j}, \omega_i | j \in \{1, ..., N\}\}$. Thus, each population optimizes parameters for a particular gene. During the evaluation, the best individuals from other populations create the total solution with all the parameters that are evaluated over the given data set (Fig. 2).

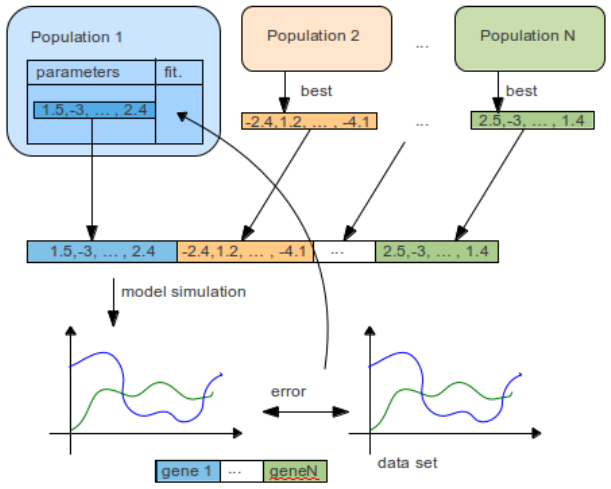


Figure 2. The first approach for problem division in NPC coevolution

The second approach, used in [10], shares model parameters in the same way by use of N populations. However, when evaluating the parameters for a particular gene, the following modified expression for the regulation input is used:

$$Z_i(t) = \sum_{j=1}^{N} W_{i,j}(t) Y_j(t)$$
$$Y_j(t) = \begin{cases} x_j(t), & j = i \\ \hat{x}_j(t), & \text{otherwise} \end{cases} \quad (9)$$

where $\hat{x}_j(t)$ is the expression of j-th gene at time t obtained by simulation of the model whose parameters are taken from the best individuals from the previous generation. So, at the end of each iteration the model is simulated using the best individuals and thus generates $\hat{x}_i(t)$ for each gene, which is then used in calculating the regulation inputs for the next generation (Fig. 3).

C. Test sets

A test set refers to a set of one or more data sets, which are used for testing the algorithms. Their overview is given in table I.

Data sets Tominaga2 and Tominaga5 are artificially obtained using S-systems that were presented in the paper [4] and are often used for measurement of algorithms performance. Data set Tominaga2 is shown in figure 4. The x axis indicates time; the time *intervals* between the two measurements can vary, but for LTV model only their ordinal number (T) is relevant.

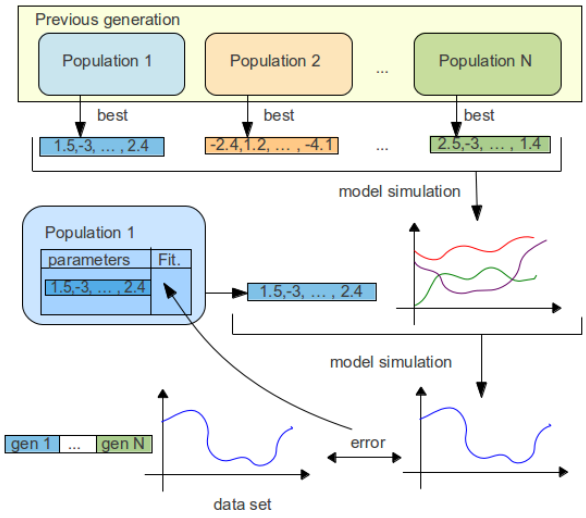


Figure 3. The second approach for problem division in NPC coevolution

Table I
TEST SETS OVERVIEW

Name	Gene count (N)	Measurement count (T)	Data sets count (K)	Data sets type
IS-param	5	20	1	Tominaga5
IS-t2	2	11	1	Tominaga2
IS-t5	5	11	10	Tominaga5
IS-yeast6	6	18	1	Spellman d.s.

The y axis indicates the gene expression levels and contains values between 0 and 1.

Test set IS-yeast6 contains experimental measurements of two cell cycles of the *Saccharomyces cerevisiae* organism, and the data was taken from Spellman data set [3].

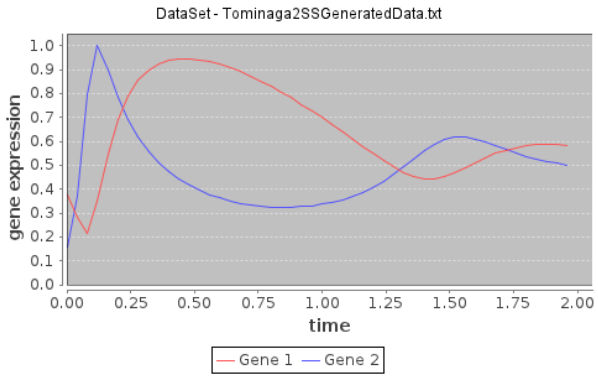


Figure 4. Tominaga2 data set

D. Overview of related work

The problem of modeling GRN network is one of the actual problems in systems biology field of research. Application

[3]available in the Kegg database (http://www.genome.jp/kegg/)

of evolutionary computation to optimize the parameters of the model is only one possible approach. Table II gives the representation of used algorithms, models, fitness functions and data sets from several papers related to application of evolutionary algorithms for GRN model inference.

The paper [5] gives comparison of the different evolutionary algorithms in modeling GRN network with an S-system. Seven algorithms were compared: GA, MOGA (multiobjective GA), GA + ES, GA + ANN (GA with artificial neural networks), PEACE1 [6] GLSDC [11] and DE. GA + ANN and DE proved to be the best on real world GRN networks. Tests were carried out on artificially generated networks and real DNA microarray data (Spellman data set).

Table II
OVERVIEW OF THE RELATED WORK

Paper title	**Evolutionary algorithm**	**Model type**	**Data sets (size)**
[8]	SA-DE	LTV	Generated (5), E. Coli (6)
[10]	coev-GLSDC	SS	Generated (5,30), Theromophilus (25)
[5]	GA, DE, GA+ANN, MOGA, PACE1, GLSDC, GA+ES	SS	Generated (10-50), Spellman d.s.
[12]	GA-simplex	diff. equ.	Rice (1)

The paper [8] used a linear time-varying model with differential evolution algorithm and self-tuning parameters. Tests were performed on Tominaga5, E.coli SOS and cAMP data sets with additional 5% and 10% noise. The algorithm has achieved satisfactory results after a relatively short running time.

In [10] a cooperative coevolution was used in combination with GLSDC algorithm on Tominaga5, S-system with 30 genes with 10% noise and *Thermus thermophilus* HB8 MAD data sets. The approach was successful in overcoming the problem of dimensionality with a larger number of genes. The results obtained for a set Tominaga5 are roughly $2 \cdot 10^{-3}$, and approach proved more successful than the usual problem decompositions.

V. ALGORITHM EVALUATION

The first series of experiments served to determine the efficiency of each of the applied algorithms: genetic algorithm (GA), differential evolution (DE), particle swarm optimization (PSO) and a hybrid algorithm based on GA and Hooke-Jeeves local search (HIB). For every evaluated algorithm we need to specify:

1) algorithm parameters,
2) model parameter intervals,
3) test set,
4) stop condition.

Algorithm parameters used for evaluation were determined with param-IS test set, while the other sets were used for comparison of performance. Stopping condition and the model parameter intervals have been specified separately for each of the tests.

A. Algorithm parameters search

Each of the four tested algorithms has its own set of parameters that determine its behavior for a given test set. Optimal parameters were determined over the IS-param test set with best solution stagnation and a time limit as a stopping condition, and 10 repetitions for each parameter value.

The algorithm would stop if the fitness of the best individual in 100 iterations has not increased by at least 0.01 or if the time limit of 3 minutes has been reached. The value of a parameter that would give the best average value of fitness of best individuals from each of the trials, would then be taken as optimal. For testing, independence of the parameters has been assumed and the parameters were determined one after another. At each test, the optimal values of the previously determined parameters were used. Initial and final optimal parameter values are given in tables (III - VI). Parameters are determined from left to right as they appear in the tables.

For population size (n) used values were 50, 100, 150, 200 and 250. Crossover probability (p_c) in the genetic algorithm and hybrid Hooke-Jeeves GA assumed values 0.5, 0.6, 0.7, 0.8, 0.9, and mutation probability (p_m) was 0.02, 0.04, 0.06, 0.08, 0.1 and 0.12. For differential evolution three types of differentiation were tested (introduced in paper [13]) with uniform and exponential crossover. For values of the parameter F the following values were used: 0.5, 1, 1.5, 2 and 2.5. In particle swarm optimization the size of the neighborhood was 2, 4, 8, 16, 32, and parameters C_1 and C_2 assumed values 0.5, 1, 1.5, 2, 2.5 and 3.

After the estimation of each algorithm's parameters, two coevolutionary problem division approaches (see Sec. IV-B) were investigated, as well as the number of iterations for the base algorithm that is executed over each subpopulation in a cooperative environment. The second approach proved to be superior (Table VII shows the average error), and results for the number of iterations are shown in Table VIII.

Table III
GA ALGORITHM PARAMETERS

	n	p_c	p_m
initial	200	0.8	0.05
optimal	250	0.8	0.06

Table IV
DE ALGORITHM PARAMETERS

	n	differentiation	crossover	F	p_c
initial	200	DE/best/1	uniform	0.5	0.9
optimal	200	DE/best/1	uniform	0.5	0.7

B. Evaluation results

The aim of the second series of experiments is the comparison of all the algorithms and their coevolutionary versions

Table V
PSO ALGORITHM PARAMETERS

	n	k	C_1	C_2
initial	200	2	2	2
optimal	250	8	0.5	1

Table VI
HIB ALGORITHM PARAMETERS

	n	p_c	p_m	Δx_0	ε
initial	50	0.8	0.05	1	10^{-6}
optimal	50	0.7	0.1	-	-

Table VII
COMPARISON OF COEVOLUTION PROBLEM DIVISION APPROACHES

	DE	GA	HIB	PSO
first approach	-1.1254	-0.076895	-4.1002	-6.0824
second approach	-0.099778	-0.025158	-0.21125	-0.023291

Table VIII
NUMBER OF ITERATIONS FOR COEVOLUTIONARY ALGORITHM

	DE	GA	HIB	PSO
initial	5	5	5	5
optimal	2	10	2	5

on different data sets. In all the following tests a stopping condition of maximum $3 \cdot 10^6$ evaluations and a time limit of 10 minutes was used. Each test has been repeated 50 times.

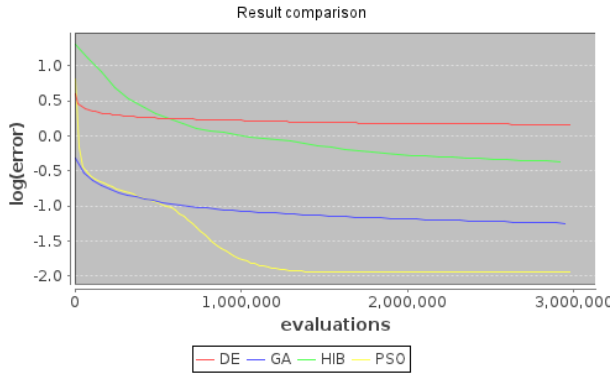


Figure 5. Average fitness for IS-param test set

1) Test set IS-param: Figure 5 shows the average value of the best solutions of all 50 repetitions, depending on the number of evaluations for the IS-param test set. PSO algorithm gives best solutions on average, and hybrid algorithm found the best overall solution with an error of 0.0022.

For further comparison of the results a boxplot graph type was used [14]. The graph shows the median, upper and lower quartile and a minimum and maximum value that is within the interval size of 1.5 IQR (interquartile range, the difference between the upper and lower quartiles). Outliers are marked with circles.

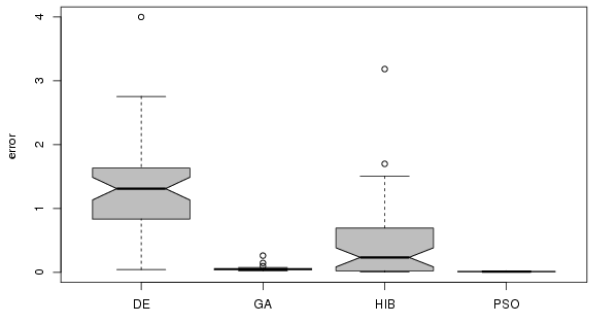


Figure 6. Result comparison for IS-param test set without coevolution

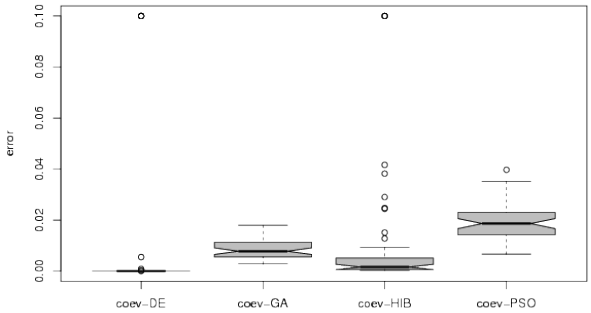


Figure 7. Result comparison for IS-param test set with coevolution

Fig. 6 shows that GA and PSO have the smallest and DE the greatest discrepancies in the results.

Coevolution results show improvements in all algorithms except the PSO [4]. The best solution is found with the DE algorithm with error of $2,122 \cdot 10^{-7}$, which is a significant improvement. Also from the Fig. 7 it is noticeable that the solution quality of DE is rather uniform, as opposed to tests without coevolution.

2) Test set IS-t2: The best average error and the best solution for the IS-t2 test set was achieved by the hybrid algorithm (Fig. 8). As in the previous case improvement with the use of coevolution is visible in all algorithms except for PSO (Fig. 9). Improvements are somewhat lower than in the previous set, which can be attributed to a small number of model parameters, where algorithms without coevolution give good results in the first place.

3) Test set IS-t5: IS-t5 test set has the most data sets and from this point of view is more difficult compared to other sets. On average, the best solution was obtained by PSO algorithm, and the best overall solution was found by the DE algorithm. Fig. 10 shows that error for the hybrid algorithm has

[4] when comparing algorithms with and without coevolution the average error was observed

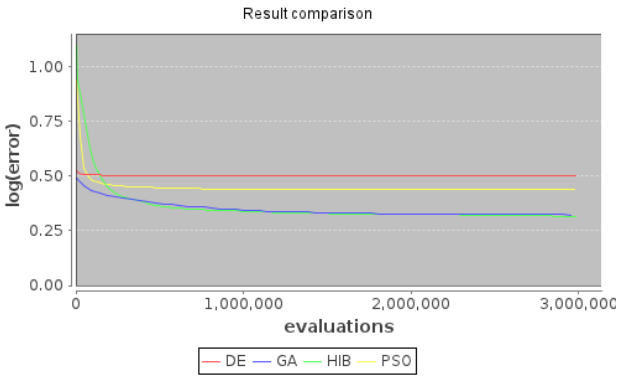


Figure 8. Average fitness for IS-t2 test set

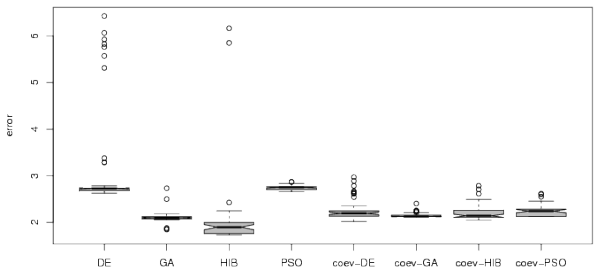


Figure 9. Result comparison for IS-t2 test set

the slowest decline, from which it can be assumed that with increasing number of evaluations it could have better results. With the application of coevolution all algorithms except PSO were improved (Fig. 11).

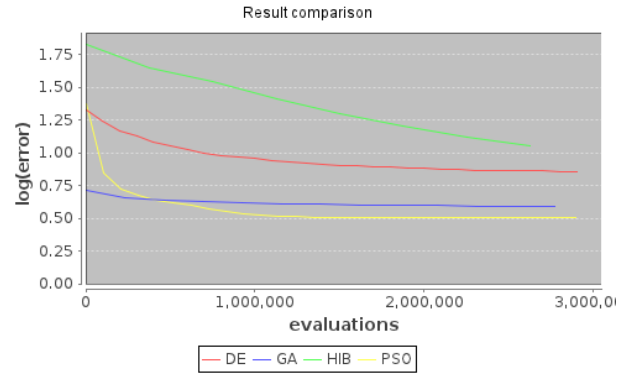


Figure 10. Average fitness for IS-t5 test set

4) Test set IS-yeast6: The best results in the last test set were obtained by PSO, and the evolution rate is shown in Fig. 12). It is interesting that the application of coevolution in this test set also improves all algorithms except the PSO. One possible explanation of this behavior is that the PSO does not adapt to constant changes in the solution space. In coevolution the fitness of each individual is calculated in

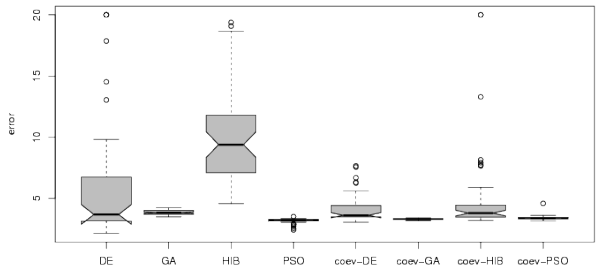


Figure 11. Result comparison for IS-t5 test set

relation to the other populations, which are also changing. The PSO algorithm contains additional information about the particle momentum, which could have a negative impact at the next iteration after the other populations change.

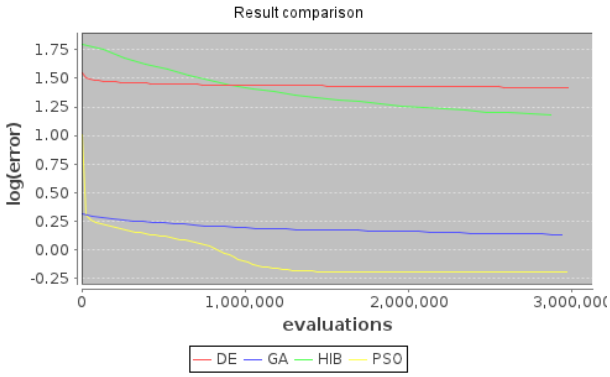


Figure 12. Average fitness for IS-yeast6 test set

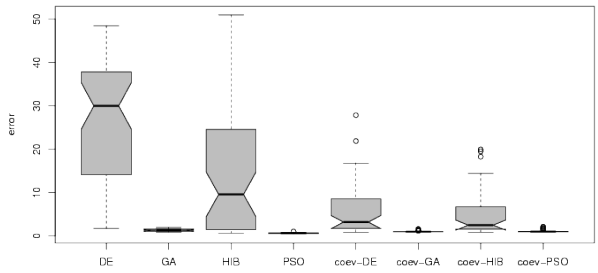


Figure 13. Result comparison for IS-yeast6 test set

C. Statistical analysis on applied coevolution

The impact of the application of the coevolution is further evaluated using T-test for the solutions obtained with and without coevolution. This type of test gives the probability that two samples originated from the same population with

the same mean. For each algorithm 50 solutions obtained without and 50 obtained with the use of coevolution were compared for each of the test sets, and the results are shown in Table IX. It is evident that most of the values are very small, indicating that the application of coevolution produces statistically significant differences. In PSO algorithm, however, the difference comes at the expense of coevolution. The values for the hybrid algorithm for two test sets are slightly larger and it is generally difficult to conclude whether in this case coevolution is a benefit.

Table IX
T-TEST RESULTS ON APPLIED COEVOLUTION

	DE	GA	HIB	PSO
IS-param	$< 10^{-6}$	$< 10^{-6}$	$2.342 \cdot 10^{-5}$	$< 10^{-6}$
IS-t2	$< 10^{-6}$	0.011	0.132	$< 10^{-6}$
IS-t5	0.053	$2.902 \cdot 10^{-4}$	0.106	$< 10^{-6}$
IS-yeast6	$< 10^{-6}$	$< 10^{-6}$	$2.87 \cdot 10^{-5}$	$< 10^{-6}$

VI. CONCLUSION

Modeling of gene regulatory networks is currently an active area of research in the field of systems biology. Creating successful models provides greater insight into cellular processes and improves the possibility of their predictions. Evolutionary computation algorithms allow search for optimal parameters of the model and, according to recent contributions, give promising results.

It is difficult to pinpoint the algorithm that would obtain the best results for every data set, so a viable approach suggests the use of multiple algorithms. The hybrid algorithm provided a more stable convergence, although requiring a larger number of evaluations. Coevolution was also proved beneficial, as it has caused improvements for all algorithms (except PSO) in each test set. We believe that the additional data structures the PSO maintains with every individual may be the main cause of its deterioration with the use of coevolution.

Possible further research includes assessing additional types of algorithms and model types, testing larger data sets and conducting experiments in the presence of noise.

REFERENCES

[1] S. K. Pal, S. Bandyopadhyay, and S. S. Ray, "Evolutionary computation in bioinformatics: a review," vol. 36, no. 5, pp. 601–615, 2006.

[2] T. Schlitt and A. Brazma, "Current approaches to gene regulatory network modelling," *BMC Bioinformatics*, vol. 8, no. Suppl 6, pp. S9+, 2007.

[3] P. Baldi, G. W. Hatfield, and W. G. Hatfield, *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*, 1st ed. Cambridge University Press, 2002.

[4] D. Tominaga, M. Okamoto, Y. Maki, S. Watanabe, and Y. Eguchi, "Nonlinear Numerical Optimization Technique Based on a Genetic Algorithm for Inverse Problems: Towards the Inference of Genetic Networks," in *German Conference on Bioinformatics (GCB)*, 1999.

[5] A. Sirbu, H. Ruskin, and M. Crane, "Comparison of evolutionary algorithms in gene regulatory network model inference," *BMC Bioinformatics*, vol. 11, no. 1, p. 59, 2010.

[6] S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi, and M. Tomita, "Dynamic modeling of genetic networks using genetic algorithm and s-system," *Bioinformatics*, vol. 19, no. 5, pp. 643 – 650, 2003.

[7] J. Kim, D. G. Bates, I. Postlethwaite, H. P. Harrison, and K. H. Cho, "Linear time-varying models can reveal non-linear interactions of biomolecular regulatory networks using multiple time-series data," *Bioinformatics*, vol. 24, no. 10, pp. 1286–1292, 2008.

[8] M. Kabir, N. Noman, and H. Iba, "Reverse engineering gene regulatory network from microarray data using linear time-variant model," *BMC Bioinformatics*, vol. 11, no. Suppl 1, p. S56, 2010.

[9] S. Luke, "Essentials of metaheuristics," pp. 103–125, 2009.

[10] S. Kimura, K. Ide, A. Kashihara, M. Kano, M. Hatakeyama, R. Masui, N. Nakagawa, S. Yokoyama, S. Kuramitsu, and A. Konagaya, "Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm," *Bioinformatics*, vol. 21, no. 7, pp. 1154–1163, 2005.

[11] S. Kimura and A. Konagaya, "High dimensional function optimization using a new genetic local search suitable for parallel computers," *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, vol. 1, pp. 335 – 342, 2003.

[12] P. Koduru, S. Das, S. Welch, and J. L. Roe, "Fuzzy dominance based multi-objective ga-simplex hybrid algorithms applied to gene network models," in *GECCO (1)*, 2004, pp. 356–367.

[13] K. V. Price, "An introduction to differential evolution," pp. 79–108, 1999.

[14] T. Schlitt and A. Brazma, "Current approaches to gene regulatory network modelling," *BMC Bioinformatics*, vol. 8, no. Suppl 6, pp. S9+, 2007.

Evolution of Vehicle Routing Problem Heuristics with Genetic Programming

Matija Gulić, Domagoj Jakobović
Protok d.o.o., Zagreb, Croatia, e-mail: matija.gulic.hr@gmail.com
University of Zagreb, Faculty of Electrical Engineering and Computing, e-mail: domagoj.jakobovic@fer.hr

***Abstract*—Increasingly complex variants of the vehicle routing problem with time windows (VRPTW) are coming into focus, alleviated with advances in the computing power. VRPTW is a combination of the classical traveling salesman and bin packing problems, with many real world applications in various fields – from physical resource manipulation planning to virtual resource management in the ever more popular cloud computing domain. The basis for many VRPTW approaches is a heuristic which builds a candidate solution that is subsequently improved by a search or optimization procedure. The choice of the appropriate heuristic may have a great impact on the resulting quality of the obtained schedules. In this paper we use genetic programming to evolve a suitable heuristic to build initial solutions for different objectives and classes of VRPTW instances. The results show great potential, since this method is applicable to different problem classes and user-defined performance objectives.**

***Index Terms*—vehicle routing problem with time windows, genetic programming, heuristic scheduling**

I. Introduction

Travel logistics have always been important to minimize expenses and maintain a good quality of service. A good formal representation of real-life transportation and service delivery problems is the capacitated vehicle routing problem with time windows (VRPTW). It is a problem of having to minimize the distance that vehicles need to travel in order to deliver a service with multiple realistic constraints regarding space, time and capacity. It has been shown that solutions for this problem could be applied to real-life domains and significantly improve their efficiency – up to 30% in some cases [1].

VRPTW is an NP-hard problem [2], so the aim is to find an approximate solution of high-enough quality. There are many solutions to the VRPTW and the most promising ones, such as [3], [4], apply the so-called optimization decomposition technique to divide the search into separate constraint dimensions. The method used in this paper also applies a decomposition based on separate vehicles and then uses parallel local search for optimization within each decomposed dimension.

However, for a decomposition approach to be successful, a good initial solution is needed which the decomposition algorithm could transform into a solution of an acceptable quality. For this purpose usually simple and fast greedy heuristics are used for creation of initial solutions that are further improved by decomposition operators. Here a choice must be made to select an appropriate heuristic, which may be highly dependent on the given performance objective, as different heuristics yield very large differences in the final solution quality [5].

Instead of manually selecting (and sometimes guessing) which heuristic would be suitable for initial solutions, we propose to evolve (i.e. automatically generate) an appropriate heuristic with the use of genetic programming (GP, [6]). Genetic programming has the ability to evolve any form of algorithm, simply by defining its building elements and a measure of algorithm quality (the fitness function). With this approach, we may create heuristics tailored to the problem at hand, regardless of the given performance objective and specific constraints [7][8]. The GP evolved heuristic does not produce the final solution, but the initial solutions obtained in this way may provide much better final solutions after a decomposition phase, which is investigated in the results. This paper is an extension of our previous project [5] where only the decomposition approach was presented, without the use of GP for the creation of initial solutions.

The rest of the paper is organized as follows. Section II gives a formal problem statement and introduces the considered model, as well as listing the related work in the field. The method of creating the solution with GP generated heuristic functions is described in Section III. Evaluation procedure and numerical results are described in Section IV. Section V gives possible future research directions and concludes the paper.

II. Vehicle Routing Problem with Time Windows

The vehicle routing problem with time windows is an extension of the well-known vehicle routing problem (VRP, defined in [9]). The VRP is described as follows: "A set of n customers must be serviced from a central office using vehicles of equal given capacity". Each customer must be served from exactly one vehicle. Usually two objective criteria are used, the primary to minimize the number of vehicles and the secondary to minimize the total travel distance. An additional constraint associated with the time windows extension is that every customer must be serviced within a given time frame. If a vehicle arrives earlier it must wait for the window opening time (waitTime). If the vehicle arrives after the end of the time window, the solution is not valid. Every customer has the following parameters defined:

- readyTime - window opening time,
- dueDate - window closing time,
- serviceTime - time needed for the customer to be serviced,

- demand - customer capacity
- geographical data.

Other parameters, considered in this work, include:

- distance - geographical distance between two customers,
- timeDistance - time needed to travel from one customer to another,
- windowTime - difference between dueDate and readyTime.

In the algorithm used here to solve the VRPTW, a single objective criterion is defined with the goal to minimize the total number of vehicles multiplied with 10000 and summed with traveled distance in kilometers:

$$objective = (\text{vehicle count} * 10000) + \text{distance traveled} \quad (1)$$

Using such objective criterion primary focuses on generating solution with minimal vehicles number, and in case of equal vehicles number than criterion considers traveled distance. Generally, a smaller vehicle number implies a smaller travel distance. Nevertheless, the proposed method of heuristics generation may be used with any conceivable performance objective.

A. Related Work

A lot of work has been invested in creating efficient solvers for the vehicle routing problem. A constraint decomposition approach, where a possible solution is optimized for each of the constraints in turn and then combined, was proposed in [10], [3]. The problem space is decomposed to spatial, time and vehicle dimensions. As for the optimization techniques inside the VRP domain [11], various algorithms are used. A simulated-annealing like local search was proposed in [12]. In [13] a tabu search optimization is used. The naive ejection chain method for local neighborhood searching is proposed in [14] with a high potential for solution diversification and total vehicle number reduction. Models for evolutionary methods were presented in [15], [16]. Ant-colony optimization was used in [17]. Proposed optimization approach is based on a parallel local search algorithm once the solution space is decomposed into small enough instances. A decomposition technique which reduces the number of vehicles and can also be executed in parallel was described in [5]; this approach is used in this paper, but it requires a suitable initial solution to be constructed beforehand.

Many solving procedures rely on a reasonably good initial solution, which is then improved with metaheuristics, local search algorithms, decomposition techniques and combinations of those. The problem that researches face is the choice of an appropriate method of creation of initial solutions (usually a simple greedy heuristic), which is highly dependent on the given performance objective. To the best of our knowledge, no previous work has employed genetic programming for generation of greedy algorithms that would be used to build the initial solution. This approach allows automatic creation of different greedy heuristics that may be tailored to specific objectives and classes of the problem.

III. Initial VRPTW Solutions with Genetic Programming

A. Initial solution creation using a heuristic function

The VRPTW solving algorithm used in this paper is divided in two major parts. Firstly, an initial solution is generated using a fast greedy algorithm. Then the number of vehicles is reduced using parallel customer insertion, which is described in greater detail in [5]. One of the most important advantages of the implemented parallel method is easy patching of subresults into a global result which fully meets the defined constraints.

To create an initial solution, a greedy algorithm is used that creates a list of customers for each vehicle. The input parameters are the initial customer, the vehicle capacity and a list of unvisited customers (that still need to be visited). The pseudocode of this phase of the algorithm follows:

Algorithm 1 Create initial solution

```
while unvisitedCustomer exists do
    start newRoute;
    set newRoute.first(home depot);
    while true do
        posibleCustomers <- unvisited valid customers;
        if no valid customers then
            break;
        end if
        nextCustomer <- select one posibleCustomer with minimal OBJ value;
        newRoute.add(nextCustomer);
    end while
    set newRoute.last(home depot);
    add newRoute to result[route];
end while
```

In this phase, a list of customers to visit is built by taking into consideration the criterion expressed with the OBJ value. Here the greedy heuristic is used which provides the OBJ value for every combination of current state and customer. The heuristic is in fact reduced to a function that uses customer information (such as distance, due date etc.) to identify the next customer - the one with the lowest function value. In our previous work [5] we experimented with several functions, e.g.:

- $fun1 = distance$,
- $fun2 = readyTime$,
- $fun3 = x * distance + y * waitTime + z * (dueTime - visitedTime)$.

In this case, we use genetic programming to evolve a suitable function. The 'suitable' function may be evolved for a single test case (single set of customers) or the same function may be evolved and then used with several new (unseen) test cases - both approaches are investigated and presented in the results.

B. Genetic programming

Genetic programming [6] is an optimization and machine learning technique that uses evolutionary concept to automatically discover symbolic procedures (functions, programs) to

the problem at hand. The main idea behind GP is that the solution to the problem may be represented as a (computer) program, in most applications in the form of a tree (which allows mapping to any procedural language). The elements of the programs (tree nodes) must be predefined by the user and must be sufficient to describe the solution to the problem (e.g. mathematical and logical functions, variables, actions such as move forward, turn left etc). The algorithm randomly generates functions (potential solutions) and evaluates each function on a predefined set of test cases (e.g. how well does the function describes the data). Each potential solution thus receives its quality estimate - the fitness value - which is then used in the selection process.

The selection process imitates natural evolution where weaker individuals (solutions) are eliminated, and better individuals survive. Additionally, better individuals also participate in recombination, where two (or more) individuals are combined to form a new solution. The algorithm also incorporates a mutation mechanism, where a single individual is subject to a change, with a relatively small probability. The process continues, building new generations from old ones, until a suitable termination criterion is reached. These criteria usually include finding a solution of the desired quality or running the algorithm for a predefined amount of time. The examples of human-competitive results of genetic programming may be found in [18].

C. Creating the solution with GP generated heuristics

In this paper the genetic programming is used to automatically create a heuristic function used in initial solution creation. The function is represented as a tree, where inner nodes are operators and leaves are variables. The operator and variable set must be manually defined; the operators used here consist of arithmetic functions +, - and *, with subsequent experiments with a square root function. The variables used are the same values that represent a single VRPTW customer, denoted with single letters for GP readability:

- distance to the next customer (denoted as 'a'),
- readyTime ('b'),
- dueDate ('c'),
- demand ('d'),
- visitedTime - time when services start for customer ('e') and
- waitTime ('f').

This way the GP is able to create any of the manually created greedy heuristics, and possibly find a few better ones. GP creates a set of functions (individuals) in each generation, and each of those must be evaluated. This is performed in the following manner:

while there are individuals to evaluate **do**
  build initial solution using the current individual;
  decompose initial solution to obtain the final solution;
  assign the final solution quality as the current individual's fitness;
end while

After the whole generation is evaluated, genetic operators (crossover, mutation and selection) are performed, thus producing the next generation of individuals (new candidate functions). The whole process is repeated until a predefined stopping criterion is met.

D. Decomposition

Once an initial solution is created, the result decomposition is started which tries to reduce the total vehicle number. Each result can be divided into independent vehicles, where each vehicle has a list of customers that have to be visited. The only way to reduce the number of vehicles is to reallocate customers from a specific vehicle and thus remove that vehicle from the solution. The reduction algorithm iterates over all vehicles; for each vehicle all of its customers are assigned to other vehicles. Other vehicles can accept a customer only if the capacity and time constraints have been met. In case other vehicles can accept all these customers, the total number of vehicles is reduced by one and we say that reduction has succeeded. The process is repeated until no more reductions can be made. Specific details for this approach can be found in [5].

IV. Experimental Setup and Results

A. Implementation environment

Implementing the system which would be able to create candidate functions using GP and validate those functions required the use of two frameworks. For greedy functions evolution the Evolutionary computation framework (ECF,[19]) was used. For evaluating greedy functions the VRPTW framework (developed by authors) was used. Communication between ECF and VRPTW consists only of providing a string formatted greedy function, and after that the VRPTW would return a value which is the performance measure for generated function.

B. Experimental Setup

The experiments were conducted on Gehring and Homberger benchmark set [20] containing problems with 1000 customers. Benchmark set is divided into 6 groups (C1, C2, R1, R2, RC1, RC2). Groups named Cx contain problems in which geographical data is clustered, whereas those named Rx have randomly generated geographical data and RCx contain problems with a combination of both. Problems in groups x1 have a short scheduling horizon and allow only a small number of customers per route. In contrast, groups x2 have a long scheduling horizon and the number of customers per route is significantly bigger. For all problems, the travel time and distance is equal to the corresponding euclidean distance [21]. All problems have a central depot, as well as capacity and time window constraints.

Starting parameters for GP include mutation rate of 0.3, population of 30 individuals and stopping criterion as 20 generations without improvement. For greedy function elements we used operators +,- and *, tree depth is varied from 2 to

4, and variables are distance (a), readyTime (b), dueDate (c), demand (d), visitedTime (e) and waitTime (f).

In the first experiment all 60 problem instances were combined in one set of test cases. In other words, we try to find a single greedy function that would be used in initial solution creation for every problem instance. The result of combined instances is defined as sum of results from each instance. Best known result for all combined instances is 3418 vehicles [22], whereas the evaluation of the described method provided the result of 3763 vehicles (with the heuristic function $a + c + d + 2e - f$). In comparison, a hand-made heuristic function used in the previous work [5] achieved the result of 3781 vehicles.

Although slightly better than a hand-made heuristic, this result suggests that it is very difficult to find a single function that would yield good results on a larger set of problem instances. This may be attributed to differing characteristics of the problem set, as well as to a very large solution space, which requires much computation effort to find a good solution. Better results should be obtainable if the set of problem instances is divided into smaller groups, containing only a few (or even one) instances.

In the next phase, therefore, each of 6 groups from the benchmark set is divided into a train set (containing 5 instances for each group, with indexes 1,3,5,6,8) and a test set (containing the other 5 instances for each group, indexed 2,4,7,9,10). For each training group a separate greedy function is evolved with GP. Table I reports results generated using the evolved greedy functions on the corresponding test set. The best known total number vehicles for the 6 test sets is 1693, and with GP evolved functions the result was 1822, which is a deviation of 7.6%. Using training and test sets significantly depends on the benchmark set, and in this case in specific groups of problem instances.

Finally, in the next experiment a separate heuristic function is evolved for each of 60 instances. In this case, with maximum specialization, the obtained vehicles sum from all instances is equal to 3611, which is significantly better than a hand-made solution of 3781, and corresponds to a deviation of 5.6% from the best known solution. This indicates that the best approach is to find a function that would be used with a single test case; since this is rather time consuming, the influence of GP parameters should be investigated to reduce the required computation time.

C. Parameters of GP

After testing on combined multiple instances, our focus changed to single instances in order to tune up parameters. Among the most important ones are the stopping condition and the population size, which are tested below. Three random problem instances (C1_10_5, R2_10_5, RC1_10_5) were chosen for further testing.

To estimate the required computational effort to reach an acceptable solution quality, we try to discover an appropriate termination condition for the GP evolution, without impairing the quality of the final result. A practical termination condition usually includes stopping the evolution after no improvement has been made in a predefined number of generations, or fitness evaluations. The termination criterion was initially fixed to 50 generations without improvement. GP was then applied to the above problem instances and the evolution is repeated 50 times. For these 50 runs, Fig. 1 represents the number of runs in which the best result occured at a certain generation range. It can be seen that the majority of occurrences of the best solutions are within the first 30 generations; in the following experiments the termination condition was therefore set to 30 generations without improvement.

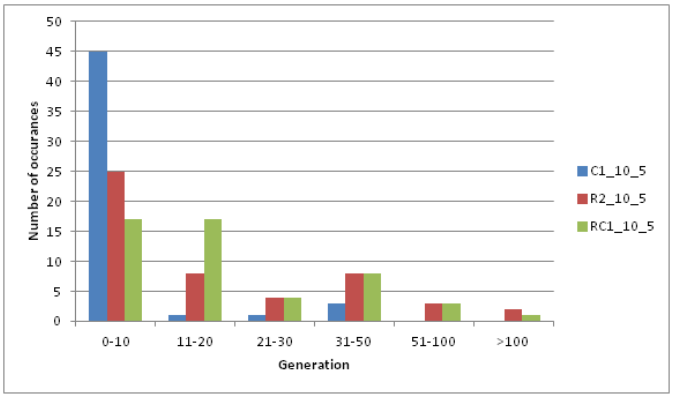


Figure 1. Influence of termination criterion

The second investigated parameter is the adeqaute population size; tables II, III and IV show minimum, average and maximum results for 3 different population sizes (10, 25 and 50 individuals) tested 50 runs for each value. In addition to the objective values, we also report the percentage of runs in which the corresponding minimum value was found (in brackets). The results are presented using the defined goal function (1). The results suggest that using 25 individuals provides a good enough average, so it would be preferable to use population size of at least 25 individuals for Gehring and Homberger test set.

Table II
POPULATION TESTS - C1_10_5

	10 (individuals)	25 (individuals)	50 (individuals)
min	1000050 (10%)	1000050 (20%)	1000050 (38%)
avg	1001859	1000058	1000056
max	1050050	1000060	1000060

Table III
POPULATION TESTS - R2_10_5

	10 (individuals)	25 (individuals)	50 (individuals)
min	200085 (2%)	200082 (2%)	190091 (2%)
avg	226673	221875	216279
max	240080	230087	230067

V. Conclusions and Future Work

This paper presents an application of genetic programming for generation of heuristic functions that guide the creation of

Table I
VEHICLES COUNT FOR 6 TRAINING AND 6 TEST SETS

Group	Train (vehicles)	Best known (vehicles)	Test (vehicles)	Best known (vehicles)	Greedy function
C1	497	482	480	459	2a + c + 7e
C2	159	145	165	144	a * a + (2f + a) * f +e + c - d
R1	492	464	479	455	c + a
R2	104	95	105	95	3e - b - f + d
RC1	486	450	485	450	[(2a + c * c) * e - d + a + e] * c
RC2	114	92	108	90	c + e + a

Table IV
POPULATION TESTS - RC1_10_5

	10 (individuals)	25 (individuals)	50 (individuals)
min	940064 (2%)	930068 (4%)	930065 (2%)
avg	955878	948072	944068
max	990093	960091	960063

initial solutions for the vehicle routing problem. Although the paper presents only preliminary results, this approach shows high potential. Simple and fast greedy functions obtained by GP provide acceptable quality of initial results for further refinement. The presented method can be used as a basis for creating heuristic functions for various performance objectives in the VRP domain, as well as for a diverse set of optimization problems. Moreover, heuristic functions that are evolved beforehand can be used in real-time VRPTW scenarios, where changes in problem parameters may occur unexpectedly (e.g. a time window is changed for a customer, or an additional customer appears). In such cases, the existing search-based algorithms may not be practical since they require a large amount of processing.

Future work will primarily focus on definition and evaluation of additional operators and variables for greedy function creation. Experimenting with different variables could provide significant improvements. Additionally, the quality of the final solution is greatly influenced with the algorithm used in the second phase, where some other metaheuristic can also be applied. Finally, the proposed approach will be suited and evalauted to meet real world problem demands.

REFERENCES

[1] T. Caric, "Improving of transport organization using heuristics methods," *(Croatian) Ph. D thesis, University of Zagreb, Croatia*, 2006.

[2] J. K. Lenstra and A. H. G. R. Kan, "Complexity of vehicle routing and scheduling problems," *Networks*, vol. 11, no. 2, pp. 221–227, Jun. 1981. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/net.3230110211/abstract

[3] R. Bent and P. Hentenryck, "A two-stage hybrid local search for the vehicle routing problem with time windows," *Transportation Science*, vol. 38, no. 4, pp. 515–530, 2004.

[4] ——, "A two-stage hybrid algorithm for pickup and delivery vehicle routing problems with time windows," *Computers & Operations Research*, vol. 33, no. 4, pp. 875–893, 2006.

[5] M. Gulic, D. Lucanin, and N. Skorin-Kapov, "A two-phase vehicle based decomposition algorithm for large-scale capacitated vehicle routing with time windows," *MIPRO, 2012 Proceedings of the 35th Internation Convetion*, pp. 1104–1108.

[6] J. R. Koza, "Genetic programming - on the programming of computers by means of natural selection."

[7] D. Jakobovic and K. Marasovic, "Evolving priority scheduling heuristics with genetic programming," *Applied Soft Computing*, vol. 12, no. 9, pp. 2781–2789, Sep. 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1568494612001780

[8] D. Jakobovic, L. J. c, and L. Budin, "Genetic programming heuristics for multiple machine scheduling," *Lecture Notes in Computer Science*, vol. 4445, pp. 321–330, 2007.

[9] M. Solomon, "Algorithms for the vehicle routing and scheduling problems with time window constraints," *Operations research*, pp. 254–265, 1987.

[10] R. Bent and P. Hentenryck, "Spatial, temporal, and hybrid decompositions for large-scale vehicle routing with time windows," *Principles and Practice of Constraint Programming-CP 2010*, pp. 99–113, 2011.

[11] P. Toth and D. Vigo, *The Vehicle Routing Problem*. SIAM Monographs on Discrete Mathematics ans Applications, SIAM, Philadelphia, 2002.

[12] H. Li and A. Lim, "Local search with annealing-like restarts to solve the VRPTW," *European journal of operational research*, vol. 150, no. 1, pp. 115–127, 2003.

[13] E. Taillard, P. Badeau, M. Gendreau, F. Guertin, and J. Potvin, "A tabu search heuristic for the vehicle routing problem with soft time windows," *Transportation science*, vol. 31, no. 2, pp. 170–186, 1997.

[14] L. Rousseau, M. Gendreau, and G. Pesant, "Using constraint-based operators to solve the vehicle routing problem with time windows," *Journal of Heuristics*, vol. 8, no. 1, pp. 43–58, 2002.

[15] G. Schrimpf, J. Schneider, H. Stamm-Wilbrandt, and G. Dueck, "Record breaking optimization results using the ruin and recreate principle," *Journal of Computational Physics*, vol. 159, no. 2, pp. 139–171, 2000.

[16] D. Mester and O. BrÃd'ysy, "Active guided evolution strategies for large-scale vehicle routing problems with time windows," *Computers & Operations Research*, vol. 32, no. 6, pp. 1593–1614, 2005.

[17] L. Gambardella, E. Taillard, and G. Agazzi, "Macs-vrptw: A multiple colony system for vehicle routing problems with time windows," in *New ideas in optimization*, 1999.

[18] J. R. Koza, "Human-competitive results produced by genetic programming," *Genetic Programming and Evolvable Machines*, vol. 11, pp. 251–284, September 2010. [Online]. Available: http://dx.doi.org/10.1007/s10710-010-9112-3

[19] *ECF - Evolutionary Computation Framework*, 2011. [Online]. Available: http://gp.zemris.fer.hr/ecf

[20] "Extended SOLOMON's VRPTW instances," http://www.fernuni-hagen.de/WINF/touren/inhalte/probinst.htm. [Online]. Available: http://www.fernuni-hagen.de/WINF/touren/inhalte/probinst.htm

[21] M. Deza and E. Deza, *Encyclopedia of distances*. Springer Verlag, 2009.

[22] "VRPTW," http://www.sintef.no/Projectweb/TOP/Problems/VRPTW/. [Online]. Available: http://www.sintef.no/Projectweb/TOP/Problems/VRPTW/

Multimedia stimuli databases usage patterns: a survey report

M. Horvat[1], S. Popović[1] and K. Ćosić[1]
[1] University of Zagreb, Faculty of Electrical Engineering and Computing
Department of Electric Machines, Drives and Automation
Zagreb, Croatia
marko.horvat2@fer.hr

Abstract – Multimedia documents such as images, sounds or videos can be used to elicit emotional responses in exposed human subjects. These stimuli are stored in affective multimedia databases and successfully used for a wide variety of research in affective computing, human-computer interaction and cognitive sciences. Affective multimedia databases are simple repositories of multimedia documents with annotated high-level semantics and affective content. Although important all affective multimedia databases have numerous deficiencies which impair their applicability. To establish a better understanding of how experts use affective multimedia databases an online survey was conducted into the subject. The survey results are statistically significant and indicate that contemporary databases lack stimuli with rich semantic and emotional content. 73.33% of survey participants find the databases lacking at least some important semantic or emotion content. Most of the participants consider stimuli descriptions to be inadequate. Overall, 1-2h or more than 24h are generally needed to construct a single stimulation sequence. Almost 84% of the survey participants would like to use real-life videos in their research. Experts unequivocally recognize the need for an intelligent stimuli retrieval application that would assist them in experimentation. Almost all experts agree such applications could be useful in their work.

I. INTRODUCTION

Emotionally annotated databases such as International Affective Picture System (IAPS) [1], International Affective Digital Sounds (IADS) [2], The Geneva Affective PicturE Database (GAPED) [3] and NimStim Face Stimulus Set [4] are often used in research of emotion processing, attention, stress resilience and mental health, yet not enough has been done to facilitate their usage and expand prevalence in the field. Multimedia documents are loosely annotated which makes semantic retrieval difficult and results in low recall and precision. High-level semantic content descriptors are informal and ambiguous due to the insufficient annotation methods which rely on unrestricted keywords. In this limited meta-retrieval framework no knowledge base, concept taxonomy or terminology exists. These problems contribute to a deficient stimulation, suboptimal elicitation and emphasize the need to thoroughly improve the elicitation process and the structure of multimedia stimuli databases [5] [6].

This research has been partially supported by the Croatian Ministry of Science, Education and Sports.

Currently stimuli from multimedia databases are extracted manually by a laborious visual examination of each stimulus and the accompanying manuals. Stimuli sequence construction is often demanding and time-consuming task.

To complicate their usage even further, multimedia stimuli databases have diverse structures, describe emotional and semantic data differently and contain various media formats and stimuli modalities [5]. There is no consensus among researchers in an optimal format or implementation of a multimedia stimulus database [6]. Because of these reasons a typical multimedia stimuli database user must be a proficient expert in the field of emotion elicitation and simultaneously skilled in technology-related tasks such as stimuli selection and extraction. Such heterogeneous fusion of different skills is difficult to master and since databases' structures are mutually different these skills must be separately trained for each multimedia stimuli database [7].

All these issues outline a need for a stimuli generating computer system which can assist experts in finding the most appropriate stimulus and to do it in the shortest time possible. Such system must be as universal as possible towards various database and media formats, efficient and user-friendly. Furthermore, to assist the expert even further the system should contain empirically derived rules for decision support and automatic generation of stimuli sequences. Also, the stimuli generator should support personalized stimulation individually tailored to specific emotional and semantic parameters.

To address these issues, and as a part of research in methods for formal representation of multimedia stimuli, a software application Intelligent Stimuli Generator (intStimGen) was developed at the Laboratory for Interactive Simulation Systems with the University of Zagreb, Faculty of Electrical Engineering and Computing. The intStimGen (Fig. 1) enables searching of multimedia stimuli databases and construction of stimuli sequences using semantic and emotion descriptors. Stimuli semantics can be described with unrestricted keywords, tag clouds, WordNet synsets [8] and SUMO concepts [9], while emotion descriptors are based on dimensional models. Integrated exploration of IAPS and IADS is supported, but the application's architecture enables other stimuli databases to be modularly added in the future.

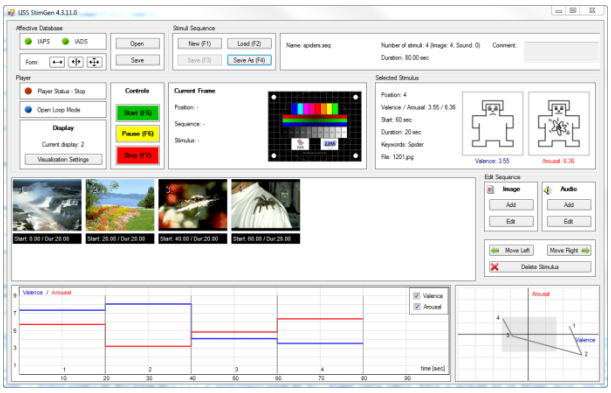


Figure 1. The main screen of the Intelligent Stimuli Generator application displaying a fear-provoking sequence constructed with semantics related to arachnophobia.

The constructed sequences can be displayed to a subject on a separate screen and his voluntary responses may be acquired through a standard human interface device such as keyboard. During exposure the application sends system messages which can be used for synchronization of physiology acquisition equipment and other systems.

However, to assess the need for such software system among the professional community a survey was needed to collect experiences and opinions among experts who use multimedia stimuli in their research.

II. SURVEY METHOD

Between 19 March and 19 April 2012 an invitation to a survey has been sent to 120 e-mail addresses of different researchers, psychologists, psychiatrists, neuroscientists and medical doctors that have been published papers in which they have used at least one emotionally annotated database. The targeted databases primarily were IAPS, IADS[2], NimStim and GAPED. Some invitees have been using Karolinska Directed Emotional Faces (KDEF) [10], Pictures of Facial Affect (POFA) [11] and Japanese Female Facial Expression (JAFFE) Database [12]. Relevant papers were retrieved using search engines Web of Knowledge, PubMed, IEEE Xplore and Google Scholar. The invitees were all outside Croatia (i.e. international) and their affiliations included Max Planck Institute for Human Cognitive and Brain Sciences, Georg-August-University Göttingen, Humboldt-Universität zu Berlin, University of Geneva, University of Gent, University of Chicago, University of Florida, University of Southern California etc. The participants could fill out a survey form in Word format or give their answers online at: http://www.surveymonkey.com/s/L79ZXV2. Out of 120 invitees 30 have completed the survey (25%). Four (4) have filled out the survey form and other 26 have completed the survey online. The participants' answers were sent in period 26 March – 28 August 2012. The answers were individually recorded and archived. The anonymity of the participants was guaranteed as was stated in the e-mail invitation. Only participants' IP addresses, the time of access and responses were recorded. Multiple answers were not allowed. Online participants finished the survey in under 5 minutes, with 2min 40sec on average.

The survey consisted of 10 question with predefined answers. Possible answers in the Likert scale were in a range 1–5 or 1–7. All questions also had "Not sure / Not applicable" ("N/A") option as one of the possible answers. Any question could be left unanswered (blank). Such responses were later processed as "N/A".

The survey questions were:

1. How difficult or easy is the image retrieval process from an emotionally-annotated database (e.g. IAPS, NimStim, IAD), KDEF, POFA, JAFFE, GAPED, etc.)?
2. How satisfied are you with the above mentioned level of difficulty in the image retrieval process from the database?
3. How much time was necessary to effectively search the database and construct one picture sequence that was used in your research?
4. Have you at any time felt that the picture set you were using is missing images with particular emotion or semantic content that would be useful for your stimuli sequence?
5. On a scale of 1 to 7, with one being extremely inadequate and 7 being extremely adequate, please rate how inadequate, ambiguous or insufficient did you find the predefined descriptions of the images you used?
6. On a scale 1 to 7, with one being extremely useless and 7 being extremely useful, please rate how useful, helpful or beneficial would a user-friendly software tool for intelligent retrieval of emotionally-annotated images be to your research?
7. Have you at any time during your research wanted to find the most appropriate emotionally-annotate image faster and more efficiently?
8. Did you construct the sequence manually or with a help of a software tool?
9. Skip this question if you did not use a software tool for retrieval of emotionally-annotated images. Did your group actually develop the tool used in your experiment or acquired it elsewhere?
10. How useful or useless a stimuli database with real-life video-clips, instead of just still images, would be to your work?

III. CUMULATIVE SURVEY RESULTS

The survey results are given in the table and figure below. The results are freely available and can be acquired by contacting the first author of the paper (i.e. the author of the survey).

TABLE I. OVERVIEW OF THE CUMULATIVE SURVEY RESULTS

Question[1]	Answers
Q1	Very difficult = 0%, Difficult = 16.67%, Neither difficult nor easy = 36.67%, Easy =20%, Very easy = 10%, N/A = 16.67%
Q2	Very dissatisfied = 3.33%, Dissatisfied = 13.33%, Neither satisfied nor dissatisfied = 33.33%, Satisfied = 30%, Very satisfied = 6.67%, N/A = 13.33%
Q3	Less than 5 mins = 0%, 5-15 mins = 3.33%, 15-30 mins = 10%, 30-60 mins = 10%, 1-2 hrs = 26.67%, 2-6 hrs = 13.33%, 6-12 hrs = 10%, More than 12 hrs = 20%, N/A = 6.67%
Q4	Yes = 73.33%, No = 20%, N/A = 6.67%
Q5	Value 1 (extremely inadequate) = 0%, Value 2 = 6.67%, Value 3 = 30%, Value 4 (neutral) = 23.33%, Value 5 = 23.33%, Value 6 = 10%, Value 7 (extremely adequate) = 0%, N/A = 6.67%
Q6	Value 1 (extremely useless) = 0%, Value 2 = 0%, Value 3 = 0%, Value 4 (neutral) = 10%, Value 5 = 23.33%, Value 6 = 13.33%, Value 7 (extremely useful) = 46.67%, N/A = 6.67%
Q7	Yes = 86.67%, No = 10%, N/A = 3.33%
Q8	Manually = 46.67%, With a software tool = 23.33%, Both = 23.33%, N/A = 6.67%
Q9	The tool was acquired = 20%, The tool is own development = 20%, N/A = 60%
Q10	Completely useless = 0%, A little bit useful = 3.33%, Neither useful nor useless = 0%, Useful = 20% Very useful = 73.33%, N/A = 3.33%

1. Text of the questions is provided in Section 2.

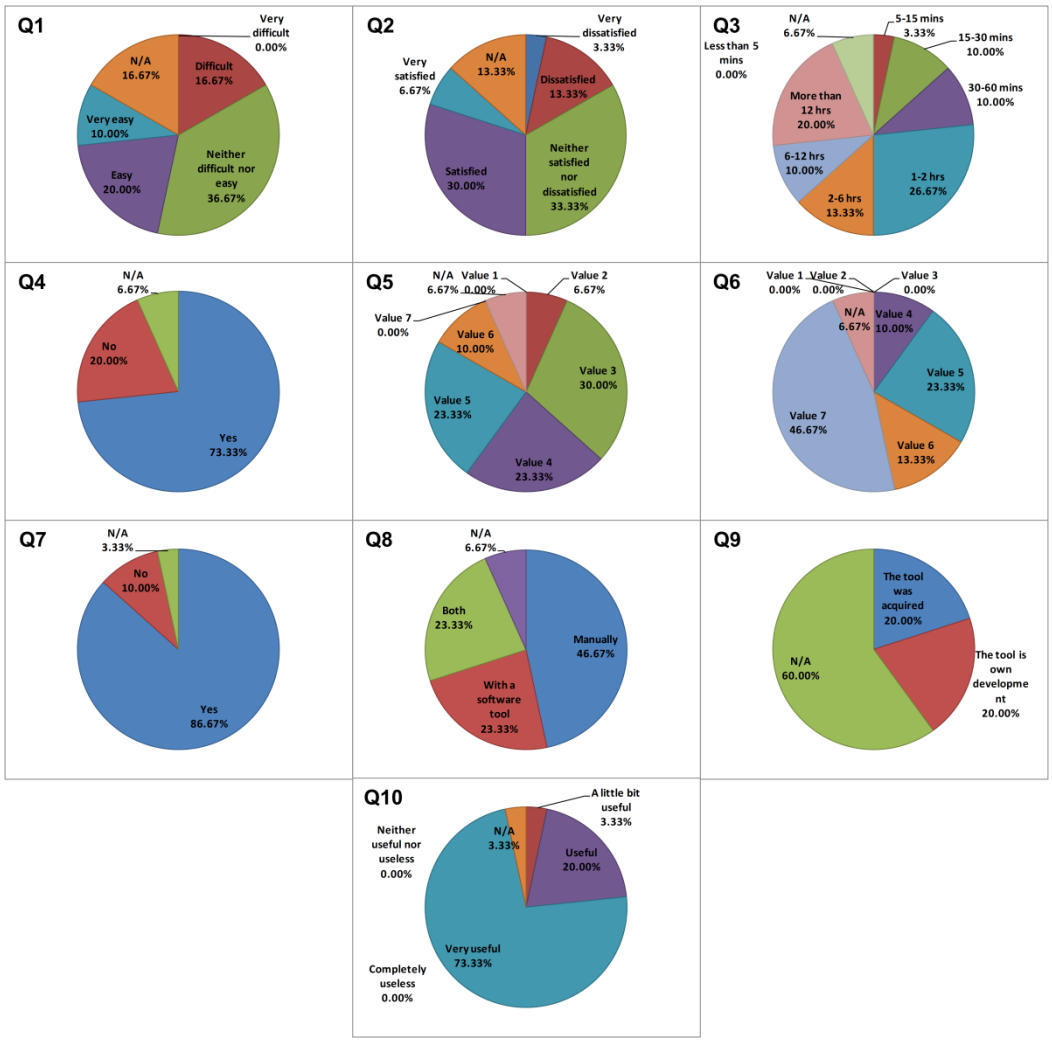


Figure 2. Corresponding charts of the cumulative survey results.

IV. DISCUSSION AND RESULTS ANALYSIS

The survey has clearly indicated that intelligent, efficient and user-friendly generators of emotionally and semantically annotated multimedia stimuli are clearly desired by the professionally community. A great majority of the participating researchers (83.33%) consider such intelligent software tools to be useful or very useful in their work with retrieval of emotionally-annotated images.

Very importantly the results show that stimuli sequences are predominantly constructed by hand – with experts thoroughly scouring through multimedia stimuli– over a period of 1-2 hours (26.67%), 2-6 hours (13.33%) or more than 12 hours (20%). The latter result is particularly alarming because it shows that construction of experiments is a technology-influenced task that may be greatly facilitated by usage of intelligent multimedia retrieval algorithms and procedures. A stimuli construction process that lasts several hours or days is very demanding and labor intensive for the researcher. It also impairs experiment flexibility and repeatability. For comparison, a document retrieval task using the intStimGen tool would last between several milliseconds and seconds depending on the database size and query complexity. In any case this is negligible compared to the typical duration of a manual stimuli sequence construction.

Construction of stimuli sequences is still almost entirely unsupported by intelligent software tools. Almost half of the all researchers explore multimedia stimuli databases only manually (46.67%), by simultaneously looking over multimedia documents and their descriptions. However, 23.33% of researches retrieve stimuli both manually and with a software tool. Some researchers use general purpose tools like Statistica and SPSS for stimuli retrieval, or their teams have developed custom-made software applications for this purpose (20%).

Most researchers are ambiguous towards practical difficulties in construction of stimuli sequences, yet a 86.67% of them would like to accomplish this task even faster and more efficiently. This answer could imply that researchers have not yet used multimedia stimuli management tools but they do recognize that the purely technical side of their work should be somehow streamlined and made easier.

A clear majority of the participants (73.33%) think that current emotionally annotated databases lack at least some stimuli with a particular semantic or emotional content that could be useful in their work. Also, 93.33% of the participants consider real-life video sequences in emotion elicitation procedures to be useful or very useful. Combined with previous answers this once again accentuates the need to expand the content of multimedia stimuli databases and create new databases with rich and personalized emotion-provoking content that provides numerous benefits in the emotion elicitation experiments [13].

Most participants (60%) deem stimuli descriptors to be at least in some way inadequate, ambiguous or insufficient in conveying the true semantic and emotional content of multimedia stimuli. These problems motivate interdisciplinary research into methods for improving annotation of multimedia stimuli. Formal knowledge representation and automated reasoning techniques can be successfully applied to impair deficiencies of stimuli databases and represent the best option for their optimization. The aforementioned intStimGen application was developed with intention to implement these advanced stimuli retrieval methods and improve construction, utilization and proliferation of multimedia stimuli databases.

Finally, it should be mentioned that a number of participants spontaneously expressed the support for the development of intelligent tools for multimedia stimuli retrieval and wished to remain in contact after the survey was completed and evaluate the tools in their research.

V. CONLUSION

Although not large by the complexity of posed questions or by the number of participants, this survey may be considered a success. Aggregated answers follow Gaussian distribution and present statistically significant indications of the target population opinions. Motivation of the experts who have taken part in the survey demonstrates the importance of the issues involved, i.e. current trends in usage patterns of multimedia stimuli databases and intelligent tools for their management.

The survey was initially conceived as a test for validity of the intelligent stimuli sequence generator concept, but apart from succeeding in its verification the survey provided an insight in the needs and preferences of the professional community.

The collected results may be considered as an unmistakable empirical indication of the validity of the intelligent stimuli generator concept and provides an encouragement for continuation of the work in this area.

The author of the survey would like to cordially thank all experts who have taken part in the survey.

REFERENCES

[1] P. J. Lang, M. M. Bradley, B. N. Cuthbert, “International affective picture system (IAPS): Affective ratings of pictures and instruction manual, ” Technical Report A-8, University of Florida, Gainesville, FL, 2008.

[2] P. J. Lang, M. M. Bradley, B. N. Culthbert, “International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings (Tech. Rep. No. B-2),“ The Center for Research in Psychophysiology, University of Florida, U.S.A., 1999.

[3] E. S. Dan-Glauser, K. R. Scherer, “The Geneva Affective PicturE Database (GAPED): A new 730 picture database focusing on valence and normative significance,” Behavior Research Methods, vol. 43(2), pp. 468–77, 2011.

[4] N. Tottenham, J. W. Tanaka, A. C. Leon, T, McCarry, M. Nurse, T. A. Hare, D. J. Marcus, A. Westerlund, B. J. Casey, C. Nelson, “The NimStim set of facial expressions: Judgments from untrained research participants, ” Psychiatry Research, vol. 168:3, pp. 242–249, 2009.

[5] M. Horvat, S. Popović, N. Bogunović, K. Ćosić, “Tagging multimedia stimuli with ontologies,” Proceedings of the 32nd International Convention MIPRO 2009, Croatian Society for Information and Communication Technology, Electronics and Microelectronics – MIPRO, Opatija, Croatia, pp. 203–208, 2009.

[6] M. Horvat, S. Popović, K. Ćosić, “Towards semantic and affective coupling in emotionally annotated databases,” Proceedings of the

35nd International Convention MIPRO 2012 Croatian Society for Information and Communication Technology, Electronics and Microelectronics – MIPRO, Opatija, Croatia, pp. 1003–1008, 2012.

[7] K. Ćosić *et al*., "Virtual reality adaptive stimulation in stress resistance training", In Proceedings RTO-MP-HFM-205 on "Mental Health and Well-Being across the Military Spectrum", Bergen, Norway, April 2011.

[8] C. Fellbaum, "WordNet: An Electronic Lexical Database," MIT Press, 1998.

[9] I. Niles, A. Pease, "Towards a Standard Upper Ontology," In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17–19, 2001.

[10] D. Lundqvist, A. Flykt, A. Öhman, "The Karolinska Directed Emotional Faces – KDEF, " CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, 1998.

[11] The Paul Ekman Group, LLC, retrieved from: https://face.paulekman.com/face/productdetail.aspx?pid=1

[12] M.J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, "Coding Facial Expressions with Gabor Wavelets," In: Proc. Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara Japan, IEEE Computer Society, pp. 200–205, 1998.

[13] K. Ćosić, S. Popović, D. Kukolja, M. Horvat, B. Dropuljić, "Physiology-driven adaptive virtual reality stimulation for prevention and treatment of stress related disorders," CyberPsychology, Behavior, and Social Networking. Vol. 13:1; pp. 73–78, 2010.A. Damasio, "The Feeling of What Happens: Body and Emotion in the Making of Consciousness", William Heinemann, London, 1999.

Navigational Evaluation of Breadth First Search Spanning Trees

Denis Helic*, Markus Strohmaier*†,Weronika Wójcik*
* Knowledge Management Institute
Graz University of Technology
Graz, Austria
Email: {dhelic,markus.strohmaier}@tugraz.at, weronika.wojcik@student.tugraz.at
† Know-Center, Graz University of Technology, Graz, Austria

***Abstract*—Decentralized navigation is one of the most important functions of complex networks. A wide variety of different networks such as communication, social, or information networks possesses certain structural clues which allow navigational agents to efficiently navigate those networks even with local knowledge of the network only. Such structural clues include node degrees and their centralities, similarities between nodes, or node clustering coefficients. In practice, those properties may be combined and abstracted in the form of a distance metric on the network node set – a typical representation of such a distance metric is a hierarchy of network nodes where the most central nodes are situated in the upper levels of the hierarchy and the hierarchy links capture the node similarities or their clustering. Recently, a number of algorithms for extracting such hierarchies have been introduced. The majority of those algorithms is based on complex and computationally intensive methods such as hierarchical clustering. In this paper we analyze several simple spanning tree algorithms and their ability to extract sound hierarchies for network navigation. In particular, we are interested in correlation of structural node properties (such as node degree) and the navigational quality of spanning trees that are rooted at those nodes. Our work sheds light on the ability of spanning trees to serve as a distance metric for decentralized navigation in networks. Our results are relevant for scientists interested in the navigability of complex networks and for engineers who are interested in fast and simple extraction of hierarchies for supporting navigation in networks.**

I. Introduction

Decentralized navigation in social or information networks represents a task where an agent navigates a network with limited knowledge. The agent's goal is to navigate from a starting node s to another target node t without having detailed knowledge about the full network topology. Informally, we say an agent uses *local* knowledge for navigation when the agent has knowledge about the nodes within the immediate one-hop neighborhood of the current location only, but no detailed knowledge of the network beyond that. We refer to these nodes as *candidate* nodes, because they are candidates for navigation. We say an agent uses *background* knowledge for navigation when the agent has some intuitions about the candidate nodes - for example which candidate nodes might bring her closer to a target. In previous research, Kleinberg [1]–[3] and others have introduced a simulation-based navigation framework called *decentralized search* that is formally capable of capturing these and other aspects of navigation.

In this paper, we turn our attention to spanning trees as one particular technique for constructing background knowledge. We define the problem formally, and investigate the usefulness of breadth first search constructed spanning trees for navigation. We prove a few navigational properties of spanning trees theoretically, and expand our analysis by conducting experiments. We conclude that breadth first search spanning trees represent an interesting new option for constructing efficient trees for navigation that can be used as background knowledge in decentralized navigation. However, relation between node properties and the efficiency of their corresponding breadth first search spanning trees is not immediately visible requiring further research that should analyze further algorithms for construction of spanning trees.

II. Related Work

Research on decentralized search in social networks started with Milgram's seminal small world experiment [4] in which he aimed to study the connectedness of the US society. In his experiments he found that people are able to navigate large social networks (such as the population of the USA) efficiently. Specifically, he found that the average number of links between people in his experiment is around 6, hence a "small world". Later, Kleinberg analyzed an implicit result of the Milgram's experiment, the ability of humans to *find a short path* when there is such a path between two nodes [1]–[3]. Kleinberg concluded that social networks possess certain latent properties that humans are aware of. This *background knowledge* of network structure allows humans to find a short path between two arbitrary network nodes efficiently. Kleinberg defined an "efficiently" navigable network as a network for which a *decentralized search algorithm* exists, such that its delivery time (the number of nodes that the algorithm needs to visit before it reaches the destination node) is polynomial in $log(n)$, where n is the number of nodes in the network.

Subsequent work has investigated the nature of background knowledge that is required for efficient decentralized search algorithms. In other words: What structural properties do efficiently navigable networks possess? To that end, Kleinberg designed a number of network models such as the 2D-grid model [1], hierarchical model [3], and group model [3]. Inde-

pendently, Watts [5] introduced the notion of social identity as a membership in a number of social groups organized in hierarchies and showed the existence of efficient decentralized search algorithms by simulation.

Decentralized search has been used to model navigation in social and information networks. For example, Adamic and Adar [6] have used organizational hierarchies as background knowledge to model navigation in e-mail networks. In our own previous work, we have used decentralized search to evaluate the usefulness of topical hierarchies for navigating information networks such as social tagging systems [7], [8]. Decentralized search has also been used to model navigational trails left by users on Wikipedia [9]. However, little is known about the nature and impact of background knowledge on decentralized search. For example, we don't know how different kinds of background knowledge influence the ability of a decentralized search model to efficiently navigate networks with local information. This issue is what we want to explore in this work. In particular, we turn our attention to the efficiency of spanning trees as background knowledge for decentralized search.

III. Theoretical Framework

A. Preliminaries

Let us first define some basic terms that we use throughout the paper. $G(V, E)$ is a network with nodes from a non-empty set V and links from set E. We denote the number of nodes in a network with n and the number of links with m. In this paper we focus on the analysis of undirected networks – the analysis of the directed networks is more complicated and we leave it for the future work. A *path* in a network is a sequence of different nodes that are connected by links – it is a *route* through the network. The path length is the number of links in that path. A *shortest path* between nodes u and v is a path between u and v with the minimal length. We define distance $d(u, v)$ between u and v as the length of the shortest path connecting u and v. We call a network connected if there is a path from every node to every other node in the network. A *cycle* is a path that starts and ends at the same node. A *tree* is a connected network that has no cycles. It is easy to show that a tree with n nodes has exactly $n - 1$ links. A network can also consist of a number of disconnected parts, all of which contain no cycles. In that case the individual parts are all trees. We call then the complete network a *forest*.

A *spanning tree* of a connected graph G is a tree that spans every node from the network. A special type of a spanning tree is a *shortest path spanning tree rooted at node* r, which is a spanning tree such that for any node u, the distance between r and u is the same as in the network.

In this paper we will denote the distance between u and v in the network with $d^G(u, v)$ and their distance in a forest with $d^F(u, v)$.

B. Global vs. local knowledge

Finding a shortest path between any two nodes in a network is easy if we know about all the nodes and links; that is if we have *global knowledge of the network*. Asymptotically, global knowledge is of size $O(n^2)$, since we might have a link between any two nodes to denote distances in the network. In practice, we often do not possess such global knowledge of the network, but *local knowledge of the network* only. For instance, in a social network we typically know about our one hop neighborhood; that is we know about our friends, but not about the friends of their friends, and so on. Analogously, in an information network such as Wikipedia, we can see only the links emanating from the page that we are currently viewing. That means that in practice, our knowledge about the network structure is much smaller than $O(n^2)$ – we only have intuitions about parts of the network.

Yet, people are still able to find shortest paths in social networks even with such limited knowledge. For example, in the famous "small world" experiment by Milgram [4], randomly selected persons from Nebraska and Massachusetts were required to pass a letter to a target person in Boston, Massachusetts in a decentralized manner through their social contacts, thereby requiring a group of people to search in a very large social network by utilizing local knowledge only. One of the most important results of the experiment was the finding that people are able to efficiently find short chains from the network in a *decentralized* manner. For example, Kleinberg concluded that humans possess *background knowledge* of the network structure and that this background knowledge allows humans to efficiently find short paths [1]–[3]. Kleinberg represented such background knowledge as a hierarchy (tree) of nodes, where more similar nodes are situated closer to each other in the tree. Thus, people consult the tree and base their navigation decisions on this background knowledge, which is encoded in a tree.

It is important to note that the size of background knowledge (represented as a tree) is considerably smaller than the size of global knowledge. Specifically, there are $n-1$ links in the tree, which means asymptotically the size of knowledge encoded in the tree is $O(n)$ as compared to the $O(n^2)$ links encoded in global knowledge of the network.

Based on this difference between the size of background knowledge vs. global knowledge, an interesting research question to ask is the extent to which different kinds of background knowledge exhibit different kinds of navigational properties. In other words, how do different trees or forests differ with regard to their ability to encode the knowledge about a network for navigational purposes? To answer this question, we next formalize some of the terms that are needed for analysis.

C. Navigation

Definition 1: **Navigation (routing)**. Let $G(V_G, E_G)$ be a connected undirected network, and let $F(V_F, E_F)$ be a forest such that $V_F = V_G$, and $|E_F| \leq n - 1$. Let s be a start node and t a target node. A navigation path between s and t is defined as a path $(s, u_1, u_2, ..., u_{k-2}, u_{k-1}, t)$ from G such that $d^F(u_i, t) > d^F(u_j, t)$ for $i < j$.

With that, we can say that in a navigation path, the subsequent node in a path has a smaller *forest distance* to the target node than its predecessor. We call a network G *navigationally*

connected with respect to a forest F (that is the forest F provides the distances for the navigational paths in G) *if and only if* there exists a navigation path from every node to every other node in the network. This allows us to connect the structural description of a network and a forest in the form of navigational paths with the dynamics of an agent navigating the network. Thus, at each navigation step towards a target node, an agent consults the forest about the distances of the candidate nodes to the target node and selects a node with a smaller distance to the target node than that of the current node. We denote the current node with u, and the set of its neighbors; that is the set of the candidate nodes with $\Gamma(u)$.

Next, we demonstrate a number of ways in which forests relate to their corresponding networks from a navigational perspective.

Theorem 1: Let G be a connected undirected network, and let F be a forest such that $V_F = V_G$, and $|E_F| \leq n-1$. Then:

(a) If F is disconnected then G is navigationally disconnected with respect to F.
(b) If F is connected and $E_F \backslash E_G \neq \emptyset$ then it is possible that G is navigationally disconnected with respect to F.
(c) If F is connected, and F is a spanning tree of G; that is $E_F \subseteq E_G$ then G is navigationally connected with respect to F.

Proof:

(a) Since F is disconnected, there are at least two trees in F. Let us denote these two trees with T_1 and T_2, respectively. Let $t_1 \in T_1$ and $t_2 \in T_2$. Then by definition of the length of the shortest path $d^F(t_1, t_2) = \infty$, and thus we can not create a navigation path between those two nodes with decreasing distances.
(b) We need to construct one configuration without a navigation path between two nodes u and v. Let t be a node with a single link (v, t) in G and a single link (u, t) in F. In that case $(u, t) \in E_F \backslash E_G$ with $d^F(u, t) = 1$. There is no link between v and t in the forest, and thus $d^F(v, t) \geq 2$. Now suppose also that there is a link (u, v) between u and v in G. A path from u to t in the network must go through v and so we have (u, v, t) with the forest distances $(1, \geq 2, 0)$ (see also Figure 1).
(c) Since F is a spanning tree of G, we have a path in F from every node s to every other node t that is also a path in the network. Each subsequent node from this path has a smaller distance to target node t, and the forest distance is decreased by 1. Thus, $d^F(u_i, t) > d^F(u_j, t)$ holds for every $i < j$ and G is navigationally connected with respect to F (see also Figure 2).

■

The concept of navigational connectivity is closely related to the success rate of an agent navigating a network with the forest as the background knowledge. The success rate captures the extent to which an agent is successful in finding paths between two nodes with limited (local & background) knowledge only.

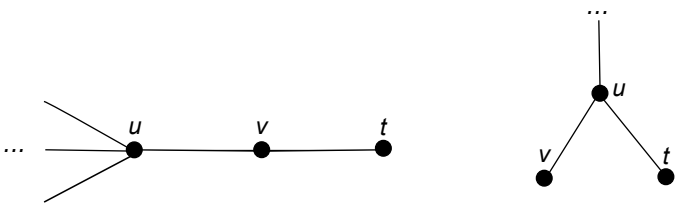
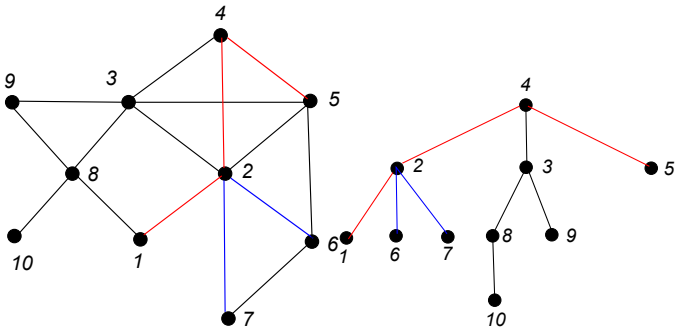


Fig. 1. A particular configuration of G (left) and F (right), which is navigationally disconnected. There exist links in F that do not exist in G – the link between u and t. The path from u to t passes through v – the distances in F are not monotonously decreasing: $(1, 2, 0)$.



(a) Sample network (b) Shortest path spanning tree rooted at 4

Fig. 2. Paths from 1 to 5 (red) and 6 and 7 (blue) exist both in the network (left) and in the spanning tree (right). The distance $d^F(u, v)$ along the paths is monotonously decreasing. Thus, the network is navigationally connected with respect to the spanning tree.

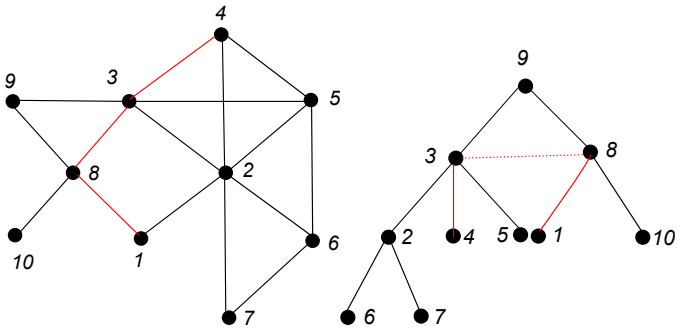
Thus, if a network G is navigationally connected with respect to the forest F, then the navigation agent will be able to find any target node starting from any other start node. However, if the network is not navigationally connected then for certain configurations of the network G and forest F the success rate will drop as for some combinations of start and target node the navigational path does not exist. This might happen in cases where F contains links that do not exist in G – those links might be exogenous to the network – for instance, two nodes are similar in some external properties but they do not share a link in the network.

D. Greedy navigation

Next, we are interested in the efficiency of the navigational agent, or in other words we are interested in the length of the navigational paths.

Definition 2: **Greedy navigation**. Let $G(V_G, E_G)$ be a connected undirected network, and let $F(V_F, E_F)$ be a forest such that $V_F = V_G$, and $|E_F| \leq n-1$. Let s be a start node and t a target node. A greedy navigation path is a navigation path between s and t defined as a path $(s, u_1, u_2, ..., u_{k-2}, u_{k-1}, t)$ from G such that $u_{i+1} \in M := \text{argmin}_{u_j \in \Gamma(u_i)} d^F(u_j, t)$ is the argument of the minimum of the distance function of candidate nodes at node u_i.

Thus, a greedy navigation path is a path where each node in the path has the smallest distance to the target node among all candidate nodes at that position. Dynamically, for each node u a navigation agent selects an adjacent node v which has



(a) Sample network (b) Shortest path spanning tree rooted at 9

Fig. 3. Path from 4 to 1 (red) is a greedy navigational path. There is a leap in the distance reduction along the path (the dashed link from 3 to 8) – this link exists in the network but not in the tree – this is an example of the greedy navigator taking a shortcut in the network. The greedy navigational path $(4, 3, 8, 1)$ is longer than the shortest path in the network $(4, 2, 1)$.

the smallest tree distance to the target node t. In cases where $|M| > 1$ we have two or more candidate nodes that all have the minimal distance to the target node we rely on a total order defined over the node set V. For example, we might count nodes and use $<$ relation over the integer representation of the nodes to select a node with the minimal encoding.

We denote the length of a greedy navigation path with $h(s,t)$ – it is the number of links in a greedy navigator path between s and t. Note that depending on the particular structure of the forest, a greedy navigational path might take shortcuts (larger leaps than 1) in decreasing the distance in a forest. On the other hand, a greedy navigation path is not necessarily a shortest path in G, as can be seen in Figure 3.

To measure the navigational efficiency of a forest F, we can compare $h(s,t)$ to $d^G(s,t)$ for every s and every t. We introduce stretch $\tau(s,t)$ as the ratio between $h(s,t)$ to $d^G(s,t)$:

$$\tau(s,t) = \frac{h(s,t)}{d^G(s,t)}, s \neq t. \quad (1)$$

Stretch $\tau(s,t)$ is equal to one if the greedy navigation path is equal to the length of the shortest path between s and t in G. Otherwise if the greedy navigation path is longer than the shortest path, we obtain stretch values larger than one. For example, a stretch value of 2 and $d^G(s,t) = 2$ means that the greedy navigation path between s and t is twice as long as the global shortest path – it is 4. To assess the navigational efficiency of a spanning tree, we will analyze the stretch in more details.

We also introduce the global stretch τ. For a navigationally connected network G with respect to a spanning tree F, we calculate the global stretch:

$$\tau = \frac{1}{n(n-1)} \sum_{s \neq t} \frac{h(s,t)}{d^G(s,t)}. \quad (2)$$

E. Navigation efficiency

Next, we turn our attention to the analysis of the navigational properties of different spanning trees. In this paper we focus on shortest path spanning trees, which can be obtained, for example, by performing breadth first search. Such spanning trees are rooted at the node from which we start breadth first search and span a tree with distances from the root node to every other node being the same as its distances to every other node in the network. It is interesting to investigate the relation between node properties in the network and the navigational efficiency of spanning trees rooted at those nodes. This would allow us to better identify spanning trees with useful navigational properties.

First, we construct a shortest path spanning tree rooted at a node v. Obviously, the efficiency of greedy navigation from any starting node s towards v is maximal, since the tree contains shortest paths to v from every other node. In other words, the stretch $\tau(s,v)$ for all s is equal to 1.

However, we are interested in exploring what happens if we navigate towards arbitrary nodes, not only towards v. Can we measure stretch induced by a shortest path spanning tree rooted at v over all search pairs by taking arbitrary s and t as the starting and target node respectively?

Let us start our analysis with a conservative estimation (upper bound) of the stretch of a given shortest path spanning tree rooted at v. For a starting node s and a target node t we can always take a path towards v first and then after we reach v we can navigate from v towards t. The total number of steps is:

$$d^G(s,v) + d^G(v,t). \quad (3)$$

If we iterate over all starting nodes fixing t as the target node, the total distance that the navigator travels is given by:

$$\sum_s (d^G(s,v) + d^G(v,t)) = \sum_s d^G(s,v) + n d^G(v,t). \quad (4)$$

If we now iterate over all target nodes t we obtain:

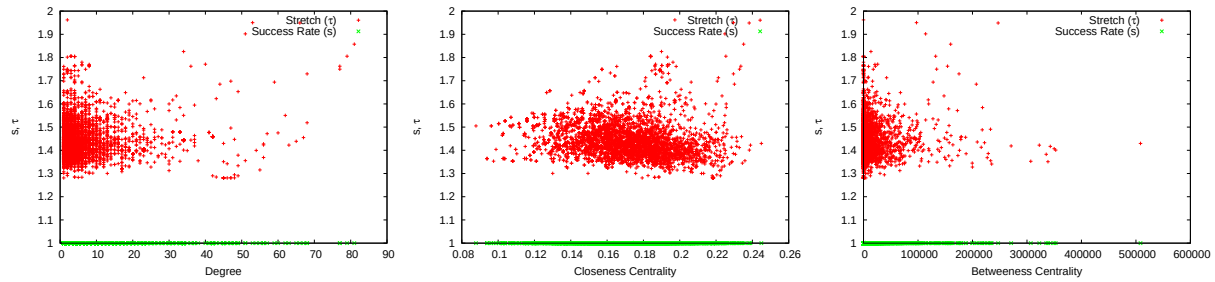
$$\sum_t (\sum_s d^G(s,v) + n d^G(v,t)) = 2n \sum_s d^G(s,v). \quad (5)$$

The estimation is proportional to the sum of shortest paths of node v. For nodes which are on average "closer" to all other nodes in the network this sum will be smaller. In fact, the sum is reciprocal of the node closeness centrality – for nodes with high closeness centrality we can expect the sum of the length of navigational paths in the network to be small.

Following the same argument we can also calculate the estimation (upper bound) of the global stretch of a shortest path spanning tree rooted at v:

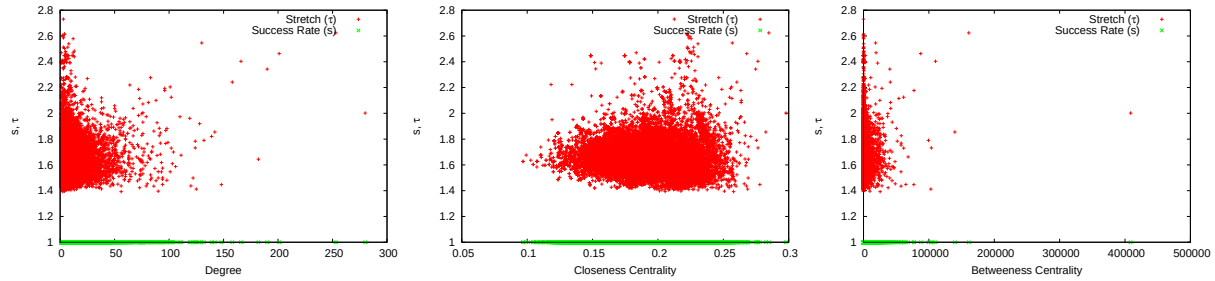
$$\tau = \frac{1}{n(n-1)} \sum_{s \neq t} \frac{d^G(s,v) + d^G(v,t)}{d^G(s,t)}. \quad (6)$$

The global stretch will become smaller for node v which lies frequently on a shortest path between other nodes in the network. In that case we have: $d^G(s,v) + d^G(v,t) = d^G(s,t)$ and therefore $\tau(s,t) = 1$. Therefore, we can expect that nodes with high betweeness centrality produce spanning trees that are more efficient than the nodes with lower betweeness centrality.



(a) Degree (b) Closeness centrality (c) Betweeness centrality

Fig. 4. Success rate (green) and stretch (red) for each shortest path spanning tree of the GR-QC network. Our searches are always producing a successful result with the stretch between 1.2 and 2, however it is impossible to notice, whether is it more advantageous, to choose shortest path spanning tree starting at the root with a greater or with a lower centrality property.



(a) Degree (b) Closeness centrality (c) Betweeness centrality

Fig. 5. Success rate (green) and stretch (red) for each shortest path spanning tree of the GR-CondMat network. Again, we observe maximal success rates (as expected from the theoretical analysis) with the stretch between 1.4 and 2.8. Consistent with the experiments with the GR-QC dataset we can not deduce a simple relation between stretch and node centrality properties.

IV. Experiments

We perform our experiments on two datasets: GR-QC (General Relativity and Quantum Cosmology)[1] and GR-CondMat (Condense Matter Physics)[2] collaboration network. The first dataset covers scientific collaborations between authors of papers submitted to General Relativity and Quantum Cosmology category, whereas the second dataset includes the collaborations between authors publishing in the field of Condense Matter. If an author i co-authored a paper with author j, the graph contains an undirected edge from i to j. If the paper is co-authored by k authors, this generates a completely connected (sub)graph on k nodes. The data covers papers in the period from January 1993 to April 2003.

For our purposes, we choose the largest connected component of the aforementioned networks containing 4158 nodes and 26850 links (GR-QC dataset), and 21363 nodes and 182628 links (GR-CondMat dataset). From the chosen component, we create shortest path spanning trees using breadth first search, where each spanning tree has the root node in a different node of the component. For evaluation, we randomly select 1000 pairs of nodes and simulate navigation using decentralized search. Each of those pairs represents a start node and a target node for a greedy navigation path.

For every tree, we perform greedy navigation on the selected pairs. Search is considered successful *iff* a navigation agent finds a navigation path between the start and the target node. Moreover for each root, we calculate degree, closeness and betweenness centrality. This allows us to observe the extent to which stretch changes according to the aforementioned properties of a node. Figures 4 and 5 depict the relationship between stretch and the various properties of the root.

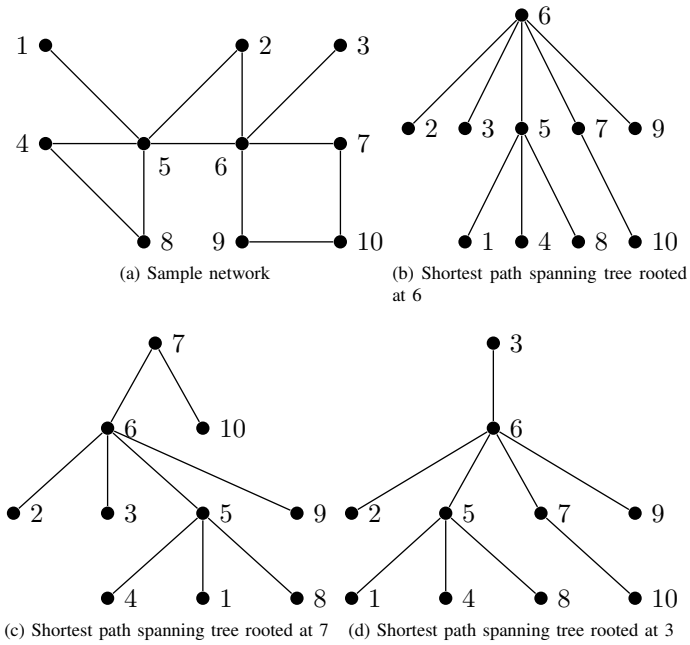
[1]http://snap.stanford.edu/data/ca-GrQc.html
[2]http://snap.stanford.edu/data/ca-CondMat.html

V. Results and Discussion

Figures 4 and 5 show that for every shortest spanning tree, our navigation agent is able to find a navigation path between the start and the target node - as expected from our theoretical discussion of the problem. The stretch for both datasets is low: for GR-QC it is less than 2, and for GR-CondMat less than 2.8.

When we focus on just one property of the root, for example degree, we cannot deduce whether the stretch is going to be better for a greater or for a lower degree of the root of a shortest path spanning tree. A simple explanation for this somewhat unexpected behavior might be easily found through a closer investigation of the structure of the spanning trees constructed by breadth first search.

Breadth first search algorithm produces in many cases structurally similar or even identical shortest path spanning trees regardless of the node from which we start the search. For example, suppose that we start breadth first search at a so-called hub i; that is a node which has a lot of links to other nodes. Now, suppose that among the nodes that are linked to that hub i, we have a node j with a single link (i, j).



(a) Sample network

(b) Shortest path spanning tree rooted at 6

(c) Shortest path spanning tree rooted at 7

(d) Shortest path spanning tree rooted at 3

Fig. 6. A sample network and three breadth first search spanning trees rooted at nodes 6, 7, and 3. Although these nodes have different degrees their corresponding spanning trees are identical. Therefore, the trees have identical navigational properties, and we are not able to relate the navigational efficiency of the trees with the structural properties of nodes at which the trees are rooted.

Obviously, starting breadth first search at node j produces an identical spanning tree to the spanning tree produced by breadth first search started at node i. In the first step of the algorithm we immediately reach node i and proceed from there in the same way as previously. For an illustration of this situation please consult Figure 6.

The same phenomenon can be observed when concentrating on closeness or betweenness centrality. We expected that trees with higher betweenness centrality of the root should have lower stretch. However with breadth first search two shortest path spanning trees constructed from roots with different betweenness centralities, can be identical and therefore can have the same stretch.

The estimations of the upper bound on stretch are in the range of 1.4 to 4.0 for the GR-QC dataset. In our experiments, the estimations turn out very precise. The stretch measured by simulation is smaller than the conservative estimation in 99.8% of cases. The small amount of cases where stretch is greater than the estimated upper bound is due to a small sample size. We sample 1,000 search pairs out of 17,000,000 possible search pairs.

VI. Conclusions

In this work, we analyzed spanning trees and their usefulness to act as background knowledge for informing decentralized search in networks. We found that the navigational quality of spanning trees rooted at certain nodes is mostly independent from the node's centrality in the network. On one hand, this can be explained by the fact that different roots can produce similar or identical spanning trees. On the other hand, we found - through inspection - that even if different spanning trees are produced, they do not necessarily differ significantly. For example, the majority of the structure of a spanning tree rooted at a high degree node can be similar to the structure of a spanning tree rooted at a low degree node. These results are interesting as they suggest that spanning trees rooted at arbitrary nodes in a network can be used efficiently as background knowledge for decentralized search. Our results are relevant for designers of user interfaces aiming to build hierarchical structures that aid users in navigating information networks (such as Wikipedia) or for organizational researchers interested in identifying organizational hierarchies that optimize navigability of social networks within organizations. For future research, we consider an expansion of our analysis to directed networks, or the investigation of other spanning tree algorithms (such as depth-first-search), as interesting routes to take. Another promising approach would be to construct spanning trees with a small, or even minimal diameter [10]. The diameter of a graph is the longest shortest path in a network, and as such it is a structural upper bound on the length of the navigational paths. One can expect a tree with a shorter diameter to be more navigationally efficient than a tree with a longer diameter.

Acknowledgment

This research was in part funded by the FWF Austrian Science Fund research project "Navigability of Decentralized Information Networks" (P 24866-N15).

References

[1] J. Kleinberg, "The small-world phenomenon: an algorithm perspective," in *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, ser. STOC '00. New York, NY, USA: ACM, 2000, pp. 163–170.

[2] J. M. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, no. 6798, p. 845, August 2000.

[3] J. Kleinberg, "Small-world phenomena and the dynamics of information," in *Advances in Neural Information Processing Systems (NIPS) 14*. Cambridge, MA, USA: MIT Press, 2001, p. 2001.

[4] S. Milgram, "The small world problem," *Psychology Today*, vol. 1, pp. 60–67, 1967.

[5] D. J. Watts, P. S. Dodds, and M. E. J. Newman, "Identity and search in social networks," *Science*, vol. 296, pp. 1302–1305, 2002.

[6] L. Adamic and E. Adar, "How to search a social network," *Social Networks*, vol. 27, no. 3, pp. 187 – 203, 2005.

[7] D. Helic, M. Strohmaier, C. Trattner, M. Muhr, and K. Lerman, "Pragmatic evaluation of folksonomies," in *Proceedings of the 20th international conference on World wide web*, ser. WWW '11. New York, NY, USA: ACM, 2011, pp. 417–426.

[8] M. Strohmaier, D. Helic, D. Benz, C. Körner, and R. Kern, "Evaluation of folksonomy induction algorithms," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 74:1–74:22, Sep. 2012.

[9] C. Trattner, P. Singer, D. Helic, and M. Strohmaier, "Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks," in *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, ser. i-KNOW '12. New York, NY, USA: ACM, 2012, pp. 14:1–14:8.

[10] R. Hassin and A. Tamir, "On the minimum diameter spanning tree problem," *Information Processing Letters*, vol. 53, no. 2, pp. 109 – 111, 1995.

Comparison of the Nodes of Knowledge method with other graphical methods for knowledge representation

Alen Jakupović*, Mile Pavlić**, Ana Meštrović** and Vladan Jovanović***
* Polytechnic of Rijeka, Business Department, Rijeka, Croatia
alen.jakupovic@veleri.hr
** University of Rijeka, Department of Informatics, Rijeka, Croatia
mile.pavlic@ris.hr
ana.mestrovic@inf.uniri.hr
*** Georgia Southern University, School of Information Technology, Statesboro GA, USA
vladan@georgiasouthern.edu

Summary – One of the research paths in the field of artificial intelligence is knowledge representation. There are different approaches, formalisms, methods and languages. They vary from simple to complex and from less semantically rich to very expressive. In their previous papers, the authors introduced a new method for knowledge representation named Nodes of Knowledge (NOK), bearing the idea that it should be simple but semantically rich. This article presents a brief example of the basic concepts in the NOK method and its comparison with the following methods: Basic Conceptual Graphs, Multi-layered extended semantic networks, Hierarchical Semantic Form and Resource Description Framework. All these methods belong to the same class as the NOK method - graphical methods for knowledge representation.

I. INTRODUCTION

Knowledge representation and reasoning is a central field within the research of artificial intelligence. It is concerned with how knowledge can be represented symbolically and manipulated in an automated way by reasoning programs [1] and it involves machine-interpretable representation of the world [2].

During the time, many different approaches, formalisms, methods and languages for knowledge representation have been developed. The most influential and the most important approaches are logic schemas, network schemas, frames and rules. In the last two decades, ontologies play an important role in the knowledge representation domain.

Different applications need different formalisms and languages that may vary from simple to complex and very expressive knowledge representation formalisms. More expressive (semantically rich) knowledge representation formalisms are easier to use and to understand on the semantic level. The disadvantage of semantically rich formalisms is a more complex inference algorithm. The trade-off between expressiveness and efficiency has to be considered while choosing or creating knowledge representation formalisms. Therefore, drawbacks of existing formalisms and languages are either poor expressiveness or high complexity and inefficiency.

The authors are introduced a new method for knowledge representation named Nodes of Knowledge (NOK). Its main features are simplicity, the ability of displaying many different types of human knowledge, the ability of automatic detection of new knowledge from existing knowledge, the ability of guided input of new knowledge and the ability of simple creation of user queries over the knowledge base.

An example of knowledge represented with the basic concepts in the NOK method is given below. Subsequently, the NOK method is compared with the following graphical methods for knowledge representation: Basic Conceptual Graphs, Multi-layered extended semantic networks, Hierarchical Semantic Form and Resource Description Framework.

II. NODES OF KNOWLEDGE METHOD

NOK uses two elements for graphical representation: nodes and links. Different kinds of nodes are used for representation of terms. Links between nodes enable grouping of terms into more complex expressions. Specially, a process node is introduced as an aggregation point for representing knowledge described in the sentences. Further, an array of interconnected process nodes can represent knowledge expressed in a sequence of sentences.

The goal of NOK is to represent a knowledge network of text-based knowledge. Furthermore, the NOK method can capture different types of knowledge: knowledge from dictionaries and encyclopaedias, knowledge from existing databases, knowledge embedded in business processes, knowledge stored in business documents, etc. Figure 1 demonstrates an example of graphical knowledge representation using the NOK method. The following types of knowledge are presented:

- according to the contextual node “studies”, the statement “The student studies mathematics” is true

- according to the process node “studies”, the statement “Marko studies the definition of natural numbers” is true, but here we also have a connection with the process node “takes”, which gives the answer to the question: “Why does Marko study?”. Therefore, the full statement is “Marko studies the definition of natural numbers because he takes”. According to the process node “takes”, the statement “Marko takes the exam” is true. Therefore, combined with the former statement finally we get “Marko studies the definition of natural numbers because Marko takes the exam.”

- The existence of three contextual connections demonstrate that the following statements are true: “Marko is a student”, “The definition of natural numbers belongs to the field of mathematics”, “The studying of Marko can be classified as studying of a college student” and “Studying of the definition of natural numbers can be categorized as studying of mathematics”.

Finally, the graphical knowledge representation on picture 1 could be translated to the following statements:

“A student studies mathematics.”

“Marko is a student.”

“Marko studies the definition of natural numbers because Marko takes the exam.”

“The studying of Marko can be classified as studying of a college student.”

“Studying of the definition of natural numbers can be categorized as studying of mathematics.”

“The definition of natural numbers belongs to the field of mathematics.”

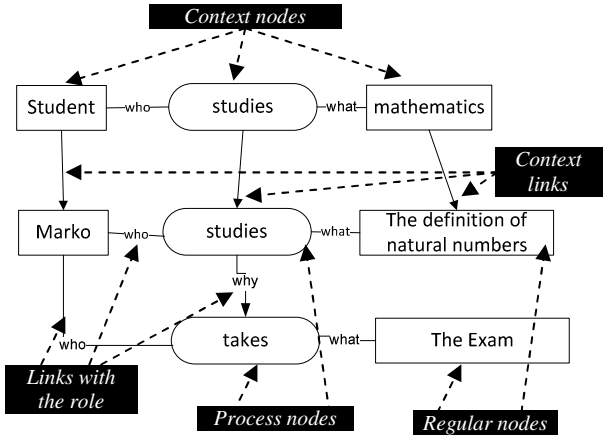


Figure 1. Example of graphical knowledge representation using the NOK method

III. COMPARISON OF THE NOK METHOD WITH OTHER METHODS FOR GRAPHICAL KNOWLEDGE REPRESENTATION

NOK is a knowledge representation method that belongs to the group of semantic networks. In this chapter, the following methods will be compared with the NOK method: BG (Basic Conceptual Graphs) [3], MULTINET (Multi-layered extended semantic networks) [4], HSF (Hierarchical Semantic Form) [5] an RDF (Resource Description Framework) [6]. All methods can graphically represent knowledge, and the resulting model consists of nodes (used to represent concepts) and links between them (used to represent relations between concepts). Various models depicting the following statement will serve as illustrative examples of differences between methods: “Peter finished the discussion”. Figure 2 represents the statement modelled with the NOK method.

Figure 2. Representation of the statement "Peter finished the discussion." using the NOK method

A. *NOK vs BG*

In the BG method, nodes are divided into conceptual and relational and they are linked with connectors. Each node has its name and it is graphically represented with a rectangle (if we are talking about a conceptual node), or an ellipse (if we are talking about a relational node).

A relational node connects conceptual nodes and defines the direction in which the model is to be read. The connection between them has a number that determines the order of conceptual nodes in a relation, which is important to know while reading knowledge. In order to represent the same knowledge (e.g. A is close to B) in reverse order it is necessary to introduce a new relational node with a reverse order of numbers. Solely based on the knowledge contained in the statement “Marko is working on a project”, it is not possible to conclude that “On a project Marko is working”, without specifying it separately.

In the NOK method, the process node is analogous to the relational node of the BG method. Similar to the BG method, a process node connects conceptual nodes. The NOK method, unlike the BG method, does not have number on connections. Instead of numbers, each connection has a role – the name of the role is a simple question. Thanks to this feature, reading is not associated with the order of appearance of a concept node in the relation, but it can be started from any node, but with the application of the corresponding link role, i.e. question.

A very important concept that characterizes the NOK method and can’t be found in the BG method is the connection between process nodes. This feature makes it possible to represent the execution order of process nodes, their logical moves, their conditioning etc.

The BG method has a concept for the representation of generalization/specialization (arrow) between nodes. However, generalization/specialization can be also specified in the name of a specialized node by indicating the name of the generalized node to which it belongs. NOK enables grouping or generalizations through contextual nodes and their connections. BG allows subnets, so that a group of nodes is declared a new node (an aggregation of nodes) and is connected to other nodes. In NOK there are no subnets as aggregations of nodes and connections are made by connecting any

nodes. Knowledge about relationships on a higher level is achieved by connecting context nodes.

Figure 3 represents the statement “Peter finished the discussion”, modeled using the BG method.

Figure 3. Representation of the statement "Peter finished the discussion." using the BG method

Numbers on connections reveal the order of conceptual nodes in the relation “Finish”. Based on this model, the following statements can be read: “Peter finish ?”, “Peter finish discussion.”, “? finish discussion” and “? finish ?”. It is not possible to read statements such as “Peter ?”, “? discussion” and “Peter ? discussion.”. Based on the model that was made using the NOK method (Figure 2), it is possible to read statements such as these and statements that were derived from the model. Therefore, the set of statements that can be read from the NOK-based model is a superset of statements that can be made from the BG-based model.

B. NOK vs MULTINET

In the MULTINET method, each node belongs to a class from a predefined classification of nodes called *conceptual ontology* (a collection of 29 classes). In the NOK method the number of classifications is not limited and the user can add them freely based on needs and field of use.

Each node has 7 predefined attributes, the values of which position it in the semantic space (this approach is motivated by the analogy with a point in the Euclidian space). NOK enables free addition of attributes and only the identifying attribute is necessary.

Connections between nodes are made using one of 89 predefined connection types. 16 of them have the so-called cognitive function and those connections describe the relation between the major participants in a given situation. In the NOK method, the connections are established using special process nodes that can be predefined as context or new nodes used to describe current knowledge.

The method does not graphically distinguish between node types as the NOK method does. The MULTINET method is extremely complex due to a large number of elements that need to be defined for each node (it has be assigned to one of 29 classes, values of 7 attributes have to be assigned and while connecting it, it is necessary to pick an adequate connection type among 89 categories).

Figure 4 represents the statement “Peter finished the discussion.” that is modelled using the MULTINET method.

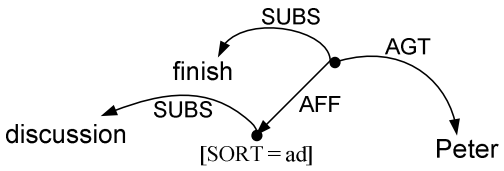


Figure 4. Representation of the statement "Peter finished the discussion." using the MULTINET method [4]

The nodes Peter, finish and discussion are visible on the model, and they belong to one of the classes of the predefined conceptual ontology. The model does not show the values of 7 attributes that belong to each node, but the connections (arrows) are visible, as well as their type. The connections AGT and AFF have cognitive roles and stand for: *agent* and *affected*. SUBS is a connection that indicates that situational concepts *finish* and *discussion* are subordinate to their related general concepts. [SORT = ad] points out the fact that the agent Peter influences the entity that belongs to the group of abstract objects (strictly speaking, to the group of dynamic abstractions).

Based on the model that is made using the MULTINET method it is possible to create all statements that can also be created using the NOK method. However, it needs to be pointed out that the MULTINET method is an extremely complex method that is primarily used to model aspects of natural languages.

C. NOK vs HSF

The HSF method consists of two basic concepts: a group and a link. These two concepts are graphically represented by circles (empty circle for groups, filled circles for links) that are mutually connected with arrows (to indicate the proper reading direction).

The concept of groups is used to represent a specific sign, group of signs, word, semantic categories and complex patterns. It is used in order to represent the sequence of selected terms on various levels of abstraction (a group represents a link to the first element in a sequence). The same group can occur on different levels of abstraction. Simply speaking, the concept of groups represents an aggregation of facts/knowledge, starting with the lowest level – signs. In the NOK method, context nodes represent aggregated knowledge. However, it is not clear how the HSF method represents knowledge that was created by generalizing knowledge on a lower level of abstraction. For example, based on the statements “The student Marko studies mathematics.” and “The student Ana studies mathematics.” a generalized statement such as “A student studies mathematics.” could be made. This statement can be easily represented in the NOK method using context nodes which are further connected to the original statements by context links.

The concept of links is used in the creation of a sequence of signs, group of signs, words, semantic categories etc. on various levels of abstraction. For example, for the pattern “Student”, the connection should represent the exact order of each sign in the pattern. It would be similar for words, sentences etc. In the NOK

method, sequence is represented by links between process nodes. It is possible to use such a sequence to represent a pattern (group) of signs, but the goal in using the NOK method is primarily to represent concepts that contain minimal knowledge.

The HSF method adds semantics to a certain sequence of terms by connecting them to a group concept that contains semantic description within itself (e.g. “is part” or “part of day”, etc.). In the NOK method, semantics of a certain sequence of concept that is represented by nodes and links (i.e. parts of a network) is represented by a sequence of context nodes (i.e. by an adequate network on a higher level of abstraction).

The HSF method uses the SOUL algorithm that gives it the ability of learning new patterns. A future research path in the NOK method is the development of an algorithm that would be able to generalize knowledge on a higher level of abstraction from facts/knowledge on a lower level of abstraction (analogous to the induction method). Moreover, a possible future research path is the use of existing generalized knowledge already in a NOK model in order to add new knowledge on a lower level of abstraction (analogous to the deduction method).

Figure 5 represents the statement “Peter finished the discussion.”, which is modelled using the HSF method.

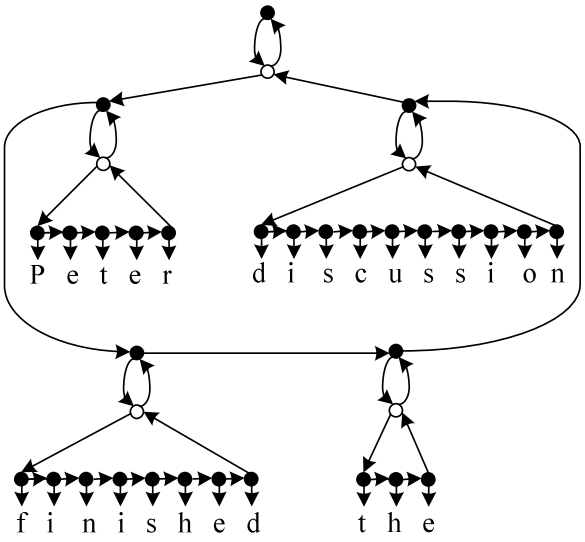


Figure 5. Representation of the statement "Peter finished he discussion." using the HSF method

From the model made using the HSF method, it is visible that the source statement is fully preserved (contrary to the NOK method where the statement is read as “Peter finish discussion”). It is clear that nodes represent specific signs, sequences of signs, and sentences. The correct direction of reading the model is guided by arrows – therefore it is possible to read it only in one direction which starts by the node on the highest level of abstraction (in the example above it is the black node on the top of the figure). In the NOK method, it is possible to start reading it from any node due to the existence of link role (i.e. the question).

D. NOK vs RDF

The RDF method represents the relation between web resources by using a named attribute and its value. Examples of web resources include various data, documents, images etc. A named attribute refers to a specific resource. The name of the attribute is defined using a Uniform Resource Identifier (URI), so that the attribute itself can be considered as a new resource. The value of the attribute can be another resource or other data (e.g. string, number, date etc.). The RDF method has many display modes, one of them being the RDF graph. It consists of three basic concepts: a node representing a resource, a node representing data and an arrowed line that connects nodes. An ellipsis graphically represents a resource, while a rectangle graphically represents data. An arrowed line represents a property and the property name is written above it. The arrow indicates the property value. Since resources have properties, on the arrowless side of the line there is always an ellipsis (a resource node), while on the other side of the line there is always either an ellipsis or a rectangle (i.e. a resource node or data).

Using the RDF method, structures of the type subject, object and predicate can be represented. Statements of this type can also be represented using the NOK method with nodes, process nodes and links between them. The RDF method cannot represent the sequence of execution of a process or logical constraints. The NOK method solves this problem using process nodes and process links. Similar to the NOK method, generalization/specialization can be represented using the RDF method (property “is part of” or “is of the type”).

The RDF method makes a distinction between two types of nodes – resource and data, while the name of the attribute is written above the line. In the NOK method all these three concepts are represented by nodes. There is no link or role name in the RDF method.

Figure 6 represents the statement “Peter finished the discussion.”, modelled using the RDF method.

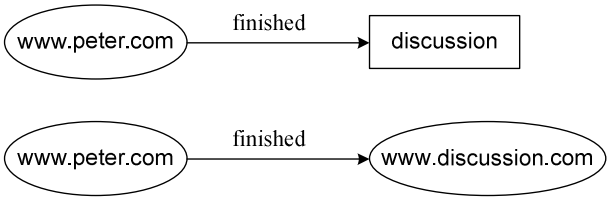


Figure 6. Representation of the statement "Peter finished the discussion." using the RDF method

The first model assumes that there is a web resource (internet page) that describes Peter. Finished is a property of this resource, and the value is discussion. In the second model, apart from the assumption that there is a web resource that describes Peter, it is also assumed that there is a web resource that describes the discussion. In this example, the first web resource has the property finished, the value of which is the second web resource. In the NOK method web resources could be represented by concept nodes, but their property would be represented by process nodes. An RDF model can be represented using a mathematical formalism as an arranged triad (predicate,

subject, object). Therefore, based on an RDF model, statements can be made that can also be made based on a NOK model, but the missing part is the role of a link, which gives it semantics.

IV. CONCLUSION

Based on the comparison of the NOK method with other methods, namely BG, MULTINET, HSF and RDF methods, the following characteristics of the NOK method can be highlighted:

1. simplicity – the NOK method has two basic elements (node and link). There are three different types of nodes (regular, context and process), and two link types (regular and context). Each link has its role (the question), which is used in reading the modelled knowledge.
2. expressiveness – the NOK method enables to represent knowledge on various levels of abstraction. The lowest level of abstraction represents relations between facts which are presented by regular and process nodes and their mutual relationships. Knowledge on higher levels of abstraction are represented by context and process nodes. The relationship between knowledge on higher and lower levels of abstraction is achieved through the use of context links.
3. Simplicity of reading – in the NOK method it is possible to start reading knowledge from any node using links roles. Therefore, the same knowledge can be reached in various ways (e.g. from the same NOK model, the following statements can be read: “Marko studies mathematics” and “Mathematics is studied by Marko”.).

The abovementioned characteristics of the NOK method will enable its simple implementation and its widespread usability. As a demonstration of its practical use, the authors of this paper started a project in order to demonstrate the possibilities in using the NOK method in business systems. It could be used in the creation of a knowledge base from an existing database.

Further paths in the research of this method include its mathematical formalization, the creation of algorithms that can conclude and suggest new knowledge from existing knowledge (using the induction method), the creation of algorithms of guided input of new knowledge on the basis of existing knowledge (use of the deduction method), and the creation of a user interface which is based on human written language.

REFERENCES

[1] R. Brachman, H. Levesque, "Knowledge representation and reasoning", The Morgan Kaufmann Series in Artificial Intelligence, Morgan Kaufmann, 2004.

[2] S. Grimm, P. Hitzler, A. Abecker, "Knowledge representation and ontologies logic", Ontologies and Semantic Web Languages, 2007.

[3] M. Chein, M. L. Mugnier, "Graph-based knowledge representation", Springer, London 2009.

[4] H. Helbig, "Knowledge representation and the semantics of natural language", Springer, London, 2006.

[5] M. Stanojević, S. Vraneš, "Knowledge representation with SOUL", Expert Systems with Applications, vol. 33, 2007; pp. 122-134.

[6] K. Tolle, "Introduction to RDF", 2000., <http://www.dbis.informatik.uni-frankfurt.de/~tolle/RDF/DBISResources/RDFIntro.html>, 12.10.2012.

Identification of the Global Text Skew Based on the Convex Hulls

Darko Brodić*, Dragan R. Milivojević**, Viša Tasić** and Zoran N. Milivojević***
* University of Belgrade, Technical Faculty in Bor, Bor, Serbia
** Institute for Mining and Metallurgy, Bor, Serbia
** Technical College Niš, Niš, Serbia
* dbrodic@tf.bor.ac.rs

Abstract - Old printed documents represent the important part of our heritage. For theirs preservation, the digitalization is indispensable. The paper proposes the algorithm for the identification of the global text skew. It is based on the connected components made by the filled convex hulls around each text element. Furthermore, connected components are enlarged by morphological dilatation. Currently, the biggest connected component is extracted. The global orientation of the document is identified by its orientation. Efficiency and correctness of the algorithm is examined by testing on datasets. Obtained results proved the algorithms correctness.

I. INTRODUCTION

Old historical documents represent the important part of our heritage. It is quite common for such documents to suffer from degradation problems. They include variable illuminated background, ink seeping, smear and strains. These phenomena make the process of image preprocessing difficult producing recognition errors. Therefore, appropriate methods should be developed to remove such noise from historical documents during digitalization.

In document automatic recognition systems, the quality of the input image is crucial to the final performance. There are a variety of interfering effects such as noise and skewing that appear during the scanning process. These components can damage the image and decrease the performance of the recognizer. Skew correction plays an important role in the image preprocessing. A small inclination in the document image can interfere in the layout analysis and consequently, in the rest of the process. Hence, the appearance of the skew in the document image leads to the significant failures in the optical character recognition (OCR) system. That's why, the identification of the object skew in the image is one of the most important tasks in digital image processing and document image analysis. It is due to OCR system sensitivity to any skew appearance in the text.

In this paper, we deal with old printed documents made by typewriter. They are characterized with shape regularity as any printed text [1], which contain letters with similar sizes. The distance between text lines is adequate, which facilitates separation of text lines. The orientation of the text lines is similar, which constitute the uniform text skew, which is called global text skew.

Due to the process of scanning, the text skew occurrence is unavoidable. To solve a problem, a large amount of methods has been proposed. They can be classified as follows [1]:

- Projection profiles methods,
- K-nearest neighbor clustering methods,
- Hough transforms methods,
- Fourier transformation methods,
- Cross-correlation methods, and
- Other methods.

Projection profile method is a straightforward method for the extraction of the text skew. It identifies the peaks in the histogram, which represent horizontal black pixel density [2]. The method is suitable for text with uniform skew only, due to its inability to identify the skew of the multi-column text.

K-nearest neighbor clustering method is based on the page layout analysis [3]. It creates the connected components according to the method of nearest neighbors clustering. The direction vectors for all pairs of the nearest neighbors are accumulated in a histogram. The peak of this histogram represents the skew angle. However, the nearest neighbor connectivity is made for each component, which represents the characters only. Hence, the incorporation of the noisy subparts of characters reduced the accuracy of the method.

The Hough transform is a widespread method for detecting lines in images [4]. It converts the image mapping points from the Cartesian space into the parameter's Hough domain. Furthermore, the voting procedure is carried out in a parameter domain. As a result, the object candidates are obtained as local maxima in a so-called accumulator space, which is constructed by the algorithm. Unfortunately, this method is complex and computer time intensive.

The Fourier transforms method is a Fourier domain representation of the projected profile method in the pixel domain. The skew angle of a document image corresponds to the direction where the density of Fourier's space becomes the largest [5]. However, its computing complexity is remarkably high.

The cross-correlation method is based on the correlation between two lines in a document image. It creates the cross-correlation function whose maximum represents the skew angle. The method is limited only to small skew angles up to 10° [6]. Interesting extension of this group of methods is given in [7]. It proposed the algorithm based on the co-operation of the log-polar transformation and cross-correlation. As a result, the cross-correlation cost function in the log-polar domain is formed. Its maximums give the optimum matching identifying the correct text skew.

The techniques classified as other methods are based mostly on morphological and geometrical transformation.

The morphological method includes the image dilatation and erosion using a line structuring element with different length [8]. This method proved its correctness with the smallest absolute error values.

Moments measure the pixel distribution in the image [9]. Accordingly, they are sensitive to the rotation. This method is suitable for text skew identification. However, the proposed technique is applicable for the single skew estimation, i.e. for the printed text documents only [10].

Method based on geometrical transformation is proposed in [11]. It detects a dominant skew in document images using piecewise covering of document objects by parallelograms. Method proved to be very accurate, but the skew angle range is limited to 15°.

Combination techniques are reputed to be the most efficient one. However, they are multistage and computer time intensive. In [12]-[13], preprocessing of document image is made by complex decision making. It is completed with complex heuristic geometrical filtering. Global text skew is identified with the cross-correlation method applied to remain connected components. At the end, local text skew is calculated with the least square method. This technique performs local skew estimation, reliable text localization without restriction of the skew angle value.

The main contribution of this paper is the algorithm suitable for the recognition of the text skew in the old printed documents characterized with dominant skew.

Organization of this paper is as follows. Section 2 describes proposed algorithm. Section 3 defines text experiments. Section 4 shows test result, and discusses it. Section 5 makes conclusions.

II. PROPOSED ALGORTHM

The algorithm identifies the global text skew of the printed documents, which represents the dominant text skew of the whole document. It consists of the steps that follow:

A. Uneven illumination reduction with binarization.

B. Convex hulls extraction.

C. Joining text objects with binary morphology.

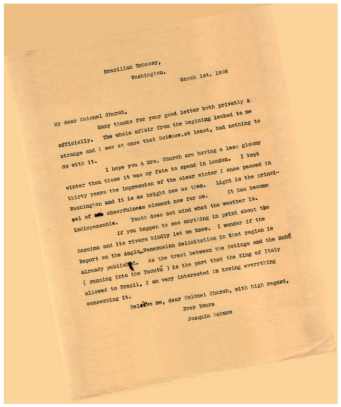


Figure 1. Old printed document skewed by 10°

D. Extraction of the longest object.

E. Skew estimation of the longest object by the moments.

F. Global de-skewing of the original document.

A. *Uneven illumination reduction with binarization*

Figure 1 shows typical old printed document skewed during the scanning process.

Along with skewness, the old documents are characterized with the uneven illumination distribution in its background. This illumination degrades the visual quality of the text image and makes it difficult to recognize the content.

In order to efficiently solve this problem, an efficient edge-based light balancing scheme (ELBS) for text images is proposed [14]. After application of this algorithm, the document text image is a grayscale without uneven illumination. It is given by matrix **D**, which consists of M rows, N columns, and L intensity levels of gray. L is the integer from $\{0, \ldots, 255\}$. $D(i, j) \in \{0, \ldots, 255\}$, where $i = 1, \ldots, M$ and $j = 1, \ldots, N$.

After the binarization, it is transformed into a binary image **B** according to the following:

$$B(i,j) = \begin{cases} 1 & if\ D(i,j) \geq D_{th} \\ 0 & if\ D(i,j) < D_{th} \end{cases}, \qquad (1)$$

where D_{th} represents the global threshold sensitivity value [15]. Accordingly, document text image is given as binary matrix **B** featuring M rows and N columns, and two intensity levels. Figure 2 shows document after the reduction of the uneven illumination as well as the binarization.

B. *Convex hulls extraction*

Instead of using bounding boxes [16], the convex hulls have been exploited. Convex hulls create smaller regions around the text compared to the bounding boxes. Hence, the possibility for touching neighbor text fragments has been reduced.

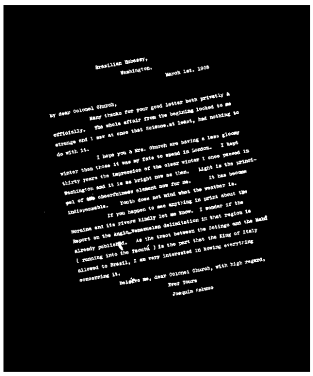


Figure 2. Document after the uneven illumination reduction and binarization

The convex hull of a set of point is the smallest convex set that contains all points. In the plane, we can visualize it as a stretched rubber band surrounding the points that, when released, takes a polygonal shape. The extreme point is the set that are the vertices of the convex polygon. For a set *S*, it can be defined as the set of points that can be expressed as convex combinations of the points in that set *S* [17]. After the extraction of the convex hulls, they are filled with white pixels. Such a text image is represented with matrix **C**. Figure 3 shows document after convex hull extraction.

C. Joining text objects with binary morphology

To form usable connected components (CC), each convex hull is growing. This step is accomplished by morphological erosion applied to **C**. This way, the adjacent CCs are merged establishing the text line. Structuring element **S** with variable width line is used. Morphological operation is given as:

$$\mathbf{Y} = \mathbf{C} \oplus \mathbf{S} \ , \tag{2}$$

In order not to touch or join separate neighbor text lines, the width of the line should be chosen carefully. It heavily depends on each CC's height. Empirically, it is given as approx. 30% of the connected component's height. Figure 4 shows CC established by erosion.

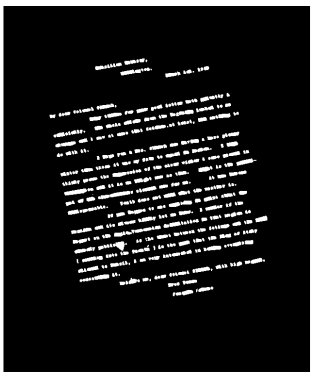


Figure 3. Document after convex hull extraction

Figure 4. Document after joining convex hull objects with erosion

Illustration of the text covering with morphologically extended convex hulls is given in Figure 5.

D. Extraction of the longest object

From **Y**, the longest connected components CC_L is extracted with the longest common subsequence (LCS) [16]. The longest connected component CC_L is shown in Figure 6.

E. Skew estimation of the longest object by the moments

Moment defines the measure of the pixel distribution in the image. It identifies global image information. This information depends on the object contour. Moments of the binary image **B** are [18]:

$$m_{pq} = \sum_{i=1}^{N} \sum_{j=1}^{M} i^p j^q \ , \tag{3}$$

where p and q = 0, 1, 2, 3, ..., n, and n represent the order of the moment. The central moment's μ_{pq} for binary image **B** can be calculated as:

$$m_{pq} = \sum_{i=1}^{N} \sum_{j=1}^{M} (i - \overline{x})^p (j - \overline{y})^q \ . \tag{4}$$

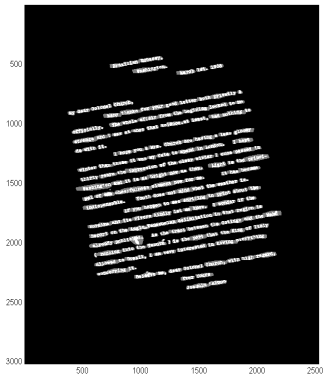


Figure 5. Document after joining convex hull objects around text elements

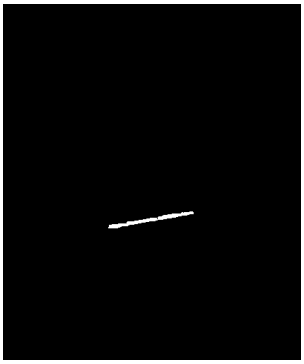


Figure 6. Extraction of the longest object

The image feature, which represents the object orientation θ can be obtained using the moments. The angle θ illustrates the angle between the object and the horizontal axis. In order to estimate that angle, the following equation is used [18]:

$$q = \frac{1}{2}\arctan\left(\frac{2m_{11}}{m_{20} - m_{02}}\right). \quad (5)$$

All these features characterize separate object.

F. *Global de-skewing of the original document*

According to the global text skew, the binary document is de-skewed, which is shown in Figure 7.

III. EXPERIMENTS

The basic goal of the experiment represents the process of the evaluation algorithm's ability to estimate text skew. It is performed on real custom dataset. Dataset consists of twelve old printed documents made by typewriter. It is shown in Figure 8.

Furthermore, samples from dataset are rotated for the angle θ, which is changed from 0° to 10° by 1° and from 10° to 45° by 5° steps around *x*-axis in positive direction. The illustration is shown in Figure 9.

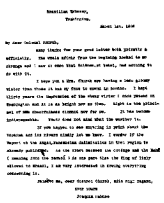


Figure 7. Document deskewing according to estimated global text skew

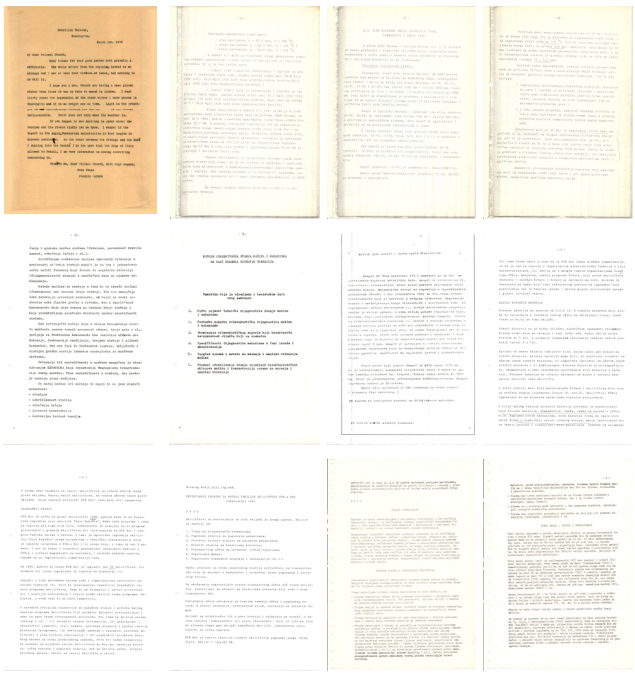
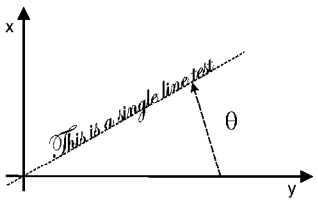


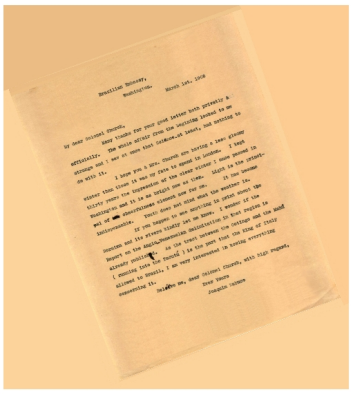
Figure 8. Dataset samples

All text samples are given in the resolution of 300 dpi as well as in lower resolutions: 150, 100, 75 and 50 dpi. The results are evaluated by the absolute deviation:

$$\Delta q_A = \left| q_{REF} - q_A \right|, \quad (6)$$



(a)



(b)

Figure 9. (a) Illustration of the text skew test, (b) Real dataset

where θ_{REF} is the referent skew of the input text sample and θ_A is the text skew obtained by the algorithm. The relative error (*RE*) is given as:

$$RE(q) = \frac{\Delta q_A}{q_{REF}}. \quad (7)$$

IV. Results and Discussion

Table I show absolute deviation of the global text skew for the real dataset.

TABLE I. Absolute Deviation

	$\Delta\theta$(°)				
θ_{REF} (°)	300 dpi	150 dpi	100 dpi	75 dpi	50 dpi
0	0.4764	0.4644	0.4167	0.4903	0.5773
1	0.4438	0.4449	0.4280	0.4059	0.5378
2	0.4385	0.4109	0.4062	0.4384	0.4943
3	0.4599	0.4046	0.4764	0.4224	0.4725
4	0.4030	0.4436	0.4472	0.5166	0.5029
5	0.4218	0.3968	0.4536	0.4674	0.5577
6	0.4148	0.4460	0.4597	0.4575	0.4262
7	0.4295	0.4075	0.4498	0.4628	0.6191
8	0.4770	0.4500	0.4529	0.4250	0.5965
9	0.4634	0.4366	0.4073	0.4218	0.6414
10	0.4500	0.4461	0.4934	0.4927	0.7326
15	0.4616	0.6302	0.6061	0.9292	1.0112
20	0.4710	0.6444	0.6452	1.1104	1.1068
25	0.4823	0.6910	0.6710	1.3061	1.5084
30	0.4827	0.7016	0.9542	1.4755	1.5822
35	0.7296	0.7738	1.3487	1.6119	1.4585
40	0.7583	0.8054	1.5281	1.6468	1.8245
45	0.7686	0.7925	1.4999	1.8097	1.9028

From Table I, the results are as follows:

- up to 10º, absolute deviation is from 0.4º to 0.48º, for images given in 150 and 300 dpi, and from 0.4º to 0.73º, for the images given in 50, 75 and 100 dpi;

- up to 30º, absolute deviation is from 0.4º to 0.7º, for images given in 150 and 300 dpi, and from 0.4º to 1.58º, for the images given in 50, 75 and 100 dpi;

- up to 45º, absolute deviation is from 0.4º to 0.8º, for images given in 150 and 300 dpi, and from 0.4º to 1.9º, for the images given in 50, 75 and 100 dpi.

Figure 9 illustrates the absolute deviation for images given in 150 and 300 dpi. It is almost the same, except for θ_{REF} between 15º and 35º, where is quite better for the images in 300 dpi. Figure 10 illustrates the absolute deviation for images given in 50, 75 and 100 dpi. It is quite similar. However, the exception is the range of θ_{REF} between 15º and 35º, where the images in the lowest resolution of 50 and 75 dpi show worse results.

Table II shows relative deviation of the global text skew for the real dataset. From Table II, the results are as follows:

- up to 10º, maximum relative deviation is from 0.04 to 0.44, for images given in 150 and 300 dpi, and from 0.05 to 0.55, for the images given in 50, 75 and 100 dpi;

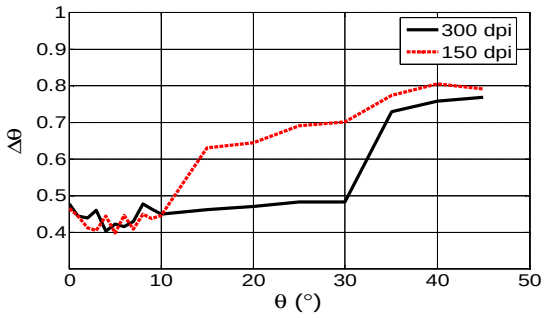


Figure 10. Absolute deviation for images in 150 and 300 dpi

TABLE II. Relative Deviation

	$RE(\theta) = \Delta\theta/\theta$				
θ_{REF} (°)	300 dpi	150 dpi	100 dpi	75 dpi	50 dpi
0	-	-	-	-	-
1	0.4438	0.4449	0.4280	0.4059	0.5378
2	0.2193	0.2055	0.2031	0.2192	0.2472
3	0.1533	0.1349	0.1588	0.1408	0.1575
4	0.1008	0.1109	0.1118	0.1292	0.1257
5	0.0844	0.0794	0.0907	0.0935	0.1115
6	0.0691	0.0743	0.0766	0.0762	0.0710
7	0.0614	0.0582	0.0643	0.0661	0.0884
8	0.0596	0.0563	0.0566	0.0531	0.0746
9	0.0515	0.0485	0.0453	0.0469	0.0713
10	0.0450	0.0446	0.0493	0.0493	0.0733
15	0.0308	0.0420	0.0404	0.0619	0.0674
20	0.0236	0.0322	0.0323	0.0555	0.0553
25	0.0193	0.0276	0.0268	0.0522	0.0603
30	0.0161	0.0234	0.0318	0.0492	0.0527
35	0.0208	0.0221	0.0385	0.0461	0.0417
40	0.0190	0.0201	0.0382	0.0412	0.0456
45	0.0171	0.0176	0.0333	0.0402	0.0423

- up to 30º, maximum absolute deviation is from 0.02 to 0.44, for images given in 150 and 300 dpi, and from 0.05 to 0.55, for the images given in 50, 75 and 100 dpi;

- up to 45º, maximum absolute deviation is from 0.01 to 0.44º, for images given in 150 and 300 dpi, and from 0.05 to 0.55, for the images given in 50, 75 and 100 dpi.

Figure 11 illustrates the relative deviation for images given in 150 and 300 dpi. It is almost identical for the image in different resolutions, in the whole range of θ_{REF}. It is the most intensive at small values of θ_{REF}, up to 10º.

Figure 12 illustrates the absolute deviation for images given in 50, 75 and 100 dpi. The results are quite similar to the previous ones.

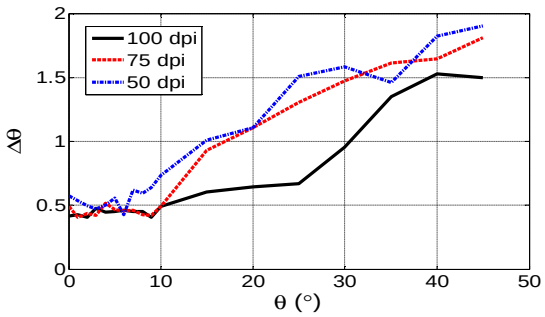


Figure 11. Absolute deviation for images in 50, 75 and 100 dpi

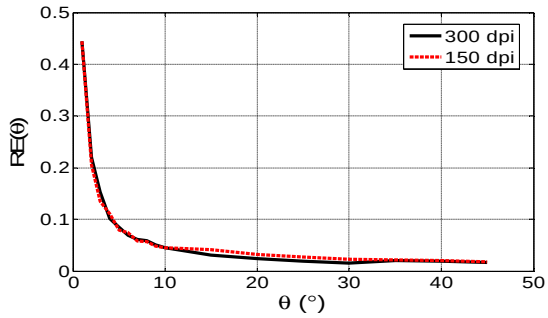


Figure 12. Relative deviation for images in 150 and 300 dpi

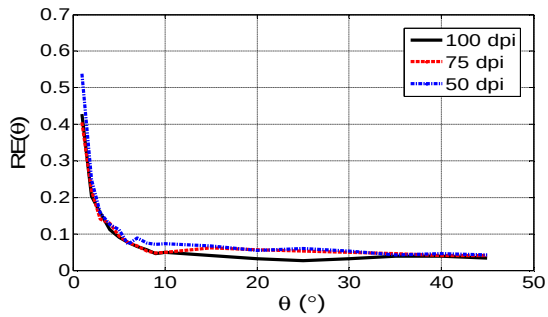


Figure 13. Relative deviation for images in 50, 75 and 100 dpi

Again, relative deviation is almost identical for the image in different resolutions and it is the most intensive at small values of θ_{REF}, up to 10º.

All above results are good ones. Using the images in lower resolution brings just slightly worse results. It is due to the use of moments in the method, which has certain insensitivity to the image resolution.

Comparison with other methods shows correctness and robustness of proposed method. This way, it is more correct than widespread projection profile method with absolute deviation < 0.75º (results dispersion approx. 0.3º) vs. < 1.5º for images in 300 dpi [19]. However, it is true especially for images given in lower resolutions. Furthermore, it should be used for multi-column text as well. Proposed method, in original form, doesn't have the versatility and correctness as multi-stage methods proposed in [12]-[13]. However, these methods are much more complex ones. They include complicated steps of geometrical filtering in preprocessing stage. Due to that, these methods exclude some "noise" elements, which aren't needed for correct identification of text skew. As a consequence, they are more computer time intensive.

In further developing steps, proposed method should be expanded. Inclusion of geometrical filtering in preprocessing stage of method should make its improvement. Hence, this step will contribute to lower dispersion of absolute error value, which will cut some of its high values.

V. CONLUSION

The paper presents the robust method for the estimation of the global text skew. It is based on the convex hulls extraction over the text as well as its extension by the binary morphology. After that, the longest object is extracted. On the basis of its contour, the moment based method estimates its orientation, which represents the global text skew. The proposed method is examined on the real dataset. The results are promising especially for the images in lower resolution. Further improvement should be made by including geometrical filtering in preprocessing stage of the method.

REFERENCES

[1] A. Amin, S. Wu, "Robust Skew Detection in Mixed Text/Graphics Documents," in Proceedings of 8th ICDAR, Seoul, Korea, 2005, Vol.1, pp. 247–251.

[2] R. Manmatha, N. Srimal, "Scale Space Technique for Word Segmentation in Handwritten Manuscripts", in Proceedings of 2nd ICSSTCV, LNCS 1682, M. Nielsen et all. (Eds.): Springer-Verlag, London, 1999, pp. 22–33.

[3] L. O'Gorman, "The Document Spectrum for Page Layout Analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993, Vol.15, No.11, pp. 1162–1173.

[4] G. Louloudis, B. Gatos, I. Pratikakis, C. Halatsis, "Text Line Detection in Handwritten Documents", Pattern Recognition, 2008, Vol.41, No.12, pp. 3758–3772.

[5] W. Postl, "Detection of linear oblique structures and skew scan in digitized documents," in 8th ICPR, Paris, France, 1986, pp. 687–689.

[6] H. Yan, "Skew Correction of Document Images Using Interline Cross-Correlation", CVGIP: Graphical Models and Image Processing, 1993, Vol.55, No.6, pp. 538–543.

[7] D. Brodić, Z. N. Milivojević, Log-polar Transformation as a Tool for Text Skew Estimation, Electronics and Electrical Engineering, 2013, Vol. 19, No. 2, pp. 61–64.

[8] B. V. Dhandra, V. S. Malemath, H. Millikarjun, R. Hegadi, "Skew Detection in Binary Image Documents Based on Image Dilatation and Region Labeling Approach", in Proceedings of 18th IEEE ICPR, Hong Kong, China, 2006, pp. 954–957.

[9] P. Shivakumara, G. H. Kumar, D. S. Guru, P. Nagabhushan, "A Novel Technique for Estimation of Skew in Binary Text Document Images based on Linear Regression Analysis", Sādhanā, 2005, Vol.30, No.1, pp. 69–86.

[10] J. Flusser, B. Zitova, T. Suk, Moments and Moment Invariants in Pattern Recognition, John Wiley & Sons, 2009.

[11] Chien-Hsing Chou, Shih-Yu Chu, Fu Chang, "Estimation of Skew Angles for Scanned Documents Based on Piecewise Covering by Parallelograms," Pattern Recognition, 2007, Vol. 40, No. 2, pp. 443–455.

[12] P. Saragiotis, N. Papamarkos, "Local Skew Correction in Documents," International Journal of Pattern Recognition and Artificial Intelligence, 2008, Vol. 22, No. 4, pp. 691–710.

[13] M. Makridis, N. Nikolau, N. Papamarkos, "An adaptive technique for global and local skew correction in color documents," Expert Systems with Applications, 2010, Vol. 37, No. 10, pp. 6832–6843.

[14] Kuo-Nan Chen, Chin-Hao Chen, Chin-Chen Chang, "Efficient illumination compensation techniques for text images", Digital Signal Processing, 2012, Vol. 22, No. 5, pp. 726–733.

[15] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Transactions on Systems, Man, and Cybernetics, 1979, Vol. 9, No. 1, pp. 62–66.

[16] D. Brodić, D. R. Milivojević, "An algorithm for the estimation of the initial text skew," Information Technology and Control, 2012, Vol. 41, No. 3, pp. 211–219.

[17] F. P. Preparata, M. I. Shamos, Computational Geometry: An Introduction, Berlin, Springer, 1985.

[18] G. Kapogiannopoulos, N. Kalouptsidis, "A fast high precision algorithm for the estimation of skew angle using moments," in Proceedings of SPPRA, Crete, Greece, 2002, pp. 275–279.

[19] D. Brodić, D. R. Milivojević, Z. N. Milivojević, "Analysis and Comparison of Different Methods for the Skew Estimation of the Printed Text," (under review).

Indexing Positively Constructed Formulas

A.A. Larionov*, I.N. Terekhin*
* Irkutsk State University, Irkutsk, Russia
ale@prisnif.org
i.terhin@gmail.com

Abstract - The calculus of positively constructed formulas (PCF) is a first-order formalism that has many features useful for solving problems of dynamic systems control. For investigation of this formalism, analysis of logical inference of formulas is used. The logical inference consist of many PCF. The analysis of big amount of formulas require special operations: search of formulas, search by template, unification search etc. Usually for solving these problems the indexing methods are used. For indexing terms as a tree-like structures the methods of indexing such as path indexing, substitution tree indexing, discrimination tree indexing etc., are used. This methods well used in many logical systems. Since PCF, like a term, have a tree-like structure, the generalization of indexing methods to the PCF indexing is possible. In this paper, we consider this problem. The modification of indexing methods and experiments concerning the analysis of logical inferences (big sets of PCF) are presented.

I. INTRODUCTION

Automated theorem proving (ATP) is one of the developing areas of artificial intelligence (AI), which is based on methods of mathematical logic (proof theory) and realized in computer programs to automatically search for logical inference. ATP is the most formalized direction of automation of deductive constructions, and automated reasoning. One of the actual applications of the ATP development is searching for open mathematical problem solution, such as described in [1].

The most state-of-art systems of ATP, such as the Otter, Vampire, EQP, E, and others are based on the well-known resolution method and its modifications. The algorithms used in the implementations of these ATP systems are suitable for other computer-oriented calculi automation.

A logical formalism of knowledge representation and processing, which is based on the positive logical calculus formalism of positively constructed formulas (PCF) [5], is described briefly in the paper. This calculus has a much greater heuristic conversation possibilities of the original logically formalized knowledge than the clause language of resolution method [2]. The possibilities allow one to develop specialized modifications of the default strategies to reflect structural and semantic properties of the formalized knowledge.

The PCF formalism was applied in [3] for improvement of a system for automatic synthesis of theorems "VLF-1" (Vector Lyapunov's Functions) by means of implementation of first-order logical equation solver. In [4] intuitionistic fragments of the calculus was used for problem solving.

II. THE CALCULUS OF POSITIVELY CONSTRUCTED FORMULAS

A. Language and calculus

The PCF calculus $\boldsymbol{JF}$ is a triple $< \boldsymbol{LF}, Ax\boldsymbol{JF}, \omega >$, where $\boldsymbol{LF}$ is the PCF language, $Ax\boldsymbol{JF}$ is the only axiom schema and ω is the unique inference rule of $\boldsymbol{JF}$.

Let us denote the set of all conjuncts as *Con* and assume that a conjunct is either a finite set of usual atoms of a first-order language, or ***False***, where ***False*** satisfies the property $A \subset \boldsymbol{False}$ for any $A \in Con$. The empty conjunct is denoted by ***True***. It is clear that if $A \in Con$ then $A \cup \boldsymbol{False} = \boldsymbol{False}$. Atoms of any conjunct (except True and False) may contain variables, constants and function symbols.

Definition 1. Let $\bar{x}$ be a set of variables (possible empty set) and A be a conjunct. Well-formed formulas of the PCF language are defined as follows:

An expression $\exists\bar{x}: A$ is a ∃-formula; an expression $\forall\bar{x}: A$ is a ∀-formula.

Let $G_1, \dots, G_k$ be a ∃-formulas, then ∀-formula is a formula of the following form: $\forall\bar{x}: A(G_1, \dots, G_k)$.

Let $G_1, \dots, G_k$ be a ∀-formulas, then ∃-formula is a formula of the following form: $\exists\bar{x}: A(G_1, \dots, G_k)$.

An expression is a well-formed PCF if it is constructed based on above Definition 1 rules only.

The variable from $\bar{x}$ are bound by a corresponding quantifiers, and these variables are called ∀-variables and ∃-variables correspondingly.

Let us define a semantic of PCFs by a common semantics of a corresponding first-order formula (FOF) in the classical first-order logic (FOL).

Definition 2. Let $A = \{A_1, \dots, A_l\}$ be a conjunct and $\bar{x} = \{x_1, \dots, x_n\}$ be a set of variables. By $A^{\&}$ we denote $A_1 \& \dots \& A_l$ and $\boldsymbol{False}^{\&} = False, \boldsymbol{True}^{\&} = True$ (propositional constants). By $\boldsymbol{F}^{FOF}$ we denote an image of PCF formula F in the FOL.

If $F = \exists\bar{x}: A$ then $F^{FOF} = \exists x_1 \dots \exists x_n (A^{\&})$. If $F = \forall\bar{x}: A$ then $F^{FOF} = \forall x_1 \dots \forall x_n (A^{\&})$.

If $F = \exists\bar{x}: A(G_1, \dots, G_k)$ then

$$F^{FOF} = \exists x_1 \dots \exists x_n \left(A^{\&} \& (G_1^{FOF} \& \dots \& G_k^{FOF}) \right).$$

The research was supported by Irkutsk State University, grant No 2012-01-02; 2012-01-01; by Federal Framework Program "Scientific and Teaching Staff of Innovative Russia for 2012-2013" State Contract № 14.B37.21.0365;

If $F = \forall\bar{x}: A(G_1, \dots, G_k)$ then

$$F^{FOF} = \forall x_1 \dots \forall x_n \left(A^{\&} \to (G_1^{FOF} \vee \dots \vee G_k^{FOF})\right).$$

Any PCF can be obviously represented by a tree structure. Each node of that tree has the following form $\exists\bar{x}: A$ or $\forall\bar{x}: A$. Therefore, for readability we will represent PCF as tree, as well as we will use the corresponding terminology: node, branch, root, leaf, etc.

If a PCF F roots from the $\forall: \boldsymbol{True}$ node and each leaf of F is a $\exists$-node then F is called a PCF in the *canonical form*. It is clear, that any PCF F can be converted to the canonical form by the following transformations:

1. If F is not canonical $\forall$-formula then $\forall: \boldsymbol{True}(\exists: \boldsymbol{True}(F)$ is PCF rooted from the $\forall: \boldsymbol{True}$.
2. If F is $\exists$-formula then $\forall: \boldsymbol{True}(F)$ is PCF rooted from the $\forall: \boldsymbol{True}$.
3. If F has leaf $\forall\bar{x}: A$ then the node $\exists: \boldsymbol{False}$ can be added as child of this leaf.

In the sequel, only PCFs in the canonical form are considered.

The axiom schema $Ax\boldsymbol{JF}$ of the PCF calculus $\boldsymbol{JF}$ has the following form:

$$\forall: \boldsymbol{True}\left(\exists\bar{x}_1: \boldsymbol{False}(\widetilde{\Phi}_1), \dots, \exists\bar{x}_n: \boldsymbol{False}(\widetilde{\Phi}_n)\right)$$

In the PCF calculus to prove F we attempt to refute its negation, that's why the axioms are identical to false. Thus, an inference process in the PCF calculus is a *refutation* process.

Definition 3. Let us assume that a question $\forall\bar{y}: A$ to a base $\exists\bar{x}: B$ has an answer Θ if and only if Θ is a substitution $\bar{y} \to H^{\infty}$ and $A\Theta \subseteq B$, where H^{∞} is Herbrand universum based on variables from $\bar{x}$, constants and function symbols that occur in corresponding base subformula.

Definition 4. If a PCF F has the structure $\forall: \boldsymbol{True}(\exists\bar{x}: (B(\Phi), \Sigma)$, where Σ is a list of other base subformulas, and Φ is a list of question subformulas, that contains a question subformula $\forall\bar{y}: A\left(\exists\bar{z}_i: C_i(\Psi_i)\right)_{i=\overline{1,k}}$, then a formula ωF is an application result of the unary inference rule ω to the question $\forall\bar{y}: A$ with an answer Θ and $\omega F = \forall: \boldsymbol{True}\left(\exists\bar{x} \cup \bar{z}_i: B \cup C_i\Theta(\Phi \cup \Psi_i\Theta)_{i=\overline{1,k}}, \Sigma\right)$.

After appropriate renaming some bound variables inside each subformula the expression ωF will satisfy the well-formed PCF requirements.

Any finite sequence of PCFs $F, \omega F, \omega^2 F, \dots, \omega^n F$, where $\omega^n F \in AxJF$, is called a *refutation* of F in the PCF calculus.

A question with $\exists: \boldsymbol{False}$ consequent node is called *goal question*.

III. INDEXING

A. Term Indexing

In information technologies, the problem of speed increase of searching the data is often solved with the indexation, which is a quite popular technique in relational databases. In our case, the main object for indexing is the general term, which is a tree-like structure, so, the methods for indexation of relational databases is not applicable.

Term indexation was researched quite well, as a practical use in ATP system, and from theoretical point of view. There are many papers, related to this theme [6, 7, 9]. The methods described in these papers allows to efficiently finding such terms, which satisfies the specified criteria: if the term is equal to query term; if the term is instantiation; if the term is a generalization or it is a unification. For developed prover we need the described operations and, because in base subformulas there are mainly ground terms, we choose to use the well-known path-indexing technique [6].

We describe path indexing method as in [6]. For every symbol in term the list of paths are created. Path – is a sequence of alternating symbols and numbers, where the number define the position of symbol among the children nodes. For example, the atom A(e,f(f(x,k),y),m) is A, A1e, A2f, A2f1f, A2f1f1x, A2f1f2k, A2f2y, A3m. I.e. for every symbol, there is a corresponding path from root to this symbol in tree representation of general term. You can see more details on Fig 1. Every path contain the pointer to the corresponding term, i.e. the term, for which we build the path. The paths stored in sorted form.

For example, we have A(a,f(x,b)) and the set of atoms {A(a,f(c,b)), A(a,f(b,e)),A(a,f(k,b)), A(b,f(e,b))}.

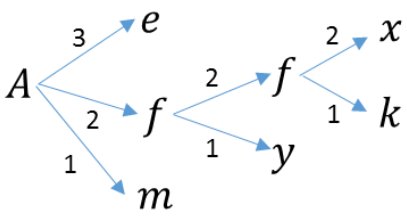


Figure 1. Tree representation of term

Every atom, which is a ground instance for specified, contains in paths the list of the following paths: A1a, A2f2b. Therefore, for ground instance search of specified atom we need to find the intersection of the sets of atoms (to which the paths is pointing). In particular, the path A1a points to set {A(a,f(c,b)), A(a,f(b,e)),A(a,f(k,b))}, and A2f2b to {A(a,f(c,b)), A(a,f(k,b)), A(b,f(e,b))}.

The intersection of these sets is a set {A(a,f(c,b)),A(a,f(k,b))}, and it is a set of instances for atom A(a,f(x,b))

The methods of term indexing is quite popular for ATP systems. They are used in almost every implementation.

B. Formulas indexing

Proving of big formulas can be processed in several stages storing the intermediate results on hard drives. The logical inference can be represented as a chain of formulas. To analyze described situations we need to check formulas for specified properties. Because PCF, as a term, has a tree structure, one can use the described above methods of path

indexing. However, in this case the term indexation need to be modified, because in case of term indexation the path is clearly defined, and in case of PCF (indexation) the number of children nodes is stored and for the symbol there is a corresponding described characteristic of PCF node (type quantifier). In our prover, we use the following features: type of quantifier, size of conjunct, lexical representation of one of terms of conjunct.

Consider the following PCF:

$$\forall: True - \begin{cases} \exists: A(e) - \forall x: A(x) - \exists: False \\ \exists: B(e) - \begin{cases} \forall x: B(x)C(x) - \exists: P(x,x) \\ \forall y: P(y,e) - \exists: False \end{cases} \end{cases}$$

In case we are characterize every node as quantifiers, the tree representation is on following Fig. 2.

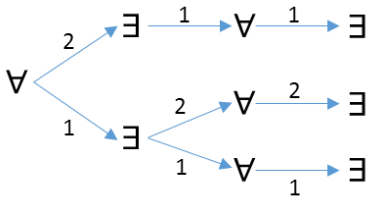


Figure 2. Nodes are quantifiers

In case we are characterize every node as a length of conjuncts, the tree representation is on following Fig. 3.

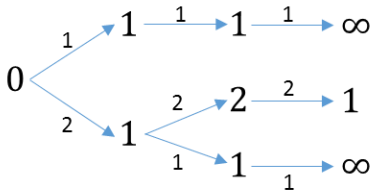


Figure 3. Nodes are length of conjuncts

Therefore, as a characteristic for the node can be any, for example, we can use lexicographical representation.

IV. Experiments

For the testing purposes, we use the problems from TPTP library [8]. This library de facto is a standard for testing the ATP systems. It contains the thousands of problems divided by rating of problem, domain of problem etc. Some problems was not solved yet. Every solution is logical inference (chain of formulas). For some problems, for example COM008+1.p, the length of logical inference is 33551 steps (i.e. chain of formulas is 33551 formulas length). All this formulas was indexed and the searching was produced. We have conduct experiments on 1000 problems. This is a first-order logic problems without equality. The length of logical inference for solved (800) problems is from 1 to ~50000 steps.

On the Figure 4, the results of testing of searching formulas are presented. We have used the searching without indexation (just a straightforward search) and searching with implemented indexation. X-axis is an amount of formulas in logical inference (i.e. length of formulas chain). Y-axis is a time of search in seconds.

As we see from Figure 4 the growth of searching time on problems with small number of logical inference steps is not much different from with indexing and without indexing methods. The greater the number of formulas in logical inference the greater the difference between methods. Hence, the application of formulas indexing is justified on long (from ~5000 steps) logical inference.

Formulas indexing as in Figure 2 allows search the formulas by given formula structure. And formulas indexing on Figure 3 allow search the formulas by given conjuncts length.

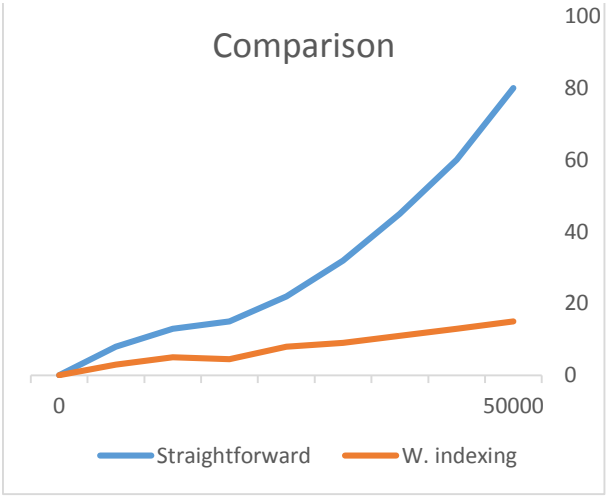


Figure 4. Comparison chart

V. Conclusion

In this paper, the PCF indexing is presented. As a basis of this indexing we take the well-known term indexing method – path-indexing [6]. We have tested the speed of search of formulas with indexing and without. Indexing allows to quickly analyzing the logical inference, and one can use this possibility for developing new knowledge about given problems. Because the PCF calculus is not well-researched and well-known, considered approach will provide the possibility to deeper investigation of the PCF calculus.

References

[1] McCune W. Solution of the Robbins Problem. Journal of Automated Reasoning, V.19(3), p.263-276, December 1997.

[2] Robinson J. A. A Machine--Oriented Logic Based on the Resolution Principle. Journal of the ACM. (12). 1965. P.23-41.

[3] Vassilyev S. N. Machine Synthesis of Mathematical Theorems. J.of Logic Programming. V.9(2\&3). 1990. P.235-266.

[4] Butyrin S. A., Makarov V. P., Mukumov R. R., Somov Ye., Vassilyev S. N. An Expert System for Design of Spacecraft Altitude Control System. Artificial Intelligence in Engineering. V.11(1). 1997. P.49-59.

[5] Vassilyev S.N., Zherlov A.K., Fedunov E.A. Fedosov B.E. Intelligent control of dynamic systems. Moscow: Fizmatlit, 2000. (in russian).

[6] McCune W. W. Experiments with Discrimination-Tree indexing and Path Indexing for Term Retrieval. Journal of Automated Reasoning. 1992. V.9(2). P.147-167

[7] Graf P. Substitution Tree Indexing. Proceedings of the 6th International Conference on Rewriting Techniques and Applications. 1995. P.117-131.

[8] Sutcliffe, G. The TPTP Problem Library and Associated Infrastructure: The FOF CNF Parts, v3.5.0 // Journal of Automated Reasoning. V43, N4, pp.337–362.

[9] Robinson, A., Voronkov, A. (eds.) Handbook of Automated Reasoning. – MIT Press, Cambridge, 2001. – 2150p.

On a Method of Solution the Equality Problem for Positively Constructed Formulas

I.N. Terekhin*, A.A. Larionov*
* Institute of Mathematics, Economics and Informatics of Irkutsk State University, Irkutsk, Russia
i.terhin@gmail.com
ale@prisnif.org

Abstract - The calculus of positively constructed formulas (PCF) is a first-order formalism that has many features useful for solving problems of dynamic systems control. This formalism is used as a basis for automatic theorem proving (ATP) systems. Many problems in the field of ATP can be formalized only with the use of equality predicate. Using this predicate will require adding extra axioms into the problem. This is not efficient for ATP systems, so the problem of equality must be solved without that axioms. In this paper, we introduce the solution of equality problem for PCF formalism. The results are integrated into the developed ATP system.

I. INTRODUCTION

Automated theorem proving (ATP) is one of the developing areas of artificial intelligence (AI), which is based on methods of mathematical logic (proof theory) and realized in computer programs to automatically search for logical inference. ATP is the most formalized direction of automation of deductive constructions, and automated reasoning. One of the actual applications of the ATP development is searching for open mathematical problem solution, such as described in [1].

The most state-of-art systems of ATP, such as the Otter, Vampire, EQP, E, and others are based on the well-known resolution method and its modifications. The algorithms used in the implementations of these ATP systems are suitable for other computer-oriented calculi automation.

A logical formalism of knowledge representation and processing, which is based on the positive logical calculus formalism of positively constructed formulas (PCF), is described briefly in the paper. This calculus has a much greater heuristic conversation possibilities of the original logically formalized knowledge than the clause language of resolution method. The possibilities allow one to develop specialized modifications of the default strategies to reflect structural and semantic properties of the formalized knowledge.

The PCF formalism was applied in [2] for improvement of a system for automatic synthesis of theorems "VLF-1" (Vector Lyapunov's Functions) by means of implementation of first-order logical equation solver.

Many problems from practical domains require the ATP system to work with equations. In this paper, we introduce the theory related to work with equation within the calculus of positively constructed formulas.

The research was supported by Irkutsk State University, grant No 2012-01-01; 2012-01-02; by Federal Framework Program "Scientific and Teaching Staff of Innovative Russia for 2012-2013" State Contract № 14.B37.21.0365

II. THE CALCULUS OF POSITIVELY CONSTRUCTED FORMULAS

A. *Language and calculus*

The PCF calculus $\boldsymbol{JF}$ is a triple $< \boldsymbol{LF}, Ax\boldsymbol{JF}, \omega >$, where $\boldsymbol{LF}$ is the PCF language, $Ax\boldsymbol{JF}$ is the only axiom schema and ω is the unique inference rule of $\boldsymbol{JF}$.

Let us denote the set of all conjuncts as *Con* and assume that a conjunct is either a finite set of usual atoms of a first-order language, or ***False***, where ***False*** satisfies the property $A \subset \boldsymbol{False}$ for any $A \in Con$. The empty conjunct is denoted by ***True***. It is clear that if $A \in Con$ then $A \cup \boldsymbol{False} = \boldsymbol{False}$. Atoms of any conjunct (except True and False) may contain variables, constants and function symbols.

Definition 1.1 Let $\bar{x}$ be a set of variables (possible empty set) and A be a conjunct. Well-formed formulas of the PCF language are defined as follows:

An expression $\exists\bar{x}: A$ is a ∃-formula; an expression $\forall\bar{x}: A$ is a ∀-formula.

Let $G_1, \dots, G_k$ be a ∃-formulas, then ∀-formula is a formula of the following form: $\forall\bar{x}: A(G_1, \dots, G_k)$.

Let $G_1, \dots, G_k$ be a ∀-formulas, then ∃-formula is a formula of the following form: $\exists\bar{x}: A(G_1, \dots, G_k)$.

An expression is a well-formed PCF if it is constructed based on above Definition 1.1 rules only.

The variable from $\bar{x}$ are bound by a corresponding quantifiers, and these variables are called ∀-variables and ∃-variables correspondingly.

Let us define a semantic of PCFs by a common semantics of a corresponding first-order formula (FOF) in the classical first-order logic (FOL).

Definition 1.2 Let $A = \{A_1, \dots, A_l\}$ be a conjunct and $\bar{x} = \{x_1, \dots, x_n\}$ be a set of variables. By $A^{\&}$ we denote $A_1 \& \dots \& A_l$ and $\boldsymbol{False}^{\&} = False, \boldsymbol{True}^{\&} = True$ (propositional constants). By $\boldsymbol{F}^{FOF}$ we denote an image of PCF formula F in the FOL.

If $F = \exists\bar{x}: A$ then $F^{FOF} = \exists x_1 \dots \exists x_n (A^{\&})$. If $F = \forall\bar{x}: A$ then $F^{FOF} = \forall x_1 \dots \forall x_n (A^{\&})$.

If $F = \exists\bar{x}: A(G_1, \dots, G_k)$ then

$$F^{FOF} = \exists x_1 \dots \exists x_n \left(A^{\&} \& (G_1^{FOF} \& \dots \& G_k^{FOF})\right).$$

If $F = \forall\bar{x}: A(G_1, \dots, G_k)$ then

$$F^{FOF} = \forall x_1 \dots \forall x_n \left(A^{\&} \to (G_1^{FOF} \vee \dots \vee G_k^{FOF})\right).$$

Any PCF can be obviously represented by a tree structure. Each node of that tree has the following form $\exists \bar{x}: A$ or $\forall \bar{x}: A$. Therefore, for readability we will represent PCF as a tree, as well as we will use the corresponding terminology: node, branch, root, leaf, etc.

If a PCF F roots from the $\forall: \boldsymbol{True}$ node and each leaf of F is a $\exists$-node then F is called a PCF in the *canonical form*. It is clear, that any PCF F can be converted to the canonical form by the following transformations:

1. If F is not canonical $\forall$-formula then $\forall: \boldsymbol{True}(\exists: \boldsymbol{True}(F)$ is PCF rooted from the $\forall: \boldsymbol{True}$.
2. If F is $\exists$-formula then $\forall: \boldsymbol{True}(F)$ is PCF rooted from the $\forall: \boldsymbol{True}$.
3. If F has leaf $\forall \bar{x}: A$ then the node $\exists: \boldsymbol{False}$ can be added as child of this leaf.

In the sequel, only PCFs in the canonical form are considered.

The axiom schema $\boldsymbol{AxJF}$ of the PCF calculus $\boldsymbol{JF}$ has the following form:

$$\forall: \boldsymbol{True}\left(\exists \bar{x}_1: \boldsymbol{False}(\widetilde{\Phi}_1), \dots, \exists \bar{x}_n: \boldsymbol{False}(\widetilde{\Phi}_n)\right)$$

In the PCF calculus to prove F we attempt to refute its negation, that's why the axioms are identical to false. Thus, an inference process in the PCF calculus is a *refutation* process.

Definition 1.3 Let us assume that a question $\forall \bar{y}: A$ to a base $\exists \bar{x}: B$ has an answer Θ if and only if Θ is a substitution $\bar{y} \to H^{\infty}$ and $A\Theta \subseteq B$, where H^{∞} is Herbrand universum based on variables from $\bar{x}$, constants and function symbols that occur in corresponding base subformula.

Definition 1.4 If a PCF F has the structure $\forall: \boldsymbol{True}(\exists \bar{x}: (B(\Phi), \Sigma)$, where Σ is a list of other base subformulas, and Φ is a list of question subformulas, that contains a question subformula $\forall \bar{y}: A\left(\exists \bar{z}_i: C_i(\Psi_i)\right)_{i=\overline{1,k}}$, then a formula ωF is an application result of the unary inference rule ω to the question $\forall \bar{y}: A$ with an answer Θ and $\omega F = \forall: \boldsymbol{True}\left(\exists \bar{x} \cup \bar{z}_i: B \cup C_i\Theta(\Phi \cup \Psi_i\Theta)_{i=\overline{1,k}}, \Sigma\right)$.

After appropriate renaming some bound variables inside each subformula the expression ωF will satisfy the well-formed PCF requirements.

Any finite sequence of PCFs $F, \omega F, \omega^2 F, \dots, \omega^n F$, where $\omega^n F \in AxJF$, is called a *refutation* of F in the PCF calculus.

A question with $\exists: \boldsymbol{False}$ consequent node is called *goal question*.

III. EQUALITY

A. General information

In this chapter, we will introduce the problems, which appears with classical approach to equality problem in ATP systems and will introduce new rule for working with it.

To work with the equality predicate we need to consider the following axioms of equality:

1. Reflexivity: $\forall x\ x = x$;
2. Symmetry: $\forall x, y\ x = y \to y = x$;
3. Transitive: $\forall x, y, z\ x = y\ and\ y = z \to x = z$;
4. Substitutivity (axioms schema): $\forall x, y\ x = y\ and\ A(x, x) \to A(x, y)$; A(x, x) – some PCF, A(x, y) – formula, where the x is substituted with y.

For resolution method in [3, 5] the direct usage of such axioms leads to the huge number of disjuncts most of which just pollute the logical inference and this fact leads to the inefficiency of ATP. Due to its features, the PCF (work with one question now) this problem only affects the speed of logical inference. However, some equality predicates (for example, $(a, f(a))$ can potentially lead to unlimited increase of base of facts, even using only questions-axioms of equality.

Definition 2.1 We say that E-extension of base conjunct B is any conjunct, which is gained from B only using the questions-axioms of equality.

Definition 2.2 We say, that all E-extensions of given base B, including base itself are congruent to each other. We will denote two congruent conjuncts A and B as $A =_E B$

Now, with excluding the questions-axioms of equality from PCF we will introduce new condition for answers, i.e.:

Definition 2.3 We say, that question $\forall Y: B$ to base $\exists X: A$ has the answer Θ only when Θ is a map (substitution) $Y \to H^{\infty}$ and $\exists A' =_E A : B\Theta \subseteq A'$, where H^{∞} is a Herbrand universum for corresponding base subformula.

So, by excluding the questions-axioms of equality, the new problem of searching the answers is introduced. This problem is already described in papers [3] and called equality matching.

Correctness and soundness of this approach follows from its equality to the classical inference with equality axioms.

Note, that this approach saved the following positive features of PCF:

1. Logical inference in the PCF calculus is based on adding the ground atoms (atoms, which does not contain variables) to the bases of PCF until they become contradictory.
2. The unique rule of inference.
3. Answers for given questions depends only from corresponding base and do not depend from other bases.

This is a fact, that equality matching problem with constants is undecidable [3]. We have developed the algorithm, which this solve problem for the PCF calculus with equality.

For implementation, we will use the methods and results of term rewriting systems (TRS) [7]. In particular, we will use the Knuth-Bendix algorithm for building the term rewriting system for equality in base [4].

We say that order on terms is not important. As an order, one can use the lexicographical or Knuth-Bendix order.

Definition 2.4 We say, that TRS is corresponding to some base B, if it is gained during the Knuth-Bendix algorithm with reduction order $<_{ord}$ for aggregation of equality atoms contained in B.

Definition 2.5 Terms t, e will be called congruent on base B, if $\exists l: t \to^*_S l, e \to^*_S l$, where $\to_S$ is a reducibility on TRS S corresponding base B. We denote congruence of t, s as a $t \approx_B e$.

Definition 2.6 Equality class on base B of term t is a subset $Cong_B(t) \subset H^\infty$ that $\forall e : e \approx_B t, e \in Cong_B(t)$

Definition 2.7 Term t norm on base B with corresponding TRS S is a term e, that $t \to^*_S e$ and $\forall r: e \not\to_S r$. We say that term e is in normal form $|e|_B$ on base B.

For unification implementation, we modify unification algorithm described here [5] to account for the congruency of terms.

Definition 2.8 We say, that set *P* of equations that has form $t\Theta \approx_S e$, where *t, e* are terms, *S* – reduction TRS, Θ is a some substitution is a equality matching problem with reduction. In case, if *S* is constructed on base *B* we say that this problem is an equality matching with reduction on base problem. If *e* is a ground term, then this problem is called as a ground matching problem. The set of substitutions that turns all equations into the true statements is a set of solutions $Sol(P)$.

We will describe the algorithm for ground matching problem with a reduction for some problem *P*. Such problems appears in case of subsumption problem (PCF case). This algorithm builds the set of answers *A(P)* for problem *P*:

1. The set of answers $A = \emptyset$.
2. If $P = \emptyset$, then this is the end of algorithm, return *A*. Otherwise, let's $t_k\Theta \approx_B e_k$ is a some equation from *P*.
3. If t_k is a ground term and $t_k \in Cong_B(e_k)$ is true then $P \coloneqq P/\{t_k\Theta \approx_B e_k\}$. If $t_k \notin Cong_B(e_k)$ return the error code.
4. If t_k is a variable, then $\forall\Theta \in A; A' \coloneqq A' \cup \Theta^\circ\{t_k \to Cong_B(e_k)\}\ A \coloneqq A'; P \coloneqq P\{t_k \to e_k\}/\{t_k\theta \approx_B e_k\}$, goto step 2.
5. If $t_k = f(a_1, a_2 \dots a_m)$, then construct $Prod(|e_k|_B)$. Let's d is a number of $f(c_1, c_2 \dots c_m)$ in $Prod(|e_k|_B)$, then $\forall l \in [1..d]\ \forall j \in [1..m]\ P_l \coloneqq P \cup \{a_j\Theta \approx_B c_j\}\ A \coloneqq A \times \cup_{o=1}^{d} Sol(P_o)$, otherwise $A = \emptyset$

The *Prod* is a set of terms such that $\forall e \in Prod(|t|_B) e \to_S |t|_B$.

Theorem 1. For solution of ground matching with reduction *P* by normalized base *B* the $\forall\Theta \in A(P) => \Theta \in$ KBCωFSol(P) is true.

Proof. Note, that transformation of form $f(a_1, a_2, \dots, a_m) \approx_B f(e_1, e_2, \dots, e_n)$ leads to $a_1 \approx_B e_1 a_2 \approx_B e_2 .. a_m \approx_B e_m$ and it is a correct transformation called decomposition without saving soundness in general case. To save the soundness we need to take account for the situations like $f(d) \approx_B f(c) \notin Cong_B(c)$, which handled in five step of described above algorithm. The work of this algorithm is equal to scheme building formed in the following way $\{t_1\Theta \approx_B e_1\ t_2\Theta \approx_B e_2 .. t_{i\Theta} \approx_B e_i\}$, where $t_{i,j} \in [1..i]$ are variables; and to find such substitution Θ, that satisfies all equations. The number of systems of such kind will be equal to the number of branches, linked with number of functions in $Prod(e_i)$. In addition, not all branches will have a solution. By branch, we understand the problem division described in step 5 of algorithm. Now we show that

1) The $\Theta \in A(P)$lead to $\Theta \in Sol(P)$*:*

It is obvious, because the substitutions in the set $A(P)$ is added only on 4 step of algorithm and only in the case when non-ground term – variable. So, substitution from $A(P)$ set is a ground substitution, which satisfies all matching equations (otherwise, it will be a contradiction like $t \approx_B e$, where t, e – ground terms and $t \notin Cong_B(e)$.

2) The $\Theta \in Sol(P)$ lead to $\Theta \in A(P)$:

Let's describe this part of proof for $P = \{t_\delta \approx_B e\}$. This proof can be naturally (by using decomposition) expanded to any matching problem with reduction by base *B*. By definition $\Theta \in Sol(P)$ leads $t\Theta \approx_B e$. Because e – ground term and $\approx_B$ – convergent relation, then $t\Theta \to_B |e|_B$. The one of the following cases must be handled:

1. $t \in Cong_B(|e|_B)\Theta = \emptyset$. In this case the algorithm will return $A(P) = \{\emptyset\}$, i.e. $\Theta \in A(P)$ is true.
2. t is a variable $=> \Theta \ni t \to l$, where $l \in Cong(|e|_B) = Cong_B(e)$. In this case the algorithm will return $A = \{t \to Cong_B(e)\}$, i.e. $\Theta \in A(P)$
3. $t = f(a_1, a_2, \dots a_i)$. It's obvious that $\exists f(c_1, c_2, \dots c_i) \in Prod_B(|e|_B)$ and Θ is the solution for $\{a_1\Theta \to c_1, a_2\Theta \to c_2, \dots, a_i\Theta \to c_i\}$. Doing the decomposition for $\forall l \in [1..i] c_l$ we receive reduction of this case to previous two. Because in step 5 of algorithm the branching without loss of soundness is performed, i.e. $\forall f(\dots) \in Prod_B(|e|_B)$, so it's obvious that $\Theta \in A(P)$

There are no other cases, because in every case $\Theta \in A(P)$ and the theorem is proved.

IV. Conclusion

In this paper, we have investigated the logical inference in the PCF calculus with equality without direct using the equality axioms. We have proven the correctness of the proposed solutions. This provides the possibility for solving more problems. This algorithm is implemented in

D programming language for the developed ATP system [6].

REFERENCES

[1] Vassilyev, S.N.: Machine Synthesis of Mathematical Theorems. The Journal of Logic programming, V.9, No.2–3, pp. 235–266 (1990)

[2] Vassilyev, S.N., Zherlov, A.K., Fedunov, E.A., Fedosov, B.E.: Intelligent Control of Dynamic Systems. Moscow: Fizmatlit (2000)(in Russian)

[3] Robinson, A., Voronkov, A. (eds.) Handbook of Automated Reasoning. – MIT Press, Cambridge, 2001. – 2150p.

[4] Wehrman I. Completion without failure / L. Bachmair, N. Dershowitz, and D. A. Plaisted // Chapter 1 of Vol. II of Resolution of Equations in Algebraic Structures, H. Ait-Kaci and M. Nivat, eds., Academic Press, 1989.

[5] Ch. Chen and R. Li, Mathematical Logic and Automatic Theorem Proving [Russian transla- tion], Nauka, Moscow (1983)

[6] A.A. Larionov, E.A. Cherkashin, and A.V. Davydov, "Theorem proving software, based on method of positively-constructed formulae", ;in Proc. MIPRO, 2011, pp.965-968.

[7] Baader F., Nipkow T. Term Rewriting and All That /F. Baader, T. Nipkow // Cambridge University Press, 1999

The Calculus of Positively Constructed Formulas, its Features, Strategies and Implementation

Alexander Larionov
Irkutsk State University
Irkutsk, Russia
Email: alexander@prisnif.org

Artem Davydov
Institute for System Dynamics
and Control Theory
Irkutsk, Russia
Email: artem@icc.ru

Evgeny Cherkashin
Institute for System Dynamics
and Control Theory
Irkutsk, Russia
Email: eugeneai@icc.ru

***Abstract*—Originally, the calculus of positively constructed formulas without function symbols was developed by Russian scientists S.N. Vassilyev and A.K. Zherlov by an evolutionary way in describing and solving of control theory problems. This calculus has features which are important, without loss of generality, for control theory applications. Later investigations shown that the calculus has features providing a good possibility to combine the proof procedure and the problem heuristics. Thus, calculus of positively constructed formulas can be characterized both as machine- and human-oriented. In this paper the calculus of positively constructed formulas with function symbols is considered. We provide the proofs of soundness and completeness for this calculus. Features, strategies, and some prototype prover implementation aspects are presented.**

I. Introduction

Originally [1], [2], the calculus of positively constructed formulas (PCF) was developed by Russian scientists S.N. Vassilyev and A.K. Zherlov by an evolutionary way in describing and solving control theory (CT) problems. In [2] PCF calculus is presented as first-order logical formalism, examples of CT problems described and solved by PCF calculus (elevator group control, mobile robot action planning and telescope guidance) are presented as well.

PCF calculus can be used in real-world applications due to some features allowing to combine automatic inference with special problem heuristics (formalized human knowledge and experience), therefore, PCF calculus can be positioned both as machine- and human-oriented.

However, in [2] this calculus is described only from theoretical and CT viewpoints. Aspects of automated deduction, proof strategies and prover implementation were not considered. Moreover, in [2] PCF calculus is presented without function symbols.

In this paper we present a further development of PCF calculus. We consider the following topics: PCF calculus with function symbols; soundness and completeness of PCF calculus; its features; special PCF strategies. In addition, we consider some aspects of our prototype prover implementation to show that PCF calculus is suitable for automated deduction.

The work was supported by Federal Framework Program "Scientific and Teaching Staff of Innovative Russia for 2009-2013", State Contract No.696.

II. The Calculus of Positively Constructed Formulas

The PCF calculus $\boldsymbol{JF}$ is a triple $\langle \boldsymbol{LF}, Ax\boldsymbol{JF}, \boldsymbol{\omega} \rangle$, where $\boldsymbol{LF}$ is the PCF language, $Ax\boldsymbol{JF}$ is the only axiom schema and $\boldsymbol{\omega}$ is unique inference rule of $\boldsymbol{JF}$.

A. The Language of Positively Constructed Formulas

Let us denote the set of all conjuncts as Con and assume that a *conjunct* either a finite set of usual atoms of first-order language or $\boldsymbol{F}$, where $\boldsymbol{F}$ satisfies the property $A \subset \boldsymbol{F}$ for any $A \in Con$. The empty conjunct is denoted by $\boldsymbol{T}$. It is clear that if $A \in Con$ then $A \cup \boldsymbol{F} = \boldsymbol{F}$. Atoms of any conjunct (except $\boldsymbol{T}$ and $\boldsymbol{F}$) may contain variables, constants and function symbols.

Definition 1: Let $\bar{x}$ be a set of variables and A be a conjunct. Well-formed formulas of the PCF language are defined as follows:

An expression $\exists \bar{x} \colon A$ is a $\exists$-formula; an expression $\forall \bar{x} \colon A$ is a $\forall$-formula.

Let $G_1, \ldots, G_k$ be a $\exists$-formulas, then $\forall$-formula is a formula of the following form: $\forall \bar{x} \colon A\,(G_1, \ldots, G_k)$.

Let $G_1, \ldots, G_k$ be a $\forall$-formulas, then $\exists$-formula is a formula of the following form: $\exists \bar{x} \colon A\,(G_1, \ldots, G_k)$.

An expression is a well-formed PCF if it is constructed on the basis of above Definition 1 rules only.

The variables from $\bar{x}$ are bound by a corresponding quantifiers, and these variables are called $\forall$-variables and $\exists$-variables correspondingly.

Any $\forall$-variable that does not occur in the corresponding conjunct is called *unconfined* variable.

Let us define a semantic of PCFs by a common semantics of a corresponding first-order formlua (FOF) in the classical first-order logic (FOL).

Definition 2: Let $A = \{A_1, \ldots, A_l\}$ be a conjunct and $\bar{x} = \{x_1, \ldots, x_n\}$ be a set of variables. By $A^{\&}$ we denote $A_1 \& \ldots \& A_l$ and $\boldsymbol{F}^{\&} = False, \boldsymbol{T}^{\&} = True$ (propositional constants). By F^{FOF} we denote an image of PCF formula F in the FOL.

If $F = \exists \bar{x} \colon A$ then $F^{\text{FOF}} = \exists x_1 \ldots \exists x_n \left(A^{\&}\right)$.

If $F = \forall \bar{x} \colon A$ then $F^{\text{FOF}} = \forall x_1 \ldots \forall x_n \left(A^{\&}\right)$.

If $F = \exists\bar{x}\colon A\,(G_1, \ldots, G_k)$ then

$$F^{\text{FOF}} = \exists x_1 \ldots \exists x_n \left(A^{\&} \& \left(G_1^{\text{FOF}} \& \ldots \& G_k^{\text{FOF}}\right)\right).$$

If $F = \forall\bar{x}\colon A\,(G_1, \ldots, G_k)$ then

$$F^{\text{FOF}} = \forall x_1 \ldots \forall x_n \left(A^{\&} \to \left(G_1^{\text{FOF}} \vee \ldots \vee G_k^{\text{FOF}}\right)\right).$$

Any PCF can be obviously represented by a tree structure. Therefore, for readability we will represent PCF as a tree, as well as we will use corresponding terminology: node, branch, root, leaf, etc.

If a PCF F roots from the $\forall\colon \boldsymbol{T}$ node and each leaf of F is a $\exists$-node then F is called a PCF in the *canonical form*. It is clear, that any PCF F can be converted to the canonical form by the following transformations:

1) If F is not canonical $\forall$-formula then $\forall\colon \boldsymbol{T}\,(\exists\colon \boldsymbol{T}\,(F))$ is PCF rooted from the $\forall\colon \boldsymbol{T}$.
2) If F is $\exists$-formula then $\forall\colon \boldsymbol{T}\,(F)$ is PCF rooted from the $\forall\colon \boldsymbol{T}$.
3) If F has leaf $\forall\bar{x}\colon A$ then the node $\exists\colon \boldsymbol{F}$ can be added as child of this leaf.

In the sequel, only PCFs in the canonical form are considered.

Some parts of a PCF have specially denoted by terms: the topmost (0-depth) node is called *root*; a 1-depth node is called *base* (of formula); a maximal subtree that rooted at the 1-depth node is called *base subformula*; a 2-depth node is called *question* (to its base); a maximal subtree that rooted at the 2-depth node is called *question subformula*; a maximal subtree that rooted at 3-depth node is called *consequent* (subformula). With respect to Definition 2 if an even-depth (odd-depth) node has more than one children then the node has *disjunctive branching* (correspondingly *conjunctive branching*).

Example 1: Let us consider a formula of FOL

$$F = \neg\big(\forall x\, \exists y P(x, y) \to \exists z P(z, z)\big).$$

An image F^{PCF} of F in the PCF language is

$$F^{\text{PCF}} = \forall\colon \boldsymbol{T} - \exists\colon \boldsymbol{T} \left\{ \begin{array}{lll} \forall x\colon \boldsymbol{T} & - & \exists y\colon P(x, y) \\ \forall z\colon P(z, z) & - & \exists\colon \boldsymbol{F}. \end{array} \right.$$

B. The Calculus of Positively Constructed Formulas

The axiom schema of the PCF calculus $\boldsymbol{JF}$ has the following form:

$$Ax\boldsymbol{JF} = \forall\colon \boldsymbol{T}\left(\exists\bar{x}_1\colon \boldsymbol{F}\left(\widetilde{\Phi}_1\right), \ldots, \exists\bar{x}_n\colon \boldsymbol{F}\left(\widetilde{\Phi}_n\right)\right).$$

In the PCF calculus to prove F we attempt to refute its negation, that's why the axioms are identical to false. Thus, an inference process in the PCF calculus is a *refutation* process.

Definition 3: Let us assume that a question $\forall\bar{y}\colon A$ to a base $\exists\bar{x}\colon B$ has an answer θ if and only if θ is a substitution $\bar{y} \to H^{\infty}$ and $A\theta \subseteq B$, where H^{∞} is Herbrand universe based on variables from $\bar{x}$, constants and function symbols that occur in corresponding base subformula.

Definition 4: If PCF F has the structure $\forall\colon \boldsymbol{T}\,(\exists\bar{x}\colon B\,(\Phi)\,, \Sigma)$, where Σ is a list of other base subformulas, and Φ is a list of question subformulas, that contains a question subformula $\forall\bar{y}\colon A(\exists\bar{z}_i\colon C_i\,(\Psi_i))_{i=\overline{1,k}}$, then the conclusion $\boldsymbol{\omega}F$ is the result of application of the unary inference rule $\boldsymbol{\omega}$ to the question $\forall\bar{y}\colon A$ with answer θ and $\boldsymbol{\omega}F = \forall\colon \boldsymbol{T}(\exists\bar{x} \cup \bar{z}_i\colon B \cup C_i\theta\,(\Phi \cup \Psi_i\theta)_{i=\overline{1,k}}\,, \Sigma)$.

After appropriate renaming some of the bound variables inside each subformula the expression $\boldsymbol{\omega}F$ will satisfy the well-formed PCF requirements.

Any finite sequence of PCFs $F, \boldsymbol{\omega}F, \boldsymbol{\omega}^2F, \ldots, \boldsymbol{\omega}^nF$, where $\boldsymbol{\omega}^nF \in Ax\boldsymbol{JF}$, is called a *refutation* of F in the PCF calculus.

A question with $\exists : \boldsymbol{F}$ consequence is called *goal question*. If a question has disjunctive branching then answering this question results in formula splitting.

For clarity, consider simple example of PCF refutation.

Example 2 (A refutation in $\boldsymbol{JF}$):

$$F_1 = \forall\colon \boldsymbol{T} - \exists\colon S(e)(Q_1, Q_2, Q_3, Q_4);$$

$$\begin{array}{l} Q_1 = \forall x\colon S(x) - \exists\colon A(a); \\ Q_2 = \forall x, y\colon C(x), D(y) - \exists\colon \boldsymbol{F}; \\ Q_3 = \forall x, y\colon B(x), C(f(y)) - \exists\colon \boldsymbol{F}; \\ Q_4 = \forall x\colon A(x) - \left\{ \begin{array}{l} \exists y\colon B(y), C(f(x)) \\ \exists\colon C(x) - Q_5; \end{array} \right. \\ Q_5 = \forall z\colon A(z), C(z) - \exists\colon D(f(z)). \end{array}$$

At the first step there exists only one answer $\{x \to e\}$ to question Q_1. After applying rule $\boldsymbol{\omega}$ with this answer, formula will have the following form:

$$F_2 = \forall\colon \boldsymbol{T} - \exists\colon S(e), A(a)(Q_1, Q_2, Q_3, Q_4).$$

At the second step there exists only one answer $\{x \to a\}$ to question Q_4. After applying $\boldsymbol{\omega}$ with this answer, formula is split, because Q_4 has disjunctive branching. And now formula has the following form:

$$F_3 = \forall : \boldsymbol{T} - \left\{ \begin{array}{l} \exists y_1\colon S(e), A(a), B(y_1), C(f(a)) - \left\{ \begin{array}{l} Q_1 \\ \cdots \\ Q_4 \end{array} \right. \\ \exists\colon S(e)A(a), C(a) - \left\{ \begin{array}{l} Q_1 \\ \cdots \\ Q_4 \\ Q_5. \end{array} \right. \end{array} \right.$$

At the third step first base can be refuted by answering on Q_3 goal question with $\{x \to y_1; y \to a\}$. Refuted base (base subformula) can be deleted from the list of base subformulas.

At the fourth step there exists the answer $\{z \to a\}$ to fifth new question Q_5. And now formula has the following form:

$$F_4 = \forall\colon \boldsymbol{T} - \exists\colon S(e), A(a), C(a), D(f(a)) \left\{ \begin{array}{l} Q_1 \\ \cdots \\ Q_4 \\ Q_5. \end{array} \right.$$

At the fifth step the only base can be refuted by answering on Q_4 goal question with $\{x \to a; y \to f(a)\}$.

The refutation is finished because all bases were refuted.

C. Soundness

By $\exists\bar{t}$ in the language of FOL, where $\bar{t} = \{t_1, \ldots, t_k\}$, we denote the list of quantifiers: $\exists\bar{t} \stackrel{\text{df}}{=} \exists t_1, \ldots, \exists t_k$.

To prove the soundness of PCF calculus the following statement is required.

Lemma 1: Let $F = \exists\bar{x}\forall\bar{y}(B(\tilde{x})\&\neg A(\tilde{y}, x'))$ be a formula of the FOL, where $x' \subseteq \bar{x}, \tilde{x} \subseteq \bar{x}, \tilde{y} \subseteq \bar{y}$, and $B(\tilde{x}), A(\tilde{y}, x')$ are conjuncts which contain all variables from $\tilde{x}, \tilde{y}$ and x' correspondingly. This formula is inconsistent if and only if there exists a substitution $\theta\colon \bar{y} \to H$ and $\mathbf{\Pi}_A\theta \subseteq \mathbf{\Pi}_B$, where H is the some set of ground terms, $\mathbf{\Pi}_A = \{A_1(\tilde{y}_1, x'_1), \ldots, A_n(\tilde{y}_n, x'_n)\}$ and $\mathbf{\Pi}_B = \{B_1(\tilde{x}_1), \ldots, B_m(\tilde{x}_m)\}$ are sets of all atoms in the corresponding conjuncts $A(\tilde{y}, x')$ and $B(\tilde{x})$.

Proof: ($\Rightarrow$) Let the formula F be inconsistent, then there exists an inference of empty clause $\Box$ by the unit binary resolution from the set of clauses $S = \{B_1(\tilde{a}_1), \ldots, B_m(\tilde{a}_m), \neg A_1(\tilde{y}_1, a'_1) \vee \ldots \vee \neg A_n(\tilde{y}_n, a'_n)\}$, corresponding to F, where $\tilde{a}_i, a'_j, i = \overline{1,m}, j = \overline{1,n}$ are Skolem constants. Let us consider such inference.

Since $B_1(\tilde{a}_1), \ldots, B_m(\tilde{a}_m)$ are positive unit clauses, then the inference is constructed by sequential cutting of literals firstly from $\neg A_1(\tilde{y}_1, a'_1) \vee \ldots \vee \neg A_n(\tilde{y}_n, a'_n)$, then from generated resolvents which will have at least one literal less than these resolvents on the previous step of the inference. Let $R_0 = \neg A_1(\tilde{y}_1, a'_1) \vee \ldots \vee \neg A_n(\tilde{y}_n, a'_n)$ and $R_1, \ldots, R_s$ be resolvents generated in the above inference, where $R_s = \Box$ and $s \leqslant n$. Let $\sigma_i, i = \overline{0,s}$ be the most general unifier for some literal $L \in R_i$ and some of $B_1(\tilde{a}_1), \ldots, B_m(\tilde{a}_m)$, then it is clear that the substitution $\lambda = \sigma_1 \circ \ldots \circ \sigma_s$ is such that $\mathbf{\Pi}_A\lambda \subseteq \mathbf{\Pi}_B$ and $\exists\theta\colon \mathbf{\Pi}_A\theta \subseteq \mathbf{\Pi}_B$ holds.

($\Leftarrow$) On the other hand, if $\exists\theta\colon \mathbf{\Pi}_A\theta \subseteq \mathbf{\Pi}_B$, then on each step of a resolution inference from the set of clauses $S = \{B_1(\tilde{a}_1), \ldots, B_m(\tilde{a}_m), \neg A_1(\tilde{y}_1, a'_1)\theta \vee \ldots \vee \neg A_n(\tilde{y}_n, a'_n)\theta\}$, corresponding to $F\theta$, either each literal $\neg A_i(\tilde{y}_i, a'_i)\theta$ has its complementary in the set $\{B_1(\tilde{a}_1), \ldots, B_m(\tilde{a}_m)\}$, or some part of $\neg A_1(\tilde{y}_1, a'_1)\theta \vee \ldots \vee \neg A_n(\tilde{y}_n, a'_n)\theta\}$ has factor complementary some of $\{B_1(\tilde{a}_1), \ldots, B_m(\tilde{a}_m)\}$, therefore, by the sequential cutting of literals an inference of the empty clause from S' can be constructed. Herbrand theorem implies that since S' is inconsistent, then the set S corresponding to F is inconsistent too. Therefore, F is inconsistent. ■

Theorem 1 (Soundness): Inference rule $\boldsymbol{\omega}$ is the equivalent transformation of PCFs, i.e., the following holds in any interpretation:

$$\models F \leftrightarrow \boldsymbol{\omega}F.$$

Proof: Let $F = \forall\colon \boldsymbol{T}(G_1, \ldots, G_n)$ be a PCF and $\boldsymbol{\omega}$ is applied to some base subformula G_i, then

$$\boldsymbol{\omega}F = \forall\colon \boldsymbol{T}(G_1, \ldots, G_{i-1}, G_{i+1}, \ldots, G_n, G_{n+1}, \ldots, G_{n+k}),$$

where $G_{n+1}, \ldots, G_{n+k}$ are new subformulas, which are the result of the split of G_i.

$$F^{\text{FOF}} = \boldsymbol{T} \to \bigvee_{j=1}^{n} G_j^{\text{FOF}}$$

$$\boldsymbol{\omega}F^{\text{FOF}} = \bigvee_{j=1}^{i-1} G_j \vee \bigvee_{j=i+1}^{n} G_j \vee \bigvee_{j=n+1}^{n+k} G_j$$

Since $G_{n+1}, \ldots, G_{n+k}$ are new disjunctive components at the right side of implication, then they doesn't affect the consistency (holding) of $F \to \boldsymbol{\omega}F$ with respect to the well known formula: $\vdash A \to A \vee B$, in case of $G_i \to G_{n+1}$ holds. If it is proved that $G_i \leftrightarrow G_{n+1}$ holds, then the reverse implication $\boldsymbol{\omega}F \to F$ holds too, since the subformulas $G_{n+2}, \ldots, G_{n+k}$ are analogous up to the indices to G_{n+1} by the structure.

Thus, it is necessary to be shown that

$$G_i \leftrightarrow G_{n+1}. \tag{1}$$

Let $G_i = \exists\bar{x}\colon B(\tilde{x})\,(\widetilde{\Phi})$, where $\widetilde{\Phi} = \{\Phi_1, \ldots, \Phi_k\}$ is a list of questions. Let $\boldsymbol{\omega}$ be applied to question $\Phi_j, 1 \leqslant j \leqslant k$:

$$\Phi_j = \forall\bar{y}_j\colon A_j(\tilde{y}_j, x') \begin{cases} \exists\bar{z}_1\colon C_1(\tilde{z}_1, y'_1, x'') & - \quad \widetilde{\Psi}_1 \\ \ldots & \\ \exists\bar{z}_m\colon C_m(\tilde{z}_m, y'_m, x'') & - \quad \widetilde{\Psi}_m \end{cases}$$

where $\widetilde{\Psi}_s,\ s = \overline{1,m}$ are lists of $\forall$-subformulas, $\tilde{y}_j \subseteq \bar{y}_j, y'_l \subseteq \bar{y}_j (l = \overline{1,m}), x' \subseteq \bar{x}, x'' \subseteq \bar{x}$. If there exists a substitution $\theta\colon \bar{y}_j \to H^\infty$ that $A_j\theta \subseteq B$, where H^∞ is Herbrand universe of G_i, then

$$G_{n+1} = \exists\bar{x}, \bar{z}'_1\colon B(\tilde{x}), C_1(\tilde{z}'_1, y'_1, x'')\theta\ (\widetilde{\Phi}, \widetilde{\Psi}_1\theta).$$

The image of the formula being proved (1) in the FOL is

$$\exists\bar{x}(B(\tilde{x})\ \&\ \widetilde{\Phi}^{\text{FOF}}) \leftrightarrow$$

$$\exists\bar{x}\exists\bar{z}'_1(B(\tilde{x})\ \&\ C_1(\tilde{z}'_1, y'_1, x'')\theta\ \&\ \widetilde{\Phi}^{\text{FOF}}\ \&\ \Psi_1^{\text{FOF}}\theta).$$

The right side is equivalent to the following formula:

$$\exists\bar{x}(B(\tilde{x})\ \&\ \exists\bar{z}'_1(\ C_1(\tilde{z}'_1, y'_1, x'')\theta\ \&\ \Psi_1^{\text{FOF}}\theta)\ \&\ \widetilde{\Phi}^{\text{FOF}}).$$

Since $(G_{n+1})^{\text{FOF}}$ differs from G_i^{FOF} only by new conjunctive component in the affirmative part of the quantifier, implication $(G_{n+1})^{\text{FOF}} \to G_i^{\text{FOF}}$ holds. To conclude this part of the proof, it remains to show that $G_i^{\text{FOF}} \to (G_{n+1})^{\text{FOF}}$.

If G_i^{FOF} is inconsistent, then $G_i^{\text{FOF}} \to (G_{n+1})^{\text{FOF}}$ evaluates to true.

Now suppose G_i^{FOF} evaluates to true in any interpretation, and $(G_{n+1})^{\text{FOF}}$ is inconsistent, i.e., there exists an interpretation I that falsify

$$I \models (B(\tilde{x})\ \&\ \exists\bar{z}'_1(\ C_1(\tilde{z}'_1, y'_1, x'')\theta\ \&\ \Psi_1^{\text{FOF}}\theta)\ \&\ \widetilde{\Phi}^{\text{FOF}}).$$

Since G_i^{FOF} evaluates to true in any interpretation, it implies that $I \models B(\tilde{x})$, and $I \models \widetilde{\Phi}^{\text{FOF}}$ holds, therefore, $I \models \exists\bar{z}'_1((C_1(\tilde{z}'_1, y'_1, x'')\theta\&\Psi_1^{\text{FOF}}\theta))$ has to be inconsistent. Question subformula Φ_j^{FOF}, which is a component of conjunction $\widetilde{\Phi}^{\text{FOF}}$, also holds in I, i.e.,

$$I \models \left(\forall\bar{y}_j\Big(\neg A_j(\tilde{y}_j, x')\ \vee\ \mathop{\&}_{l=1}^{m}\exists\bar{z}_l(C_l(\tilde{z}_l, y'_l, x'')\ \&\ \widetilde{\Psi}_l)\Big)\right).$$

Satisfying the main condition of the inference rule $\boldsymbol{\omega}$: $\exists\theta\colon A_j(\tilde{y}_j, x')\theta \subseteq B(\tilde{x})$ implies that by the Lemma 1 the FOL formula $\exists\bar{x}\forall\bar{y}_j(B(\tilde{x})\&\neg A_j(\tilde{y}_j, x'))$ is inconsistent. But,

since $B(\tilde{x})$ evaluates to true, the formula $\neg A_j(\tilde{y}_j, x')$ has to be false. Therefore, $I \models \Big(\mathop{\&}\limits_{l=1}^{m} \exists \bar{z}_l (C_l(\tilde{z}_l, y'_l, x'') \ \& \ \widetilde{\Psi}_l)\Big)$ has to hold. And among them holds the following: $I \models \Big(\exists \bar{z}_1 (C_1(\tilde{z}_1, y'_1, x'') \ \& \ \widetilde{\Psi}_1)\Big)$.

Thus, in I elements of $\bar{b}$ can be found such that $I \models (C_1(\tilde{z}_1, y'_1, x'') \ \& \ \widetilde{\Psi}_1)\delta_1$ holds, and $I \models (C_1(\tilde{z}'_1, y'_1, x'') \& \widetilde{\Psi}_1)\delta_2$ is inconsistent, where $\delta_1 = \{\bar{z}_1 \to \bar{b}\}$, $\delta_2 = \{\bar{z}'_1 \to \bar{b}\}$. This contradiction proves the theorem. ■

D. Completeness

The following two theorems prove the completeness of the $\boldsymbol{JF}$ calculus. The proofs is not presented in this paper due to space constraints. The basis of the proof provided in [2], where we prove the completeness of the calculus without function symbols. To prove the completeness of the calculus with the extended language it is necessary to consider the appropriate substitutions, and as area of interpretations will be used the Herbrand universe.

Theorem 2: If there exists a refutation of PCF F in $\boldsymbol{JF}$ then F is inconsistent.

The Theorem 2 claims that, if for the negation of a FOL formula A being checked for refutability after converting it to the language $\boldsymbol{LF}$ an inference found in $\boldsymbol{JF}$, then A is a theorem of FOL.

However, there is a problem: if A is some valid FOL formula, whether there always exist an inference for its negation (after converting it to the language $\boldsymbol{LF}$)?

The problem is solved by proving the following theorem.

Theorem 3: If PCF F is inconsistent, then there exists a refutation F in $\boldsymbol{JF}$.

III. Features of PCF Calculus

1) Any PCF has a *large-block structure* and only *positive quantifiers* $\exists$ and $\forall$.
2) PCF contains both $\exists$ and $\forall$ quantifiers, but structure of PCF is *simple*, *regular* and *predictable* due to predictable quantifiers alternation $\exists$ and $\forall$ in all the branches.
3) There is no need to preprocess (Skolemization) of the original formulas by elimination of all existantial quantifiers. The Skolemization releated to this elimination leads to elevating the complexity of terms. Aslo, this feature makes PCF calculus more human-oriented.
4) "Theoretical" quantifiers $\forall x$ and $\exists x$ usually not used in the formalisation of human knowledge, typical quantifiers $\forall x(A \to \sqcup)$ and $\exists x(A \& \sqcup)$ [5] are used instead [2].
5) The PCF calculus inference rule is unique and this property make PCF calculus more machine-oriented. In addition, inference rule has a large-block structure that make PCF calculus more human-oriented.
6) The proof procedure focuses on the root surroundings of the PCF. This is possible due to features 1, 2.
7) The proof procedure can be represented in terms of the *question-answering* procedure instead of technical terms of formal deducibility (i.e., in terms of logic connectives, atoms, etc.). Base canjunct can be interpreted as *base of facts*.
8) There is natural *OR-parallelism*, because the refutation of base subformulas are executed independently of one another.
9) Due to 1, 2, 6, 7 the proof procedure is *well compatible with heuristics*, as well as to a general proof control heuristics. Due to feature 5 proof consist of large-block steps, and it is well *observable* and *controllable*.
10) Due to features 7, 9 the proof obtained are quite interpretable by human. This interpretability of proof is quite important from the viewpoint of human-machine applications. Thus, as mentioned above, PCF calculus are not only machine-oriented, but also human-oriented.
11) Semantic of PCF calculus can be changed without any modifications of axioms and inference rule $\boldsymbol{\omega}$. Such modifications are implemented merely by some restrictions to application $\boldsymbol{\omega}$ and allow us to transform classical semantics of PCF calculus in non-monotonic, intuitionistic, etc. Examples of using such semantics are provided in [2].

These features prepossess that prover developed on the basis of PCF calculus can be quite intelligent, due to abillity of human interaction. In perspective it implies that it will be possible to incept the problem heuristics (that are not formalized in explicit form) to the prover. But first of all it is necessary to test an automated deduction ability of PCF calculus.

IV. Strategies and Prototype Implementation

To test the automated deduction ability of PCF calculus it is necessary to implement a prover. We have created a prototype implementation of PCF prover that supports strategies described in this section.

Proof States Tree.: One of the basic components of our prover implementation is a proof states tree (PST).

Each step of the inference in the PCF calculus can be considered as an addition of corresponding consequent subformulas to base subformula (facts to base; questions to question lists). Thus, a formula monotonically increasing, and splitting in case of disjunctive branching.

A proof states tree is a tree with the following properties: the root is the original PCF; other nodes are consequent subformulas (with the answer applied to it and renamed variables). In addition, the nodes contain some system information like step of inference, answer, etc. Thus, number of leafs is a current number of bases, and at current inference step a PCF denoted by path from leaf to root. This approach allows to rollback, observe proof states and share references on the terms and formulas. If a base is refuted then memory can be freed by deleting of path from corresponding leaf to the soonest branching.

In addition, in PST a data-sharing is implemented: if two base subformulas are characterized by corresponding paths, then this bases sharing data that are represented by same part of these paths.

PST for the formula from Example 2 represented in Fig. 1. PST root is an original PCF F_1. Node "2" is a Q_1 question consequent $\exists$: $A(a)$, and path from node "2" to the root node

corresponds to PCF F_2. Nodes "3" and "4" are corresponding consequentes for Q_4. Path from node "3" to the root node and the path from node "4" to the root node correspond to base subformulas for PCF F_3. For example, formulas that are denoted by paths "5"-"1" and "3"-"1" share data that are represented by nodes "1"-"2". When base subformula that is represented by path "3"-"1" is refuted, we can delete path from node "3" to soonest branching — in this case only node "3" is deleted, because nodes "2"-"1" still used for representing another base subformulas.

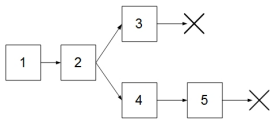


Fig. 1. Proof states tree for PCF from Example 2.

Lazy concretization.: Unconfined variables can be substituted by any term from corresponding H^∞. In general case, set H^∞ is infinite set, there is a problem what answer should be selected exactly. We suggest lazy concretization strategy. An unconfined variable has been bound with unspecified Herbrand element h that can be concretized when needed. Let us consider the following example.

Example 3:

$$\forall\colon \boldsymbol{T} - \exists\colon \boldsymbol{T} - \begin{cases} \forall x\colon \boldsymbol{T} - \exists\colon A(x) \\ \forall x\colon A(f(x)) - \exists\colon B(f(x)) \\ \forall\colon B(f(a)) - \exists\colon \boldsymbol{F} \end{cases} \qquad (2)$$

At the first step there exists an answer $\{x \to h_1\}$ to the first question, x is the unconfined variable and h_1 is unspecified Herbrand element. After this step the atom $A(h_1)$ is added to the base. At the second step there exists an answer $\{x \to h_2\}$ to the second question and h_1 is concretized to $f(h_2)$, after this step $B(f(h_2))$ is added to a base. Finally, at the third step there exists trivial answer to third question and h_2 is concretized to a.

Strategy "k,m-condition".: The answer to a question is valid if at some point of the following k steps of inference a certain condition becomes true at least m times. We use this strategy in the following cases:

1) If there exists an answer to a question with disjunctive branching, then the answer is valid if during next k steps at least m generated bases was refuted, else the process restarted with a new answer. This strategy allows to constraint a strong (highly) spawning of a base subformulas by means of control answering question with disjunctive branching.
2) The answer is valid if during next k steps at least m unspecified Herbrand elements will be concretized.

In the prototype implementation of this strategy a PST is used for potential rollbacks.

Parallel strategy.: In order to refute the original PCF, it is necessary to refute each of the bases. Note, that the refutation of each bases is mutually independent because bases contain only ground terms, and its refutation state depends only on their own questions. Thus, a base refutation procedure can be performed in an individual computation process (thread). Since refutation procedure is time consuming, we recommend to send this tasks on individual processor cores.

In addition to the OR-parallelism, a parallel schemes of algorithms for searching answers for each questions can be constructed. Since questions do not contain same variables, the answer search procedure can be performed in an individual computation process. And like in the previous strategy, second parallel strategy formulated as follows: each answer search procedure to a question is independent.

Equations.: To process equations it is very inefficient to use equality axioms (EA). For example, if formula contains binary function f and binary atom A, then in the PCF language EA are defined as the following formulas:

$$\forall x\colon \boldsymbol{T} - \exists\colon x = x$$
$$\forall x, y, u, v\colon x = y, u = v - \exists\colon f(x, y) = f(u, v)$$
$$\forall x, y, u, v\colon x = y, u = v, A(x, y) - \exists\colon A(u, v)$$

Thus, for any function symbol and any atomic symbol that occurs in PCF there exists corresponding EA question subformula. The explicit usage of EA question increases the number of inference steps and generates many extra facts in the base.

In the classical approach with respect to Definition 3 a substitution θ is an answer to a question if and only if $A\theta \subseteq B$ where A is a question conjunct and B is a base conjunct. Finding of answers is a subsumption problem. For solving this problem usually matching algorithm is used. For terms t and s matching algorithm finds θ that $t\theta = s$. In the case of the PCFs we use ground matching algorithm, because bases contain only ground terms.

To solve ground matching problem without explicit using of EA question subformulas we have to solve *ground matching with equations* problem, which is defined as follows [6]: for a set of ground equations $E(B)$, a ground term t and a term p possibly containing variables it is necessary to find the set of substitutions θ, modulo $E(B)$, over the variables in p, such that $E(B) \models t = p\theta$. By we $E(B)$ denote a set of all equations in base B. Two substitutions are equivalent if their right hand sides are pairwise congruent modulo $E(B)$.

To find some answer, modulo $E(B)$, we use an algorithm that directly uses methods of term rewriting systems theory [4].

V. Conclusion and Future Work

In this paper we completed the investigation of the PCF calculus by introducing function symbols to it and proving soundness and completeness of the calculus. In addition, main features of the PCF calculus are considered.

Implementation and testing of our prototype PCF prover confirmed that the PCF calculus is **suitable** for **automated** deduction. Further improvement of performance can be achieved, for example, by the indexing [3] or perfect data sharing techniques. However, we will focuse on infrastructure for human interaction with prover and inception of specified problem heuristics, i.e., our aim is the development and improvement of intelligent capacities of our prover.

REFERENCES

[1] Vassilyev, S.N., *Machine Synthesis of Mathematical Theorems*. The Journal of Logic programming, V.9, No.2–3, pp. 235–266, 1990.

[2] Vassilyev, S.N., Zherlov, A.K., Fedunov, E.A., Fedosov, B.E., *Intelligent Control of Dynamic Systems*. Moscow: Fizmatlit, 2000. (in Russian)

[3] Sekar, R., Ramakrishnan, I.V., Voronkov, A., *Term Indexing*. In: Robinson, A., Voronkov, A. (eds.) *Handbook of Automated Reasoning*, pp. 1853–1964. MIT Press, Cambridge, 2001.

[4] Baader, F., Nipkow, T., *Term Rewriting and All That*. Cambridge University Press, Cambridge, 1988.

[5] Bourbaki, N., *Theory of Sets*. Paris, Hermann, 1968.

[6] Moura, L., Bjorner, N., *Efficient E-matching for SMT Solvers*. Proceedings of th 21st international conference on Automated Deduction: Automated Deduction, Bremen, Germany July 17–20, pp.167–182, 2007.

Logical programming and data mining as engine for MDA model transformation implementation

A.Y. Sokolov*, I.N. Terekhin**, E.A.Cherkashin* **, S.O. Butakov***, A.A. Larionov***
* Institute of System Dynamics and Control Theory SB RAS, Irkutsk, Russia
** National Research Irkutsk State University, Irkutsk, Russia
*** Institute of Mathematics, Economics and Informatics of Irkutsk State University, Irkutsk, Russia
{alexanderysokoloff, i.terhin, s.o.butakov}@gmail.com, eugene@irnok.net, ale@prisnif.org

Abstract - In this paper, we present the description of our extension to Model Driven Architecture (MDA) methodology using the logical programming paradigm and also a practical example of application of our extension to the development process. Then we consider an induction approach of ontology from the revision control system repository. This ontology can be used in the development process as a platform independent model.

I. INTRODUCTION

Information systems (IS) frameworks are constructed on the base of a common scheme, where IS consists of at least of the following three subsystems:

- Data Warehouse (Storage) provides persistent data layer for program objects, storage formats, and productive access to the stored data.
- Application Control Layer, which usually referred to as business-logics layer; the layer mainly realizes changing the warehouse data, providing the soundness with respect to domain.
- User Interface represents stored and processed data for users and propagates events initiated by users to application control layer.

Most of the information systems also have analytical and report generation subsystems.

The growing complexity of software system leads to the creating variety of methods for automatic information system generation. One of the most applicable systems is CASE systems. They allow generating and supporting the implemented information system. Model Driven Architecture (MDA) [1] is a further development of CASE-system aimed at providing solution for the following problems of IS development:

- Rapid development of software construction technologies and programming techniques results in frequent change of software development platforms and accumulation of legacy source and binary code.
- Necessity to support a number of parallel versions of the software on various hardware platforms and operating systems; for example, most of popular Internet services have applications for mobile platforms (iOS, Android).
- Reuse of models and corresponding implementation source code in new projects, and accumulation of formalized knowledge on designing and implementation of IS subsystems.

Key concepts of MDA are CIM, PIM, PSM, and PDM.

Computation Independent Model (CIM) reflects software's external requirements, e.g., component interfaces. CIM hides internal structural elements, and therefore can be used to define specifications and checking requirements.

Platform Independent Model (PIM) is a model of the software reflecting most of the structural and some semantic aspects of the software, but this model contains no information about implementation of the structures on target program architecture. UML Class Diagram extended with some tag values and additional stereotypes is a relatively common example of PIM. The extension (marking) allows one to denote implementation nuances for structures.

Platform Specific Model (PSM) is a model, which can be directly implemented as a source code of the subsystems, e.g., it could be a physical structure of a relational database, which is translated into DDL SQL requests by means of code templates.

At the very early stages of the software development process an idea of a program is proposed. On the beginning steps basic terms and functionality requirements are iteratively collected, which correspond to MDA's CIM. The next step is a requirements analysis, which results in forming a general project outline corresponding to PIM. The developer's design activity results in detailed system design, i.e., PSM, followed by the implementation stage of the software components. MDA is a methodology for developing software by means of partial automatic source code generation of IS from visual models, so it can be considered as an approach to generative programming [3, 4].

The main distinction of MDA from the CASE-systems is that the code generation routines are not fixed and can be extended and adapted to the project requirements and developers' way of structures and functions implementation. It can, e.g., be adapted to describe even IS based on component architectures. The transformation of the PIM into PSMs is carried out under control of a

The work was supported by Federal Framework Program "Scientific and Teaching Staff of Innovative Russia for 2012-2013" State Contract № 14.B37.21.0365

Platform Description Model (PDM). PDM contains information and algorithms of PIM's structure analysis and generation of corresponding data structures in PSMs. Sometimes PSM is understood as specific variant of PIM.

Most of the entities of the UML-model are entities of ontology model (hereafter "ontology") of the domain. Usually the ontology is not formally presented during the development of the software, but informally it is the basis of developer-to-developer and developer-to-customer communications. The ontology can be the basis of code inheritance between various, e.g. parallel, projects.

There are various sources of the ontology model entities and relations. For example, revision control systems (RCS) of source code support a network of revisions marked by commit messages. One of direction of our research is to investigate the possibility to extract the ontology basis from a RCS, namely GIT system. We will describe the general scheme of thesaurus construction in further.

The extracted ontology is to be further developed in the life cycle of the software. In essence an ontology modification sometimes implies also the software modifications, thus, the ontology structure affects the model of the software.

The aim of the research is to develop set of tools of model transformation allowing one to pass pieces of models and source codes (software complexes) between projects, thus, rising level of subsystems' representation, support programming by analogy. In addition, we were interested in construction of thesaurus for domain, because it can be used as a PIM in MDA methodology, so general approach to such a construction is also in priority.

II. TRANSFORMATION IMPLEMENTATION

There are a number of approaches to MDA-transformation implementation, and the most popular are *algorithmic* approach, where all the transformation procedures are implemented with an imperative programming language; *XSLT transformations* allow representing transformation as production rules; *graph theory and graph transformation*; usage of *domain specific language* [2]. We use logical approach to define transformation as a set of productions and a pattern-directed inference engine [3] and a transformation scenario.

Patterns are represented as a mix of Prolog and Python pieces of code. Pattern query is a Prolog rule located in so-called `__doc__`-strings of Python instance methods. The bodies of the instance methods are execution parts of the patterns. Parameters passed to the methods are results of corresponding pattern queries inference. Instances itself are modules, i.e. a set of patterns and algorithms that transform part of PIM into a part of PSM. Consider the following example of patterns which are supposed to recognize and generate an SQL query to create relational database table for storing object instances of IS under development.

```
class RulesMixing: # Set of patterns
    def rule_primitive_class(self, cls, oid, oidType):
       # figure out the basics of relation coding
       # in relational tables
        """ % this starts __doc__ string
        primitive_class(Cls, OIDAttr, OidType):-
            element(Cls, 'Class'),
                \+stereotype(Cls, 'abstract'),
           stereotype(Cls, 'OODB'),
            \+internal_only(Cls),
                stereotype(Cls, 'primitive'),
            attribute(Cls, OIDAttr),
            type(OIDAttr, OidType),
            stereotype(OIDAttr, 'OIDkey'),!.
        """ # this ends __doc__ string
        self.BASE_CLASS = cls
         # we found the root of the class hierarchy
        self.BASE_CLASS_NAME = self.getName(cls)
        self.OID_NAME= self.getName(oid)
       # attribute for object reference
        self.OID_TYPE = self.coerceAttrType(oid, oidType)
        # choose a type for the object reference
        return self.BASE_CLASS_NAME, cls, oid,
                self.OID_NAME
        # the values are passed to a for statement.
    def rule_persistent_class(self, cls):
        # a class is persistent if its instances
       # are to be stored and it is not a type.
        """
        persistent_class(Cls):-
            element(Cls, 'Class'),
                \+stereotype(Cls, 'abstract'),
            stereotype(Cls, 'OODB'),\+internal_only(Cls).
        """
        # this pattern has no body
    . . . . . .
class SQLTranslator(Translator, RulesMixing): # a module
   # it generates SQL-script of database structure
   def genClass(self, cls):
        # Generate SQL-script for a class cls
        answer = [] # list of source lines
        if cls in self.generated:  # is it already
            return answer          # generated?
        for _, parent in self.query('class_parent',
                (cls, '#')): # generate all ancestors
        . . . . . .
        name=self.getName(cls)     # name of the class
        doc=cls.getDocumentation() # documentation
        if doc:                    # is it not empty?
            answer.append('/*\n%s\n*/' % doc)
        attribs = self.genSchema(cls)
       # generate table attributes
        if not self.isEmpty(attribs):
            answer.append("CREATE TABLE %s (" % name)
            answer.append(attribs) # table attributes
            answer.append(")%s;" %
                self.getTableType(cls))
        else:
```

```
        print "The class has no attributes."
    self.addFact("oodb_table('%s', '%s')" %
            (cls.getId(), name)) # assert conclusion
    # on a relation of class to table
    self.generated.append(cls)
    return answer # return generated code
```

Method *genClass* is executed by a supervisor engine for each answer *Cls* of *persistent_class(Cls)* query. Structure of the base class recognized by rule *rule_primitive_class* greatly affects a way of object references representation of the rest of the relational database tables.

This approach has cumulative advantage over above-mentioned techniques: expressive production-like transformation representation, powerful imperative and retrospection abilities of Python, existing template engines used in Python web frameworks. The approach allows one relatively simple adaptation of their way of programming to the transformational engine by means of adding new rules as additional methods and replacement of old rules as redefinition of existing methods within simple object-oriented inheritance. So the transformational engine is not tied to specific set of development environments.

To transform the UML models its XMI (XML Metadata Interchange) file is loaded. This format is a kind of XML, so DOM2 API is used to access PIM's structure. XMI is a standard data format supported by various proprietary and free software technologies and libraries. The DOM2 tree is translated into Prolog facts by means of requests from patterns. Object Constraint Language (OCL) expressions are extracted from PIM and represented as syntax trees.

At first a general reasoning about object structure is carried out, basic properties are recognized. Other modules use the reasoning results to refine implementation variants of synthesized program objects. Each module specifies PIM's structures with additional facts about existing structures and creates new objects and relations. For example, SQL database transformation module merges inherited abstract part of attributes to the class and generates table description on the base of the merge. The process is controlled by scenario represented as a list of leaf nodes to be executed.

If all nodes of the scenario are executed and all their solutions (queries) are satisfied and processed, then the set of all the facts in working memory defines PSM. The source code is generated on the base of obtained PSM. A generator module executes a query and fills in a source code template with query results.

Results of transformations and code generation are combined in object libraries. Objects from libraries are used in construction of business-logics of developed IS. Programmers supplement generated code by inheriting it in new classes. This approach partially solves the problem of generated source code modification by programmers.

The XMI file can contain not only the structure information, but also some semantic values for its elements. This information used to increase the control over the transformation procedure, in particular, for filtering information by using some criteria. UML have the following semantic definition language structures:

- Stereotypes to create new elements of UML;
- Tagged values to create properties for the elements;
- Constraints to formally define logical constraints, invariant, pre- and post-conditions for a method invocation.

Let us consider a simple example. Assume that there exists a class in an UML Class Diagram that has at least a string field *name*. If we mark the class with, e.g., «Reference Book» stereotype, then all many-to-one relations to the class can be interpreted in relational database context as many-to-one relation and corresponding tables, and reference fields *ID* are generated. Having recognized the stereotype and the relation, user interface generator can construct a widget and its controller (in sense of Model-View-Controller paradigm) to select appropriate record from the reference book and store the reference book *ID* in corresponding table and object. Now, the generated code in various subsystems is logically and mutually depended.

In order to adopt the transformation engine and its knowledge base to developer's instrumental software and technologies, one imports Python module, inherits and modifies its set of patterns and generation modules, specifies new module in scenario. In the application example in the following section, we used inheritance to refine a generic SQL relational table transformation to specific properties of MySQL server.

III. THE PROJECT STRUCTURE ANALYSIS

To take advantage of information represented in ontology the ontology should be constructed at first. One way of the construction is to deduce the ontology as thesaurus, a vocabulary of terms hierarchically ordered. Programmers to describe their work in various text messages use thesaurus. For example, in an open-source software project the messages are located in RCS systems, forums, documentation and mailing lists. We offer to deduce the thesaurus and the project structure (project-subproject relationship) on the base of the branching structure of GIT commit network (repository) of the project source code. Under project structure, we consider tree-like representation of the project branching structure; the tree reflects hierarchical division of the project to inferior subprojects. We also suppose that the hierarchical structure of ontology in some extend correlate with the project structure.

The source information for the investigation is a network graph structure representing all development, branching, and merging modes of the project sources stored in its GIT repository. The network reflects a time line of the project development process. Each new branching node can be interpreted as starting or continuation point of a subproject, but many of the branches are just error fixations. Merge points can be interpreted as finishing or committing points of a subproject. Therefore, the analysis of the commit network

is, in general, a pattern recognition problem under control of heuristics. In our case, the heuristics are defined as mutual relevance of branches.

GIT is a distributed version control system developed by Linus Torvalds specially to manage the development of the Linux kernel as very complex project. GIT supports rapid split and merge versions of the source code, includes tools for visualization and navigation. GIT also supplies each developer with a local copy of the entire history of the development. The changes that made locally can be propagated from one repository to another by various ways. In addition, this system allows representing the history of the changes (the network) graphically.

Nodes of GIT network are versions of the source code and are identified by the SHA1-hashed strings and accompanied with commit messages. Every commit is a fixation of changes made in a working copy of the main repository of the project, as well as a description of the changes. In the graph representation the nodes are connected to each other, each combination of the connections form either linear intervals of the development either splitting (branching) subprojects or merging of the subprojects or their results in a superior subproject.

The branching and merging in GIT can appear as results of two different kind of programmer's activity. The first activity is cloning a repository, e.g., creating local copy of the main project repository and incorporating the changes made in the main repository. The second case, which is of our interest, is the project development itself. In this case, according to developer recommendation of GIT programmer should create a lightweight branch of the repository, develop the feature or correct an error, and then merge results to a superior branch, in particular in the main branch. In the essence, the merged results cloning and modifications are the usual case of branch formation.

Besides branching and merging nodes, the development timeline contains linear intervals. Linear intervals reflect most probably a logical sequence of the development events related to the same subproject. The same is true for a sequence of merging events coming from an auxiliary subproject branch. Branching frequently denote creation of a subproject.

To recognize branches in GIT-network an analytical program is being developed in the programming environment XSB [13]. XSB is chosen due to its powerful ability to index dynamic facts on their arguments. The idea is to develop a set of Prolog rules which will recognize template subgraphs and construct markup structure denoting for each node its role and inclusion. The source data for the analysis is a set of Prolog facts about GIT-network obtained by means of translation of the network structure.

A. Recognition of basic structures

At the first step of the structure recognition a translation subsystem transforms the structure of the project represented in a GIT repository into a text format, which can be loaded by XSB predicates *consult/1* or *load_dyn/2*, i.e., into a collection of facts. The facts in this case are the SHA1-hashed commits identifiers. The facts establish parent-child relationship between nodes. The versions in a GIT repository can be marked with tags, which are labels assigned to a specific commit of the source code. The tag label as a compound string value can express a symbolic name for a group of committed files, and a common version of each file in the group in some general versioning principle. This property of tags gives us a possibility to consider the inheritance of the individual parts of the project, and different versions of the project as well.

The following prolog sentences are examples of a GIT-network translation:

```
node('df145gt6984tgh1254fr').
node('4684dsf457fesae4452').
node('qw78er96ty54ui12opo').
node('as558df74g110hyu058').
parent('df145gt6984tgh1254fr',
       '4684dsf457fesae4452').
parent('df145gt6984tgh1254fr',
       'qw78er96ty54ui12opo').
parent('4684dsf457fesae4452',
       'as558df74g110hyu058').
parent('qw78er96ty54ui12opo',
       'as558df74g110hyu058').
```

The following clauses are basic structure recognition rules.

```
branch(Z,X,Y):-
   parent(Z, X), parent(Z, Y), X\=Z.
merge(X,Y,Z):-
   parent(X, Z), parent(Y, Z), X\=Z.
tag('v3.3-rc2', '4564dfgert84tgdre4gs').
recent('v3.3-rc2', 'v3.3-rc1').
tagged(T,R):-tag(T,R),!.
tagged(T,R):-parent(P,R), tagged(T,P).
```

The facts of the structure node/1 denote existence of the nodes in the GIT network. Facts *parent/2* denote inheritance relation between existing nodes. Facts *tag/2* associates a tag value with a node, and *recent/2* defines a versioning order between tags. Rules branch/3 and *merge/3* are base rules for recognition of branching and merging; rule *tagged/2* reveals whether a node belongs to a tagged set of commits. In order to extract the project structure it is necessary to run all recognition rules and write down all combinations of variable values obtained because of Prolog's logical inference.

The following prolog rules recognize chain of nodes corresponding to a linear development.

```
merged(X,Y):-
      merge(X,_,Y),!.
nbr(X):-
      branch(_,_,X), !, fail.
nbr(X):-
      merge(_,_,X), !, fail.
nbr(_).
strict(X,X, K, K, 1):-
      nbr(X).
```

```
strict(X,Y, K, [Z|T], N):-
      nbr(X),
      parent(Z,X),
      strict(Z,Y, K, T, M),
      N is M+1.
/*              Recognition of chains */
/*              between branch & merge.*/
chain(X,Y, [X,Z|T], N):-
      branch(Z,_,X),
      \+ foundchain(Z, _,_, _),
      strict(Z,R,[Y],T, N),
      merge(R,_,Y),
      cache(foundchain(Z,X,Y,N)).
chain(X,Y, [X,Y], 0):-
      branch(Y,_,X),
      \+ foundchain(Y, _,_, _),
      merge(X,_,Y),
      cache(foundchain(Y,X,Y,0)).
/*              -//- between merge & branch.*/
chain(X,Y, [X,Z|T], N):-
      merge(Z,_,X),
      \+ foundchain(Z, _,_, _),
      strict(Z,R,[Y],T, N),
      branch(R,_,Y),
      cache(foundchain(Z,X,Y,N)).
chain(X,Y, [X,Y], 0):-
      merge(Y,_,X),
      \+ foundchain(Y, _,_, _),
      branch(X,_,Y),
      cache(foundchain(Y,X,Y,0)).
```

Predicate *chain/4* defines relation between a starting node, an ending node, a path and the length of the path. A path is, e.g., a linear piece of GIT network, which starts from branching, ends at merging, has no branching and merging inside. Predicate *nbr/1* denotes the argument is neither branching, nor merging node; strict/5 recognizes linear subpath inside branch under recognition. Predicate (dynamic fact) *foundchain/4* stores the result of the recognition. Its arguments are start node, end node, path identifier, and path length. The path identifier is a node inside path, which distinct the path between various paths having common branching and merging nodes and the same length. Length is used to debug algorithms: we filter program output with command line *grep* utility and calculate statistics by number of processed branches according to their lengths. Auxiliary predicate *cache/1* stores dynamic fact *foundchain/4* in working memory and shows the progress of recognition.

The recognized linear parts of the GIT network form common groups of messages, that are implied to describe the programmers' activity of developing a single feature, and, thus, it is implied that the terms used describe the activity only. The messages are joined in texts, which are the source information for thesaurus recognition. Now we will describe the thesauruses recognition.

On the first step a set of independent texts are prepared. The texts are stored in fast access database. After that full text index of the stemmed words (hereafter terms) of the texts is produced. The index should contain forth and back references between terms and texts.

On the second step, the terminological basis construction algorithm looks through the texts and tries to join adjacent terms found in texts in a compound term. The full text index frequency for the compound term is figured out. The join succeeds if the frequency of the term decreases not more than in 0.7 times (the coefficient is arbitrary and subject of experiments). If an original, single term is used at least once as first word of a compound term, then it is removed from the terminological basis.

On the next step, full text index is transformed into terminological index, where the terms are mapped into their frequencies of appearances in the texts. Now the texts are represented as points in property space, where each of the property is a term of the terminological basis. If term appearances frequencies are similar for two texts then the texts are similar and located near to each other in the property space. The set of texts are processed by some hierarchical clustering algorithm, which induced the hierarchy of classes of semantically similar texts. We use agglomerative clustering methods as it has lower computation complexity with comparison to divisive methods. On the last step in each join node of cluster the terminological index is recalculated for the texts belonging to the node. The most frequent term is selected among the terms appeared in the texts. This term denotes the semantic of the node. Also among most frequent terms a set of terms can be chosen, the set will also describe the semantic of the texts. The resulting hierarchy clustering also denotes class-subclass (term-subterm) relationship between terms as required by thesaurus.

IV. CURRENT DEVELOPMENT OF THE TRANSFORMATION TOOLS

Further development of the transformation tool is aimed at raising performance and expressive abilities of the transformation system, as well as its reliability. Main problem is that it is hard to provide efficient and in the same time sound integration of Prolog and Python: both systems have its memory management units but different memory management strategies. We decided to shift the system to use powerful expressive abilities of LogTalk Prolog [7] macro package supplying it with an imperative subsystem and a template engine. LogTalk is an object-oriented logic programming language that can use most Prolog implementations as a back-end compiler and inference engine. As a multi-paradigm object-oriented language, it includes support for prototype and class inheritance, protocols/interfaces descriptions, component-based programming through category-based composition, event-driven programming, and high-level multi-threading programming.

Let's consider a code example written in the LogTalk programming language. We have the following interface class for access to the loaded XMI DOM2 tree. This class is used in Model class, that is a primarily recognize the model in XMI file.

```
:- object(class, instantiates(class)).
      :- private(attributes_list/1,
operations_list/1,
```

```
. . . . . .
parse_operations/2).
      :- public(new/5,
new/2,
name/1,
operations/1,
attributes/1).
. . . . . .
:- end_object
```

Next code is an example of a transformation procedure

```
:- object(TransformToSQL(Class, Type) )
. . . . . .
%Parametric object allows to customize some parameters
      database(Database) :-
                parameter(1, Database).
      class(Class) :-
                parameter(2, Class).
      databasetype(DatabaseType) :-
                parameter(3, DatabaseType).
gen_sql(Name, Attributes, Output) :-
append("CREATE TABLE", class::Name,
Output),
append(Output, ::gen_attributes(
class::attributes,
Attributes), Output),
append(Output, ::databasetype, Output).
. . . . . .
:-end_object
```

Call to the g*en_sql* method looks like *TransformToSQL*(Student, "Inno DB").

In the new implementation [8] of the transformation, we take advantage of the same object-oriented hierarchical modular architecture as before, but set up the following goal: to support transformation in both directions: from abstract PIM to source code and from the source code to abstract models.

In addition, we want to integrate the new scheme of code generation to the powerful existing modelling tools. The results of the investigation supposed to be realized as plugin for the Eclipse Modelling Tools. The integration engine is to be based on change propagation [9]: the modifications made are recognized by the environment and pushed to other utilities.

A. Eclipse Modeling Project

Eclipse Modeling Project (EMP) is one of the popular and evolving tools for visual modelling tools developing. The platform consists of few projects that realize the ideas in domain of visual modelling.

The core of the project is an Eclipse Modelling Framework (EMF) that is a technology for describing the metamodels of new visual languages. The EMF project contains the full powered set of tools for creation and manipulation of metamodels.

Metamodels in EMF are defined using the special language called Ecore. EMF allows defining Ecore-metamodel by the set of different methods including XMI description, XML schema, annotated Java source code, database schema, UML models, *etc*.

The projects in EMP use the power of EMF and provide the simplified mechanism for metamodels manipulations (transformations). There is a list of projects in EMP:

- Graphical Modelling Framework (GMF) allows creation graphical modeling tools for new developed modelling languages without writing any code using the EMF for storing the models.
- Model Development Tool (MDT) is oriented to realize the industry standards in field of graphical modeling, e.g., tools like UML2, BPMN2, and OCL.
- Model to Model Transformation (M2M) provides the possibility for model transformation defined by EMF.
- Mode to Text Transformation (M2t) provides the model transformation into a text representation.
- Xtext oriented for models described in text form and can be used without EMF.

In addition, there are some tools for code and documentation generation. There are three projects related to that topic:

- JET provides the code generation by defined model.
- Acceleo implements the OMG standard MOFM2T for translation the models into the textual representation.
- XPand is a static typed language that can be used for description a language of code generation.

B. Change Propagation

Definition of change propagation in [9] from the MDA point of view can be interpreted as a way of transformation. Transformation engine compares two versions of PIMs, recognizes the difference and corrects PSM and source code in the corresponding points of change. The transformation approach uses specially stored links between objects from PIM and corresponding generated object in PSM. When an object changes or deleted its image in PSM traced by its link.

We suggest extending this idea to allow doing the propagation in both directions, including from PSM to PIM. This should results in the following additional advantages:

- Record the stages of the development as complexes of abstract models and corresponding source code fragments;
- Allow designers to transfer the model complexes between projects.

To deal with stated extension and planned features we proposed to apply the theory of systems of complexes (configurations), successfully used in geography research [10], to software life cycle, implying that the software

development is a natural process. The theory shows how to represent model elements, relations between elements, transformations and links as complexes, as an element of a category. In [10] we shown that as the model complex and a change are elements of the same set of complexes, hence the change is expressed with the same language structures as the model complex. Similar arguments are true for transformation modules; they can transform both the model and its change.

Some practical examples have been found, like *diffutils* package of any Unix distribution, which demonstrate the investigation results. The package allows programmer figuring out the differences between two versions of a source code text file (patch file) and apply the patch file to the sources. The package contains two main utilities *diff* and *patch*. The first utility compares two ASCII or Unicode texts and produces a new text file with a representation of the difference between input files.

```
--- t1.cpp 2012-02-12 11:50:42.668030039 +0900
+++ t2.cpp 2012-02-12 11:52:44.944957992 +0900
@@ -2,6 +2,7 @@
      char * name;
 };

-class Pupil:Person {
-     Class * cls;
+class Student:Person {
+     char * stud_no;
+     Group * group;
 }
```

This example shows that the difference is shown with the same ASCII or Unicode characters: one character right shifts the text so the first column became control column, where characters define a modification. The space character denotes a context of the text part under change. Minus character denotes part that to be removed, and plus character denotes the new text composition to be added into the context. As in one patch file the whole project change set can be stored, special format substrings "---" and "+++" are used to denote the particular files, and "@@" is used to denote general relative position of the first line of the context in the files.

Links from PIM to PSM express context of changes more precisely than patch file format, and allow back transformation engine to recognize the differences. We suggest using an intermediate text format between PSM and generated source code. This format reflects the links and account that the generated source code in a general case is not a flat text anymore. The format can be constructed on the base of a D.Knuth's Literate programming [11]. Literate program itself is a tagged tree representation of computer program, and those tags can be generated from PIM and PSM and reflect the links and abstract objects of the model. In opposite direction the changes of the source code now can be more precisely recognized having tags at hand. There are well developed tangling and untangling algorithms (reverse direction), e.g. in [12].

V. CONCLUSION

We have considered an existing implementation of transformation engine for Model Driven Architecture approach to software development in the case of information systems. The approach is based on mixing high-level object-oriented programming language Python and logic language Prolog. The transformation is represented as network of modules.

One of the aims of the research is to construct a software development tools based on analogy. For example, having stores in a revision control systems all the states, models and stages of MDA software development as change complexes, it probably be possible to construct new sequence of differences for new original model. The further developments are described and the small review of Eclipse Modelling Tools is presented.

In addition, we considered software for a project structure analysis of a software project revision control system commit network. Logic recognition procedures have been developed and implemented as Prolog program. Procedures for ontology extraction and labeling the project tree from the commit messages have been implemented.

REFERENCES

[1] D.Frankel "Model driven architecture : applying MDA to enterprise computing". - New York: Wiley, 2003.

[2] Si Alhir. Understanding the Model Driven Architecture (MDA) URL: http://www.methodsandtools.com/archive/archive.php?id=5. (access date-29.09.2012)

[3] M.M.Gorbunov-Posadov, "The way to grow a program", Open Systems Journal, 2000, N. 10, pp. 43–47. (in Russian) [English version URL: http://www.keldysh.ru/pages/softness/growE.htm (access date-29.09.2012)]

[4] K.Czarnecki, U.Eisenecker. Generative Programming: Methods, Tools, and Applications, Addison-Wesley, Reading, MA, USA, 2000. 864 p.

[5] V. Zakharova, A.V. Melnikov, J.A. Vokhmitsev "An approach to automated ontology building in text analysis problems". Workshop on computer Science and Information Technologies CSIT'2006, Karlsruhe, Germany, 2006. P.177-178.

[6] "Ontologies and Thesauruses – Institute of Distance Education INTUIT", http://www.intuit.ru/department/expert/ontoth (in Russian)

[7] P.Moura "Logtalk: Design of an Object-Oriented Logic Programming Language". PhD thesis. Universidade da Beira Interior, 2003. [See also, "Logtalk. Open source object-oriented logic programming language". URL: http://logtalk.org/ (access date-29.09.2012)]

[8] E. A. Cherkashin, S. A. Ipatov. "Logical Approach to UML-model processing of Informational Systems". Journal of Conterporary Techologies. System Analysis. Modelling. 2009. N 3 (23). pp. 91–97. (in Russian)

[9] M. Alanen, I. Porres, "Change Propagation in a Model-Driven Development Tool," in Presented at WiSME part of UML 2004, 2004.

[10] E.A.Cherkashin, V.V.Paramonov, et al, "Model Driven Architecture is a Complex System", E-Society Journal Research and Applications. Volume 2, Number 2, 2011, pp. 15-23.

[11] D.E.Knuth. "Literate programming". The Computer Journal, 27(2), 1984, pp. 97-111.

[12] E.K.Ream. "Leo's Home Page" URL: http://webpages.charter.net/edreamleo/front.html (access date-29.09.2012)

[13] D.S.Warren, "Programming in Tabled Prolog (very) DRAFT" Stony Brook, NY 11794-4400, U.S.A. (unpublished) URL:

http://www.cs.sunysb.edu/~warren/xsbbook/ (access date – 29.08.2012

Building Process of SCT Generators

D. Radošević, I. Magdalenić, and T. Orehovački
University of Zagreb, Faculty of organization and informatics, Varaždin, Croatia
danijel.radosevic@foi.hr; ivan.magdalenic@foi.hr; tihomir.orehovacki@foi.hr

Abstract - This paper presents the building process of generators based on the SCT dynamic frames model. The SCT generator is defined by its Specification (S), Configuration (C) and Templates (T). These elements are represented graphically by Specification diagram and Configuration diagram. The building process starts by initial steps and continues by a spiral application development, based on Boehm's spiral model of software development. To evaluate the proposed building steps, an example of building SCT generator was given. The example uses the software tool developed with a purpose to implement the building steps of SCT based generators.

I. INTRODUCTION

The paper introduces a method of building code generators according to the SCT generator model [1]. The method is supported by our generator development tool, named as *Generator builder*.

Development tools for building of applications that are based on the principles of generative programming is a real need. A development tool makes it easy to build applications and prevents the appearance of typical errors in the design as described in [2] and in sintax [3]. Applications based on generative programming are special in that they usually use a domain-specific language for defining the characteristics of the application. A usage of domain specific language introduces a new level where errors can occur. To minimize errors and to help generator developers we have already created error messaging tool described in detail in [4]. Error messaging tool serves to detect typical errors that occur in phase of the source code generator creation. Defects in the development process can be avoided by following good practices and by using of an appropriate development tools. There are several already tried and tested techniques of generative design as it is shown in [5]. This paper presents a new technique of generative design that is specially adapted to SCT generator model [1]. In this paper are described some typical steps in building process of generators based on the SCT dynamic frames model. For that purpose we have developed a graphic development tool which implements this process. The process of building of the source code generator is the main step in building of generative applications. The inputs to this process are previously created prototypes of applications on the one hand and the collected knowledge about user requests and desired application features on the other. The generator developer creates Specification and the tool assists him in creation of appropriate source code templates and Configuration rules.

The remainder of the paper is structured as follows. Section 2 provides literature review of the most prominent techniques aimed for generating of programming code. Features of the SCT generator model are described in the third section. Building steps of SCT generators are explained in the fourth section. Discussion on design issues in defining generator building steps is offered in the fifth section. Section 6 gives conclusions.

II. BACKGROUND TO THE RESEARCH

Software product lines (SPL) emerge during development of multiple versions of the same software system for different types of users. The objective of the SPL approach is to lower development expenses and concurrently increase software quality and productivity. Instead of developing each software system from the scratch, SPL approach supports reuse of product line assets such as the domain model, product line architecture, and generic components [6]. Although SPL members have many characteristics in common, they differ in certain users' requirements and implementation details called variants. The set forth variability in the SPL domain could result in a large number of possible variant combinations thus making them difficult to handle. Variability mechanisms that support automated composition and adaptation of reusable SPL assets represent an effective way of addressing the problem of managing variants. This section offers a brief overview of automated code generation techniques for handling variants in the SPL approach.

GenVoca [7] is a domain independent software development model aimed for generation of hierarchical SPL families. Fundamental features of GenVoca are realms composed of plug-compatible layers, type equations, and symmetric layers [8]. Standardized interfaces called virtual machines are a set of classes, their objects, and methods used for the implementation of SPL functionalities. Layer or component represents an implementation of particular virtual machine. If a set of layers implement the same virtual machine, then they constitute a realm or library. Each layer exports the virtual machine of the realm to which it belongs and imports interface of the realm of which parameter it contains. A particular layer is symmetric if it exports and imports the

same virtual machine. Layers encapsulate transformation that maps operations and objects between export and import virtual machines. A named composition of layers used for modeling a particular software system is called a type equation. Realms and their layer define a grammar whose sentences are SPL members.

A modular framework which enables model-driven development of SPLs is called openArchitecturWare (oAW)[9]. The essential part of the oAW is a workflow engine that enables the definition of transformation workflows together with prebuilt workflow components meant for instantiation of models, validating their semantic correctness, their transformation into other models, and eventually code generation [10][11]. Workflow components are XML files which specify steps that need to be executed in a generator run. oAW has built-in support for operating on UML, UML2, EMF/Ecore, XML, Visio, and JavaBeans-based metamodels [12]. For semantic validation of models based on the set forth metamodels, oAW offers a declarative constraints checking language called Check where tests are specified as First-Order Logic formulas. After checking model constraints, valid models are combined into one model by employing model-to-model (M2M) transformations implemented using a textual language called Xtend. Result of the M2M transformation is used as a starting point for code generation which is done using an object-oriented template language called Xpand. More recently, openArchitecturWare has moved to the Eclipse Modeling Project [13].

XML-based Variant Configuration Language (XVCL) [14] is a general purpose template language based on Bassett's frame technology [15]. XVCL works on the principle of composing custom artifacts (e.g. code) from a base of generic, adaptable, and reusable domain product line assets called meta-components or x-frames. Meta-components are XML files where variation points are instrumented by XVCL commands thus facilitating automatic customization and evolution. Normalized layered hierarchy architecture of meta-components is called an x-framework. The specification frame (SPC) is the topmost x-frame which manages composition and adaptation process of a product line member. Starting with the SPC, the XVCL processor traverses an x-framework, interprets XVCL commands embedded in visited x-frames, and by conducting necessary adaptations assembles components of a specific system.

With an aim to overcome the barriers of the SPL approach adoption, Heradio et al. [16] have proposed the exemplar driven development process (EDD). The concept underlying EDD is analogy oriented development based on the similarities among domain products. EDD process starts with an implementation of domain exemplar which represents an intersection of all the product requirements within domain. Requirements which are out of the intersection scope are formally defined during the exemplar flexibilization phase. Exemplar flexibilization results in domain specific language (DSL) compiler which is used during application engineering phase for automatic generation of software products.

Similarly as XVCL, our approach is frames based. The essential difference is that SCT generator model is based on dynamic frames [1] that are created during the generation process while XVCL is based on static frames. In addition, the generator building process proposed in this paper starts from application prototype which is analogous to a development process suggested by Heradio et al. [16].

III. GENERATOR OVERVIEW

Generators are meant for generation of code artifacts in diverse programming languages. Heretofore, they have been used in the development of web applications [17][18], web services [19][20], as well as for generation of lab based assignments [21]. The SCT generator model [1] defines the generator of source code from three core elements: Specification (S), Configuration (C) and Templates (T). Specification contains the features of the generated application in form of attribute-value pairs. Templates contain source code in a target programming language together with connections (replacing marks for insertion of variable code parts). Configuration defines the connection rules between Specification and Templates. All three model elements together constitute the SCT frame.

A particular SCT frame produces source code that could be either stored in a specific data file or included in another SCT frame. The basic idea of the generation process is shown in Figure 1. The initial SCT frame contains the initial source code template that includes connections. Source code template is file that contains source code and connections. Connections used in code templates define inclusion of content that can be from another code template, or source, if code template is omitted. Each connection has to be replaced with other source code template or value from Specification during the process of source code generation. The source code generator creates a new SCT frame for each connection. The source code of SCT frames located deeper in the hierarchy is included as the integral part of its superior SCT frame. The source code of the initial SCT frame is stored in a data file.

Since an average application contains more data files, the SCT model implies the existence of Handler. The Handler is the part of the SCT source code generator which aims to make the generator scalable in a way that it can produce more pieces of program code (e.g. program files) from the same set of Specification, Configuration and Templates. The SCT dynamic frames model enables the generation of various program units (e.g. files, classes, functions etc.) from the same Specification. Moreover, it enables the generation of different types of code e.g. JavaScript, PHP, XML, Python, Java, etc.

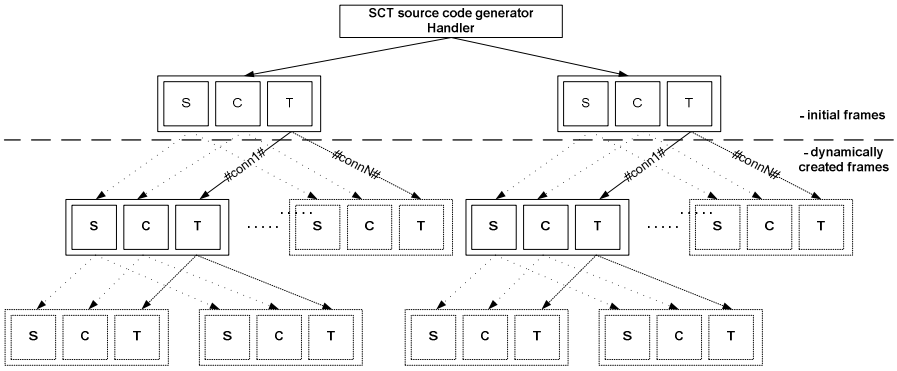


Figure 1. The generation process

All three constitutive elements of SCT model are presented in case study of application which aim is dealing with dynamic list in C++. The same application is used in next section for description of building steps of a generator.

Elements of Specification are *attribute-value* pairs as shown in following example:

```
OUTPUT:out1
out1:output/Linked_list.cpp
field_int:student_id
field_char:surname_name
field_int:year_of_study
field_char:note
```

This Specification defines main features of application. The application is going to be generated in program file named *Linked_list.cpp*. Each dynamic list element consist of four attributes and their data types: *student_id* with data type *integer*, *surname_name* with data type *char*, *year_of_study* with data type *integer*, and *note* with data type *char*.

Configuration consists from Configuration rules defined by three elements: Connection, Source, and Code Template. The following example of Configuration defines 5 rules, where each rules is defined in one line.

```
(1) #1#,,main.template
(2) #fields_declarations#,field_*,field_*.template
(3) #data_entry#,field_*,data_entry.template
(4) #field#,field_*
(5) #print_data#,field_*,print_data.template
```

The first rule defines the initial source code template *main.template*. The second rule defines replacement of connection *#fields_declarations#* with template *field_*.template* for each occurrence of attribute *field_** in Specification. For example, for attribute *field_int* is used template *field_int.template*. The third rule defines replacement of connection *#data_entry#* with template *data_entry.template* for each occurrence of attribute *field_** in Specification. Similar functionality has the fifth rule. The fourth rule defines replacement of connection *#field#* with value from specification. For example, *field_int* is replaced with *student_id* in first occurrence of this connection in source code template.

Templates are program code fragments which contain connections. Typical application is generated using several templates. For example, a template *data_entry.template* defines part of application that deals with input of user data.

```
cout << "#field#: ";
cin >> new_element->#field#;
```

The presented template has two connection *#field#* which are replaced with values from Specification. This template is used for each occurrence of attribute *field_* in Specification as defined by third rule in Configuration. The final code of that part application looks as follows:

```
cout << "student_id: ";
cin >> new_element->student_id;
cout << "surname_name: ";
cin >> new_element->surname_name;
cout << "year_of_study: ";
cin >> new_element->year_of_study;
cout << "note: ";
cin >> new_element->note;
```

All templates of example application are available online[1].

IV. BUILDING STEPS OF A GENERATOR

The process of building new generators starts from application prototype, similar to approach described in [16]. The prototype is being transformed/ decomposed into SCT model elements through several steps described in this section. The SCT generator uses these elements in automatic assembling of different application variants, which is shown on example of program in C++ that deals with simple linked list. Program variations refer to different structure of linked list element, as shown in Figure 2.

[1] SCT Generator Builder example
http://gpml.foi.hr/SCT_Generator_Builder/

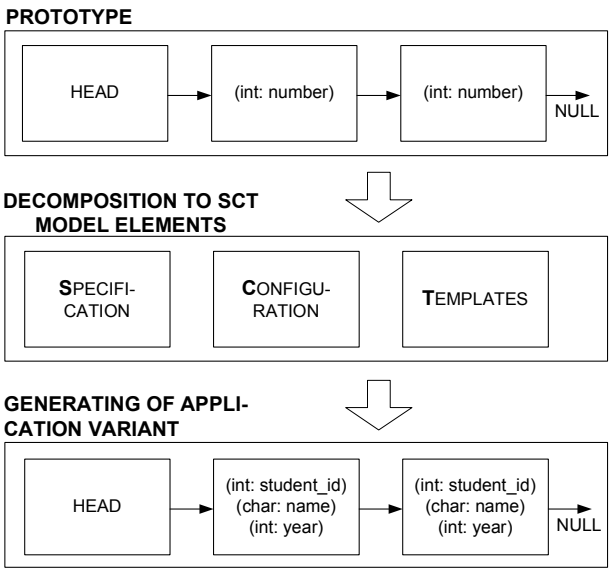


Figure 2. Transformation of prototype into application variant

Steps in a process of generator building, including example:

0. Prerequisite. The prerequisite for building of SCT generator is the application prototype in some textual form. This includes different types of programming code, regardless of programming language, mark-up code like HTML and XML, documentation text etc. The prototype is being relaxed through the next step in order to enable generation of variants.

1. Choosing new main templates and output types. Templates are program code fragments that contain *connections* (tags for further inclusion of code). The main templates are specified in the initial part of Configuration and define the type of code to be generated [1], e.g.:

#1#,,main.template

Each main template from Configuration is connected to corresponding output type in Specification. Output types are defined in the beginning of Specification, usually together with the names of output files to be generated, e.g.:

OUTPUT:out1 - *output type*
out1:output/program.cpp - *output file*

2. Creating of Specification. Specification consists from *attributes* and their values. There is also a hierarchy of *attributes* that can by represent by a *Specification diagram* [1] (Figure 3).

Developer should specify attributes and their initial values for further linking to corresponding *connections* in Templates.

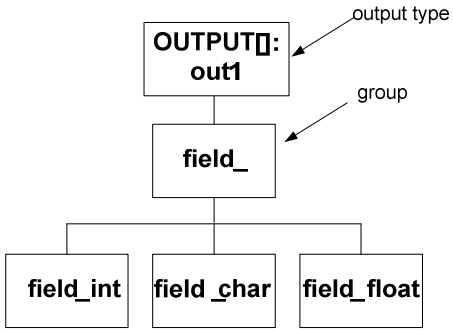


Figure 3. Specification diagram

3. Identifying of variable program parts. Variable program parts depend on Specification, so they should be later replaced with *connections*. This step is a key for *separation of concerns* [21], which is here, according to the SCT generator model, separation of program artifacts (Templates) from Specification and connection rules (Configuration). For example, the specification of used data structure contains a variable program part:

```
struct TList{
  int number;              // variable part
  TList *next;
};
```

4. Relaxing of prototype. Program parts that are identified as variable are being replaced by *connections* (in #-es). In case of data structure declaration this could look as follows:

```
struct TList{
  #fields_declarations#    // variable part
  TList *next;
};
```

5. Adding new rule to Configuration. Each *connection* created in the previous step has to be added into Configuration in form of a configuration rule. The configuration rule specifies respectively all three elements of the SCT model: connection, specification attribute and used code template. In the example this could look as follows:

#fields_declarations#,field_*,field_*.template

Specification attribute *field_** represents the number of occurrences of all attributes having name starting with *field_*. Template name specified as *field_*.template* represents usage of filenames that correspond to attribute names (e.g. usage of *field_int* designates the usage of *field_int.template*). This relation can be represented by a Configuration diagram [1] as shown in Figure 4.

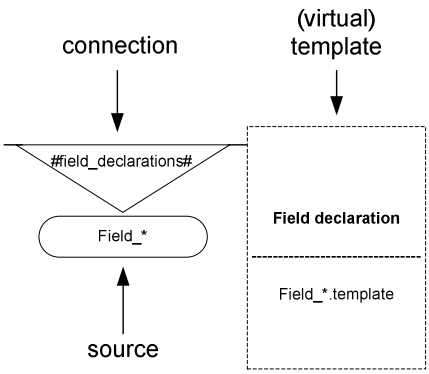


Figure 4. Part of the Configuration diagram representing single configuration rule

Usage of virtual template implies dynamic binding of template file (during the process of generation), analogous to appropriate concept of dynamic polymorphism in OOP.

6. Building of new code templates. The previous step anticipates the usage of different code templates. In case of virtual templates, there are possibly several real templates that have to be built. For a case shown in Figure 4, the templates are related to used data types, e.g.:

int #field#; *// Field_int.template*

This example requires adding of appropriate rule to Configuration (step 5):

*#field#,field_**

The third element, template, is here omitted, which specifies direct replacement of *connection* by attribute value.

7. Generating, testing and adjusting in a generative development process. The whole process of generators development and applications building consists of the repetition of operations described in the preceding steps. This can be represented by a spiral model, similar to Boehm's spiral model of software development [23], as shown in Figure 5.

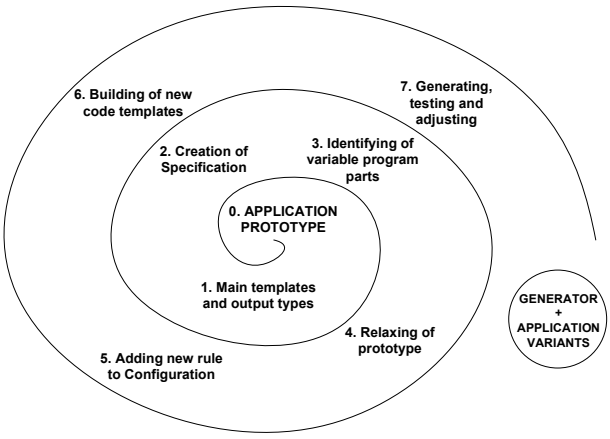


Figure 5: Spiral generator/application development

The development process starts with an application prototype and results by a generator that can be used for automatic assembling of different application variants. Modifications can start at each step of this process, following by the remaining steps, where some step could be omitted, except the step 7 (gives the final application).

V. DISCUSSION

The paper describes a method for building of code generators, according to previously introduced SCT generator model [1]. The proposed steps of generator building are also supported by a software tool named as Generator builder (Figure 6).

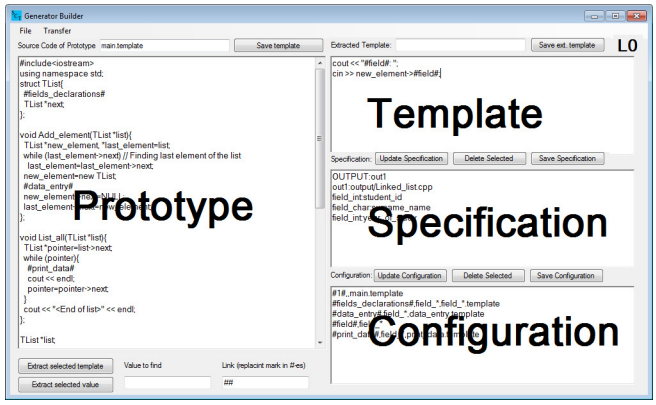


Figure 6. Generator builder

There were some design issues in defining of generator building steps and also in defining of their order. While the starting point (*application prototype*) is rather obvious, as well as the step 1(*choosing new main templates and output types*), there were some dilemmas about the step 2. The development of a generator could start from prototype decomposition, but this approach was proved to be problematic. The problem arises from a fact that each *connection* established in a prototype (in a form of replacing marks in '#'-es) has corresponding line in Specification and also corresponding line in Configuration. Some another *connection* could be assigned to the same Specification and Configuration lines, which could lead to their repetition. To avoid repetition, it seems that it's better approach to make the Specification first. In this case, each new *connection* has to be assigned to some existing Specification line. To avoid repetition in Configuration, it's enough just to check the existence of the same rule.

Other design issues were mostly concerned to design of a software tool that could be usable in building of generators. The most important effort was to adapt the user/developer interface to the need of SCT generator model, especially in performing tasks that are not very convenient for the standard programming editors. Some of the examples are simultaneous visibility of different SCT model elements, automatic updating of Specification and Configuration, easy extracting of code templates and easy building of template variants.

There also some remaining issues like generator documenting in a form of SCT diagrams (Specification diagram and Configuration diagram) [1] and building of

model elements repository, which could be some of the goals of the future work.

VI. CONCLUSION

The paper defines the development process of a SCT based code generator through seven steps. The steps are repetitive, and the whole process can be represented as a kind of spiral model of software development, similar to [23]. The appropriate software tool aimed for building of SCT based code generator, named as *Generator builder*, was also developed. For the purpose of testing the method and also the software tool, an example of generator development was given. The example starts with the prototype of a simple program in C++ that deals with linked list. The prototype was decomposed to the SCT model elements (Specification, Configuration and templates) through proposed seven steps. By using of these elements, an application variant was generated, compiled and tested.

There were some issues in design of generator building steps and software tool. Most of these were related to the order of development steps and to the design of software tool user interface. Finally, although there is a still space for improvements, the method and a software tool were found to be usable. The issues for the future work include working with the SCT diagrams and building of model elements repository.

REFERENCES

[1] D. Radošević, and I. Magdalenić, "Source Code Generator Based on Dynamic Frames", Journal of Information and Organizational Sciences, vol. 35, no. 1, pp. 73-91, July 2011.

[2] F. Heidenreich, J. Johannes, and M. Seifert, "Generating safe template languages", Proceedings of the 8th International ACM SIGPLAN Conference on Generative Programming and Component Engineering. Denver: ACM, 2009, pp. 99-108.

[3] J. Arnoldus, J. Bijpost, and M. Van Den Brand, "Repleo: A syntax-safe template engine", Proceedings of the 6th International Conference on Generative Programming and Component Engineering. Salzburg: ACM, 2007, pp. 25-32.

[4] D. Radošević, I. Magdalenić, and T. Orehovački, "Error Messaging in Generative Programming", Proceedings of the 22nd Central European Conference on Information and Intelligent Systems", Varaždin: FOI, 2011, pp. 181-187.

[5] V. Singh, and N. Gu, "Towards an integrated generative design framework", Design Studies, vol. 33, no. 2, pp. 185‑207, March 2012.

[6] H. Zhang, and S. Jarzabek, "XVCL: a mechanism for handling variants in software product lines", Science of Computer Programming, vol. 53, no. 3, pp. 381-407, December 2004.

[7] D. Batory, and S. O'Malley, "The Design and Implementation of Hierarchical Software Systems with Reusable Components", ACM Transactions on Software Engineering and Methodology vol. 1, no. 4, pp. 355-398, October 1992.

[8] D. Batory, "Validating Component Compositions in Software System Generators", Proceedings of the 4th International Conference on Software Reuse. Orlando: IEEE Press, 1996, pp. 72-81.

[9] openArchitectureWare Project, 2009, http://www.openarchitectureware.org/

[10] I. Groher, H. Papajewski, and M. Voelter, "Integrating Model-Driven Development and Software Product Line Engineering", 2007, http://www.eclipsesummit.org/summiteurope2007/presentations/ESE2007_Model-Groher-MDDAndPLE.pdf

[11] M. Regensburger, C. Buckl, A. Knoll, and G. Schrott, "Model Based Development of Safety-Critical Systems Using Template Based Code Generation, Proceedings of the 13th IEEE International Symposium on Pacific Rim Dependable Computing. Melbourne: IEEE Press, 2007, pp. 89-92.

[12] P. Friese, and B. Kolb, "Validating Ecore models using oAW Workflow and OCL", 2007, http://www.eclipsecon.org/summiteurope2007/presentations/ESE2007_Model-Friese-oAWAndOCL.pdf

[13] Eclipse Modeling Project, 2013, http://www.eclipse.org/modeling/

[14] T. W. Wong, S. Jarzabek, M. S. Soe, R. Shen, and H. Y. Zhang, "XML Implementation of Frame Processor," Proceedings of the 2001 Symposium on Software reusability: putting software reuse in context ACM: Toronto, 2001, pp. 164-172.

[15] P. G. Bassett, "Framing software reuse - lessons from real world". Toronto: Prentice Hall, 1997.

[16] R. Heradio, D. Fernandez-Amoros, L. de la Torre, and I. Abad, "Exemplar driven development of software product lines", Expert Systems with Applications, vol. 39, no. 17, pp. 12885–12896, December 2012.

[17] D. Radošević, M. Konecki, and T. Orehovački, "Java Applications Development Based on Component and Metacomponent Approach", Journal of Information and Organizational Sciences, vol. 32, no. 2, pp. 137-147, December 2008.

[18] D. Radošević, T. Orehovački, and M. Konecki, "PHP Scripts Generator for Remote Database Administration based on C++ Generative Objects", Proceedings of the 30th MIPRO International Convention on Intelligent Systems. Opatija: MIPRO, 2007, pp. 167-172.

[19] I. Magdalenić, D. Radošević, and Z. Skočir, "Dynamic Generation of Web Services for Data Retrieval Using Ontology", Informatica, vol. 20, no. 3, pp. 397-416, 2009.

[20] I. Magdalenić, D. Radošević, and T. Orehovački, "Autogenerator: Generation and execution of programming code on demand", Expert Systems with Applications, 2013, http://dx.doi.org/10.1016/j.eswa.2012.12.003, in press.

[21] D. Radošević, T. Orehovački, and Z. Stapić, "Automatic On-line Generation of Student's Exercises in Teaching Programming", Proceedings of the 21st Central European Conference on Information and Intelligent Systems. Varaždin: FOI, 2010, pp. 87-93.

[22] K. Czarnecki, and U. W. Eisenecker, "Generative Programming Methods, Tools, and Applications". Boston: Addison-Wesley, 2000.

[23] B. W. Boehm, "A Spiral Model of Software Development and Enhancement", Computer, vol. 21, no. 5, pp. 61-72, May 1988.

A Proportional Distribution of AUVs under Decentralized Control

N.V. Nagul
Institute for System Dynamics and Control Theory, Irkutsk, Russia
sapling@icc.ru

Abstract - The discrete-event model of a group of *autonomous underwater vehicles* (*AUVs*), operating under control of distributed supervisor (coordinators) in some large water area, is considered. The Lyapunov functions technique, developed for discrete-event systems, is exploited to establish the accessibility of the set describing a proportional distribution of the AUVs. Such distribution takes into consideration priorities of the areas on which the surveyed water area is divided. The ring topology with shortcuts, describing the placement of coordinators in the areas, is considered.

I. INTRODUCTION

Currently in oceanographic research and applied underwater technologies, groups of autonomous underwater vehicles (AUVs) are intensively used. There are a large number of problems associated with the implementation of collaborative and coordinated actions of AUVs in a real underwater environment, which need to be solved. Uncertainty of a priori information about the environment, variable underwater situations and the restrictions on communication and navigation are objective factors and should be taken into account. These factors cause a situational approach for solving problems of control of AUVs groups monitoring water areas and mapping seafloor [1]–[4].

The technologies for application of the automated distributed systems are rapidly developing in various fields of human activity and there is already an experience of cooperative work of mobile robots: "Martha" project, being developed in the Laboratory of System Architectures Analysis of France [5], "AMADEUS" project, proposed by Japanese developers [6], the DARS system, developed at the University of Nagoya. In the field of underwater robotics, similar investigations are conducted in AUSI (Autonomous Undersea Systems Institute, USA) [7]. The study is based on the ideas implemented in the joint project "SAUV" of IMTP FEB RAS and AUSI [8]–[11].

This paper is devoted to the problem of proportional distribution of the AUVs over the water environment under decentralized distributed control. To research water space of considerable size by a group of AUVs, as a rule, it is necessary to divide it into a number of smaller *areas*. It is supposed that there is a *coordinator* assigned for each specific area to control the number of AUVs in the area, shifting vehicles between the areas it oversees in order to achieve the proper distribution of AUVs across all the areas in the surveyed water space (Fig. 1). Wide variety of objects is able to play the coordinator role: selected vehicles of the AUVs group, specialized underwater vehicles, overwater buoys and seafloor beacons, ships, etc.

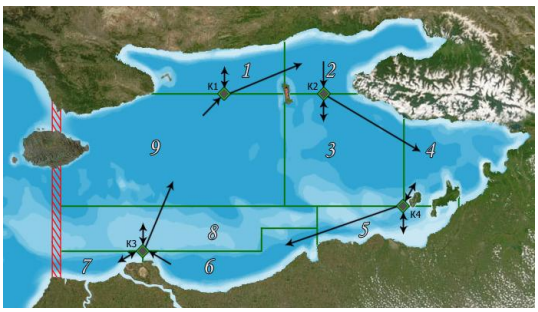


Figure 1. Water space devided into 9 areas. Coordinators K_i issue commands to move AUVs from one area to another

For such system it is convenient to employ the model presented in [12, 13]. In what following we consider some properties of that model and suggest some new results. The area–coordinator interconnection is convenient to describe by a bipartite graph in which the areas form one set of nodes and the coordinators form the other set. In this graph a coordinator node is connected only to the nodes of the areas for which it is responsible. In [12] the specific topology of bipartite graph – a unidirectional ring – is considered. In this topology there are an equal numbers of area and coordinator nodes, which alternate in sequence to form a ring comprising all of the nodes in the graph. The ring is unidirectional in the sense that the coordinators are only allowed to transfer vehicles around the ring in one direction (i.e. each coordinator is limited to moving vehicles from one specific area to another specific area). Unidirectional ring topology possesses obvious advantages. First of all, only one coordinator can transfer vehicles in an area and only one can remove vehicles from it. Secondly, there is no collision between several coordinators trying simultaneously to balance number of AUVs in assigned areas. In [12, 13] the problem of global balancing of mobile vehicles among areas is solved using the discrete-event approach. The discrete-event system (DES) has been built, which includes algorithms for distribution. Formal DES model is a generalization of the known classes of mathematical models, as automata networks, Petri nets, process algebras and others. Note that for AUV with a hierarchical control system DES can describe both the lower level of control, insuring safe movement of the vehicle, and the upper levels that are responsible for the

Supported by RFBR (grant no. 11-07-00655-a) and by Project no.10 of the Presidium of SB RAS.

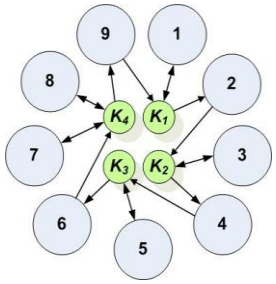


Figure 2. Bipartite graph illustrating "ring" topology for the model of Fig.1 with opportunity of "shortcuts" in the ring

constructing of an action plan of the group in general. An important field of application of the DES models is simulation and mission planning for groups of mobile robots. This problem for objective reasons is gaining more and more attention nowadays, so the development of DES properties research methods is a relevant issue.

II. THE DES MODEL OF THE AUVS DESTRIBUTION PROCESS

In what following, we borrow notations from [12, 13] for convenient reference. Let $A=\{1,\ldots,N_a\}$ be a set of areas, on which water environment of interest is divided, N_a is the number of them. Areas are denoted by variables *a*. Let $C=\{1,\ldots,N_C\}$ be a set of coordinators, N_C is the number of them, the coordinators are designated by letters *c*. In the case of the unidirectional ring topology $N_a = N_C$.

While it does achieve global balancing, unidirectional ring setup possesses a number of essential lacks. For example, in case of need to balance number of AUVs in areas with numbers N_a and *1*, it is necessary to transfer "superfluous" AUVs all the way around the ring, instead of at once transfer AUVs from area N_a to area 1. Besides, the number of coordinators should coincide with number of areas that frequently unrealizable in practice. For solving this problem, in the frame of DES model the algorithm has been developed to use "shortcuts" in a ring, which areas form [12].

Following [13], denote by $C(a)$ the coordinator that is in the right to remove AUVs out of the area *a*. For each area it is defined uniquely. Denote $S(c) \subset A$ a set of areas from which *c* can remove AUVs, that is $S(c) = \{a \in A : C(a) = c\}$. $D(c) = \{a \in A : C(a-1) = c\}$ denote the set of the areas in which the coordinator *c* is able to add AUVs, i.e. if the coordinator can remove the AUVs from the area *a-1*, it can add AUVs to the area *a*. Thus, the areas form a sort of the "ring" and coordinators can send vehicles along this ring so that they can visit each area. There also may be "shortcuts" in the ring formed by the areas, so visiting each area by the same AUV is not necessary (Fig. 2).

We suppose that coordinators do not communicate with each other and therefore they do not possess information about each other's actions. Then in general coordinator's list of vehicles in the area overseen will be inaccurate due to the range-limited nature of the communication between the vehicles and coordinators and the fact that more than one coordinator may alter the number of vehicles in an area. The aim of coordinators is to provide equal number of AUVs in every area.

For stated problem, in [12, 13] the DES models for the described system were built. Formally, *(logical) discrete-event system* [14] is a quintuple

$$G = (X, E, \{f_e\}_{e \in E}, g, E_v).$$

Here X is a set of states of the system; E is a set of events denoted by the letter e with some index; $g: X \to 2^E \setminus \{\varnothing\}$ is so-called *enable function*: $g(x)$ defines all events enabled in a current state $x \in X$. The definition of logical DES does not contain time (in common sense, like, for example, systems of differential equations). However, a kind of analog of a time variable is a counter that shows how many events have passed and, therefore, how many times states have changed. So, let there is a discrete "time" scale like a scale of the counter values $\boldsymbol{N} = \{0,1,2, \ldots\}$, and $x_k \in X$ present a state of DES at the "time moment" $k \in \boldsymbol{N}$ (i.e. when the counter is in the state *k*).

For each $e \in E$, the operator $f_e: X \to X$ transfers the system from one state to another whenever an event e occurs. It is natural to require $f_e(x)$ being defined if only $e \in g(x)$. If at this moment an event $e_k \in E$ occurs (of course, in the case of $e_k \in g(x_k)$), then the next state x_{k+1} is given by application of the operator f_{e_k} , i.e. $x_{k+1} = f_{e_k}(x_k)$.

Any sequence $x=\{x_0, x_1, x_2, \ldots\} \in X^{\boldsymbol{N}}$ such that for all k $x_{k+1} = f_{e_k}(x_k)$, where $e_k \in g(x_k)$, is said to be a *state trajectory (phase trajectory)*. Any sequence $e = \{e_0, e_1, e_2, \ldots\} \in E^{\boldsymbol{N}}$ such that there exists a state trajectory $x \in X^{\boldsymbol{N}}$ (where for all k $e_k \in g(x_k)$ and $x_{k+1} = f_{e_k}(x_k)$) is called an *event trajectory*. Therefore, to each initial state x_0 and each event trajectory e determining the order of applying the operators f_{e_k} , such that $e_0 \in g(x_0)$, the unique phase trajectory corresponds (the reverse is not true, in general).

A set of all event trajectories is denoted with $\boldsymbol{E}$ ($\boldsymbol{E} \subset E^{\boldsymbol{N}}$). By $E_v \subset \boldsymbol{E} \subset E^{\boldsymbol{N}}$ a set of *valid event trajectories* is denoted, that are physically possible in the system. Sometimes it may be useful to introduce a set $E_a \subset E_v$ of *allowed event trajectories*.

As a rule, during the DES functioning some sequences of events are undesirable. To avoid them supervisory control of DES is widely used, offered by P.J. Ramadge and W.M. Wonham [15] and highly developed now by numerous researchers. Supervisory control allows solving various control problems but in the context of AUVs group control it is, first of all, such problems as the coordination of joint AUVs actions under decentralized control. We consider the number of

vehicles in each of area as the DES state, called *plant* in the supervisor control theory: $x_P(k) = [x_1(k), \ldots, x_{N_a}(k)]^T$. Let $\tilde{x}_a^c(k)$ be a coordinator *c* estimate of number of AUVs in area *a* at the time moment *k*, and this variable is defined only if $c \in C(a) \cup C(a-1)$.

Since for some areas two coordinators are in the right of changing the number of AUVs in that area, inequalities

$$x_a(k) \geq \tilde{x}_a^c(k), \forall a \in S(c) \quad (1)$$

$$x_{a+1}(k) \leq \tilde{x}_{a+1}^c(k), \forall a \in D(c) \quad (2)$$

are correct.

The network of coordinators represents the distributed supervisor with states described by a vector

$$x_C(k) = [\tilde{x}_{a_{11}}^1(k), \ldots, \tilde{x}_{a_{1q_1}}^1(k), \ldots, \tilde{x}_{a_{N_c1}}^{N_c}(k), \ldots, \tilde{x}_{a_{N_cq_{N_c}}}^{N_c}(k)]^T,$$

$q_i \geq 2$. The state of the closed-loop system plant-supervisor is described by a vector $x(k) = [x_P^T(k), x_C^T(k)]^T$, and $X = \boldsymbol{N}^{|x|}$ is the DES state space. For the topology, where numbers of areas and coordinators are equal, $X = \boldsymbol{N}^{3N_C}$.

Whenever AUV checks in with the coordinator, not being in its list (or the coordinator himself transfers AUVs to corresponding area), the coordinator adds it in the list. Whenever AUVs, being on the list of the coordinator, doesn't check in during fixed period of time (or when the coordinator himself moves AUVs from area), the coordinator removes it from the list. The coordinator estimation of the number AUVs in area it is simple length of it's list. Thus, in considered system only following events are of interest:

- $e_c^{\alpha_{a \to a^*}}$ denote the partial event of coordinator $c = C(a)$ transferring α vehicles from the area *a* to the area a^*;
- $e_c^{\beta_a}$ denote the partial event of coordinator *c* has detected β vehicles in area *a* that it has not seen before so which earlier have been not entered in its list;
- $e_c^{\gamma_a}$ denote the partial event of the coordinator *c* has noted the absence of γ AUVs in the area *a*.

Then the total event space E is defined as the union of the sets of all partial events minus the empty set:

$$P(\{e_c^{\alpha_{a \to a^*}} : \alpha \in \boldsymbol{N}, c = \overline{1, N_C}, a \in S(c)\}),$$
$$P(\{e_c^{\beta_a} : \beta \in \boldsymbol{N}, c = \overline{1, N_C}, a \in S(c)\}),$$
$$P(\{e_c^{\gamma_a} : \gamma \in \boldsymbol{N}, c = \overline{1, N_C}, a \in S(c)\}),$$

where $P(A)$ is the power set of A.

It is supposed that $x(k) \in X_L$ for all *k*, where

$$X_L = \{x \in X : \sum_{a \in A} x_a = L\},$$

i.e. the number of AUVs in surveyed water area is constant and equal to *L*.

The distribution of vehicles is considered to be proper if the number of vehicles in each area is globally balanced. This means that the number of vehicles in any area differs from the one in any other area by no more than one, i.e.

$$X_I = \{x \in X_L : x_a \in \{m, m+1\} \text{ for all } a \in A\},$$

where $m = \lfloor \frac{L}{N_a} \rfloor$, and $\lfloor x \rfloor$ denote the integer part of x.

Unlike [12, 13], we will show how the results of the stability theory [18], adapted for DES in [14], allow to investigate properties of the considered model. First we give some notions and necessary theorems from [14].

Let $\rho: X \times X \to R$ denote a metric on X, and $\{X; \rho\}$ is a metric space. Let $M \subset X$ and $\rho(x, M) = \inf\{\rho(x, x') : x' \in M\}$ denote the distance from the point x to the set M. As r-neighbourhood of some set $M \subset X$ a set $S(M; r) = \{x \in X : 0 < \rho(x, M) < r\}$ is ment, where $r > 0$. As a fuctional a mapping of an arbitrary set to the set of real numbers *R* is meant.

Definition 1. A set $X_m \subset X$ is called *invariant* with regard to DES G if $x_0 \in X_m$ imply $h(x_0, e, k) \in X_m$ for all $e \in E_v(x_0)$ and $k \in \boldsymbol{N}$.

Definition 2. A closed invariant set $X_m \subset X$ of the system G is called *Lyapunov stable* with regard to E_a if for all $\varepsilon > 0$ there exist such $\delta > 0$ that while $\rho(x_0, X_m) < \delta$ it will be $\rho(h(x_0, e, k), X_m) < \varepsilon$ for all $e \in E_a(x_0)$ and $k \in \boldsymbol{N}$. Moreover, if $\rho(h(x_0, e, k), X_m) \to 0$ while $k \to \infty$ for all $e \in E_a(x_0)$, then a closed invariant set X_m of the system G is called *asymptotically stable* with regard to E_a.

Theorem 1 [14]. In order to a closed invariant set $X_m \subset X$ of the system G be Lyapunov stable with regard to E_a it is necessary and sufficient that there exists a functional V , defined in some sufficiently small neighbourhood $S(X_m; r)$ of the set X_m , with the following properties:

1) for any arbitrarily small $c_1 > 0$ such $c_2 > 0$ may be found that for any $k \in N$ and for all $x \in S(X_m; r)$ such that $\rho(x, X_m) > c_1, V(x) > c_2$;
2) for any arbitrarily small $c_4 > 0$ it may be found such $c_3 > 0$ that for $x \in S(X_m; r)$ and $\rho(x, X_m) < c_3, V(x) \leq c_4$;
3) if for all $x_0 \in S(X_m; r)$ and all $e \in E_a(x_0)$ $V(h(x_0, e, k))$ do not increase while $k \in \boldsymbol{N}$ increase, as soon as $h(x_0, e, k) \in S(X_m; r)$.

Theorem 2 [14]. In order to a closed invariant set $X_m \subset X$ of the system G be Lyapunov asymptotically stable with regard to E_a it is necessary and sufficient that there exists a functional V, defined in some sufficiently

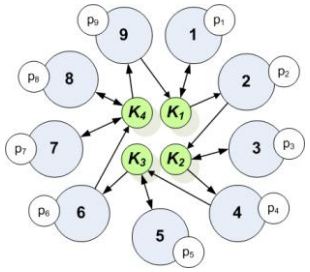


Figure 3. Areas endowed with priorities

small neighbourhood $S(X_m;r)$ of the set X_m with the properties 1), 2), 3) of Theorem 1, and, moreover, for all $e \in E_a(x_0)$ $V(h(x_0,e,k)) \to 0$ while $k \to \infty$, as soon as $h(x_0,e,k) \in S(X_m;r)$.

Obviously, in the statesments of the above definitions and theorems the set E_a may be replaced with E_v, as for any DES E_a is considered coincide with E_v, if not stated otherwise.

To imply Theorems 1 and 2 for investigating properties of the model considered in [13], an appropriate functional V should be chosen. Let $m(k) = \min_{a \in A} x_a(k)$ denote the minimum number of AUVs in each area at the moment k. Similarly let $M(k) = \max_{a \in A} x_a(k)$ denote the maximum number of AUVs in each area at the moment k. Let's consider the function $V(x(k)) = M(k) - m(k)$. Using the results from [12, 13], it is easy to show that $V(x(k))$ does not increase in time and satisfies all conditions of Theorem 1 on stability of invariant sets of DES and, therefore, the following theorem holds.

Theorem 3. For the DES, described in [13], the closed invariant set X_I is Lyapunov stable with reference to E_v.

Next, consider the function $V(x(k)) = \lfloor (M(k) - m(k))/2 \rfloor$. It can be shown that $V(x(k))$ do not increase in time, and, moreover $V(x(k)) \to 0$ while $k \to \infty$. So the function $V(x(k))$ satisfies all conditions of Theorem 2 on asymptotical stability of invariant sets of DES and the following theorem holds.

Theorem 4. For the DES, described in [13], the closed invariant set X_I is Lyapunov asymptotically stable with reference to E_v.

III. PRIORITIZED BALANCING

Natural generalization of the described model is assigning to areas some priorities – a certain portion of the total number of AUVs that the area should have (Fig. 3). Let for each area a the priority p_a is defined. In [12] the algorithm is proposed, allowing, though locally, to balance number of AUVs in the adjacent areas considering their priorities. The proposed algorithm is constructed for the unidirectional ring without shortcuts. Now we construct a DES that provides prioritized balancing for the ring topology with shortcuts.

The DES, built in [13], is modified in the following way. The enable function $g: X \to 2^E \setminus \{\varnothing\}$ is now defined as follows: if $e(k) \in g(x)$ then

1) the coordinator c transfers AUVs only once during the combined event $e(k)$, $c = 1,\ldots,N_c$;

2) for all $e_c^{\alpha_{a \to a^*}} \in e(k)$ the following conditions should be satisfied:

   a) the coordinator $c = C(a)$ moves AUVs from the area $a \in S(c)$ to the area $a^* \in D(c)$, which is chosen according the rule

$$\frac{\tilde{x}_a^c(k)}{p_a} - \frac{\tilde{x}_{a^*}^c(k)}{p_{a^*}} = \max_{a' \in S(c), a'' \in D(c)} \left(\frac{\tilde{x}_{a'}^c(k)}{p_{a'}} - \frac{\tilde{x}_{a''}^c(k)}{p_{a''}} \right),$$

   and transferring to the right next area is considered first-priority;

   b) the number of AUVs to transfer is determined according the rule

$$\text{if } \frac{\tilde{x}_a^c(k)}{p_a} \le \frac{\tilde{x}_{a^*}^c(k)}{p_{a^*}}, \text{ then } \alpha(k) = 0;$$

$$\text{if } \frac{\tilde{x}_a^c(k)}{p_a} > \frac{\tilde{x}_{a^*}^c(k)}{p_{a^*}}, \text{ then}$$

$$1 \le \alpha(k) < \frac{p_a(1 - \tilde{x}_{a^*}^c(k)) + p_{a^*}(1 + \tilde{x}_a^c(k))}{p_a + p_{a^*}}.$$

3) for all $e_c^{\beta_a} \in e(k)$ and $e_c^{\gamma_a} \in e(k)$ estimates take place:

$$\beta_a \le x_a(k) - \tilde{x}_a^c(k), \forall a \in S(c),$$

$$\gamma_a \le \tilde{x}_a^c(k) - x_a(k), \forall a \in D(c).$$

As E_v we shall consider the event trajectories, in which it is not more than a finite amount of time T takes place between all communication sessions of each AUV with its coordinator.

At any time k when an event $e(k) \in g(x(k))$ occurs, it changes the state of the system according to the function $x(k+1) = f(x(k), e(k))$ defined by the rules

$$x_a(k+1) = x_a(k) + \alpha_a^+(k) - \alpha_a^-(k),$$

$$\tilde{x}_a^c(k+1) = \tilde{x}_a^c(k) + \alpha_a^+(k) - \alpha_a^-(k) + \beta_a - \gamma_a,$$

where $\alpha_a^+(k)$ is the number of AUVs added to a, and $\alpha_a^-(k)$ is the number of AUVs removed from a.

The proportional distribution of vehicles by areas in accordance with the priorities is approximately described by the set

$$X_P = \{x \in X_L : \frac{x^c_{a'}(k)}{p_{a'}} - \frac{x^c_{a''}(k)}{p_{a''}} < \frac{1}{p_{a'}} + \frac{1}{p_{a''}},$$
$$\forall a' \in D(c), \forall a'' \in S(c),\ c = 1,\ldots,N_c\}.$$

The meaning of this set may not be obvious and will be described later.

Theorem 5. For the DES, describing the model of prioritized balancing, the closed invariant set X_P is Lyapunov stable with reference to E_v.

Proof. Due to the lack of space we give only hints of the proof. For simplicity we consider the case when coordinator's estimates of the number of vehicles in areas are correct, i.e. $x_a(k) = \tilde{x}^c_a(k) = \tilde{x}^{c-1}_a(k)$. A general case is obtained using inequalities (1) and (2) but require some additional computations.

To proof the theorem, consider the function

$$V(x(k)) = \left\lfloor \left(\sum_{a \in A} \max\{|m_a - x_a(k)|,\ 1\} \right) / N_a \right\rfloor,$$

where $m_a = \frac{L}{P} \cdot p_a$ determine ideal, though not in general realizable, portion of vehicles in the area *a*, $P = p_1 + \ldots + p_{N_a}$. We need to show that $V(x(k))$ satisfy conditions of Theorem 1. Conditions 1) and 2) are met by the choice of $V(x(k))$. For 3) we should show that while $x(k) \notin X_P$, $V(x(k))$ do not increases in time, i.e. $V(x(k+1)) \le V(x(k))$ whatever event $e(k) \in g(x(k))$ occur at time *k*. To do so, consider the enable function $g : X \to 2^E \setminus \{\varnothing\}$ determining the algorithm of distribution. The essential part of the proof is the following considerations. Every time the event $e_c^{\alpha_{a \to a^*}}$ occur, it changes two components of the vector $x(k)$: instead of $x_P(k) = [\ldots, x_a(k), \ldots, x_{a^*}(k), \ldots]^T$ we obtain

$$x_P(k+1) = [\ldots, x_a(k) - \alpha(k), \ldots, x_{a^*}(k) + \alpha(k), \ldots]^T.$$

Note that $\alpha(k)$ is chosen to make equal quantities $\frac{x_a(k)}{p_a}$ and $\frac{x_{a^*}(k)}{p_{a^*}}$. So we want to find such a number η that $\frac{x_a(k) - \eta}{p_a} = \frac{x_{a^*}(k) + \eta}{p_{a^*}}$. Solving this equation and taking into account $\frac{p_{a^*}}{p_a} = \frac{m_{a^*}}{m_a}$, we obtain

$$\eta = \frac{p_{a^*} x_a(k) - p_a x_{a^*}(k)}{p_{a^*} + p_a} = \frac{m_{a^*} x_a(k) - m_a x_{a^*}(k)}{m_{a^*} + m_a}.$$

Since only integer number of AUVs can be moved from one area to another, we add to η unity and choose $\alpha(k)$ as was described above. Then

$$\frac{x_a(k) - \alpha(k)}{p_a} = \frac{x_a(k) - (\eta + \Delta)}{p_a} = \frac{x_a(k) - \eta}{p_a} - \frac{\Delta}{p_a},$$

$$\frac{x_{a^*}(k) + \alpha(k)}{p_{a^*}} = \frac{x_{a^*}(k) + (\eta + \Delta)}{p_{a^*}} = \frac{x_{a^*}(k) + \eta}{p_{a^*}} + \frac{\Delta}{p_{a^*}}$$

and

$$\frac{x_{a^*}(k+1)}{p_{a^*}} - \frac{x_a(k+1)}{p_a} = \frac{\Delta}{p_a} + \frac{\Delta}{p_{a^*}} < \frac{1}{p_a} + \frac{1}{p_{a^*}},$$

what explains the nature of the set X_P.

Let's show that

$$|x_a(k) - \eta - m_a| + |x_{a^*}(k) + \eta - m_{a^*}| \le$$
$$\le |x_a(k) - m_a| + |x_{a^*}(k) - m_{a^*}|. \quad (3)$$

It is not hard to derive that

$$x_a(k) - \eta - m_a = \frac{m_a((x_a(k) + x_{a^*}(k)) - (m_a + m_{a^*}))}{m_{a^*} + m_a},$$

$$x_{a^*}(k) + \eta - m_{a^*} = \frac{m_{a^*}((x_a(k) + x_{a^*}(k)) - (m_a + m_{a^*}))}{m_{a^*} + m_a}.$$

Now it is clear that the signs of $x_a(k) - \eta - m_a$ and $x_{a^*}(k) + \eta - m_{a^*}$ depend on the sign of expression $m_a + m_{a^*}$). That's why considering combinations of $x_a(k)$, $x_{a^*}(k)$, m_a and m_{a^*}, we will be able to prove (3). For example, if $(x_a(k) + x_{a^*}(k)) > (m_a + m_{a^*})$ then it should be $x_a(k) - \beta > m_a$ and $x_{a^*}(k) + \beta > m_{a^*}$. Consider all combinations then $(x_a(k) + x_{a^*}(k)) > (m_a + m_{a^*})$ take place, with an additional condition $m_a > m_{a^*}$. It is possible if

1) $x_a(k) > m_a$ and $x_{a^*}(k) > m_{a^*}$,

2) $x_a(k) > m_a$ and $x_{a^*}(k) < m_{a^*}$,

3) $x_a(k) < m_a$ and $x_{a^*}(k) > m_{a^*}$.

In the third case it is not possible to simultaneously have $x_a(k) - \eta > m_a$ and $x_{a^*}(k) + \eta > m_{a^*}$ so this case is not suitable.

In the first case we obtain $x_a(k) > m_a$ and $x_a(k) - \eta > m_a$, $x_{a^*}(k) > m_{a^*}$ and $x_{a^*}(k) + \eta > m_{a^*}$, so the sum $|m_a - x_a(k)| + |m_{a^*} - x_{a^*}(k)|$ do not change.

In the second case $x_a(k) > m_a$ and $x_a(k) - \eta > m_a$ at the same time as $x_{a^*}(k) < m_{a^*}$ and $x_{a^*}(k) + \eta > m_{a^*}$, therefore the sum $|m_a - x_a(k)| + |m_{a^*} - x_{a^*}(k)|$ decrease and (3) holds.

The symmetrical case $m_a < m_{a^*}$ and the situation $(x_a(k) + x_{a^*}(k)) < (m_a + m_{a^*})$ are similar and omitted due to the lack of space. Using (3) and passing to $\alpha(k)$ instead of η, $V(x(k+1)) \le V(x(k))$ may be derived.

Consideration of partial events $e_c^{\beta_a} \in e(k)$ and $e_c^{\gamma_a} \in e(k)$ is rather simple due to the estimates on β_a and γ_a in the definition of the enable function $g(x)$. That finishes the main line of proving the theorem.

It should be noted that the rate of convergence to the set X_P will be the highest if

$$\alpha = \max\left\{\left\lfloor \frac{p_a(1-\tilde{x}_{a^*}^c(k)) + p_{a^*}(1+\tilde{x}_a^c(k))}{p_a + p_{a^*}} \right\rfloor, 1\right\}.$$

IV. CONCLUSION

In the paper it was shown how the results of stability theory and discrete-event modeling approach help to develop and prove the convergence of the algorithm of prioritized balancing of mobile agents (AUVs) under decentralized control of distributed supervisor in water environment where communication between coordinators is strictly limited. The known results on prioritized balancing were generalized to embrace the case of ring topology with shortcuts. Further investigations will be directed to computation of the bound on the time of convergence to the balanced state of the system described and considering topologies other than ring or line topology.

REFERENCES

[1] H. Stommel, "The slocum mission," Oceanography, vol. 2(1), pp. 22–25, 1989.

[2] T. Curtin, J. Bellingham, J. Catipovic, and D. Webb, "Autonomous oceanographic sampling networks," Oceanography, vol. 6(3), pp. 86–94, 1993.

[3] T. Kaga and T. Fukuda, "An oscillation analysis on distributed automations robotic system," in Proc. IEEE Int. Conf. on Robotics and Automation, vol. 4, Leuven, Belgium, May 1998, pp. 2846–2851.

[4] M. Ageev, "Organization of an automated network of oceanographic measurements based on the solar AUV," Underwater Investigation and Robotics, vol. 2, pp. 5–12, 2006, (in Russian).

[5] R. Alami, S. Fleury, M. Herrb, F. Ingred, and F. Robert, "Multi-robot cooperation in the MARTHA project," IEEE Robot. Autom. Mag., vol. 5(1), pp. 36–47, 1998.

[6] T. Karada and K. Oikawa, "AMADEUS: A mobile, autonomous decentralized utility system for indoor transportation," in Proc. IEEE Int. Conf. on Robotics and Automation, vol. 4, Leuven, Belgium, May 1998, pp. 2229–2236.

[7] S. Chappell, R. Komerska, D. Blidberg, C. Duarte, G. Martel, D. Crimmins, M. Beliard, R. N. J. Jalbert, and R. Bartos, "Recent field experience with multiple cooperating solar-powered AUVs," in Proc. of the 15th Annual International Symposium on Unmanned Untethered Submersible Technology, Durham, USA, Aug. 2007.

[8] J. Jalbert, P. Irazoqui-Pastor, S. Miles, D. Blidberg, J. Darwin, and M. Ageev, "AUV technology evaluation and development project," in Proc. of the 10th International Symposium on Unmanned Untethered Submersible Technology, Durham, USA, Sep. 1997, pp. 75–87.

[9] D. Blidberg, J. Jalbert, and M. Ageev, "Experimental results; the AUSI/IMTP solar powered AUV project," in Proc. of the Ocean Community Conference, Baltimore, USA, Nov. 1998.

[10] M. Ageev, J. Jalbert, and D. Blidberg, "Description and analysis of a solar autonomous underwater vehicle," MTS Journal, vol. 32(4), 1999.

[11] M. Ageev, J. Jalbert, D. Blidberg, C. Melchin, and D. Troop, "Results of the evaluation and testing of the solar powered AUV and its subsystems," in Proc. of the 11th International Symposium on Unmanned Untethered Submersible Technology, Durham, USA, Sep. 1999.ind Speed and Direction. Moscow, Russia: Nauka, 2005, (in Russian).

[12] B. Moore and K. Passino, "Decentralized redistribution for cooperative patrol," Int. J. Robust Nonlinear Control, vol. 18, pp. 165–195, 2008.

[13] ——, "Distributed coordination strategies for wide-area patrol," Journal of Intelligent and Robotic Systems, vol. 56, pp. 23–45, 2009.

[14] Michel A. N., Kaining W. K., Bo H. *Qualitative Theory of Dynamical Systems*, Marcel Dekker, Inc., N.Y., 2001.

[15] P. Ramadge and W. Wonham, "Supervisory control of class of discrete event processes," SIAM J. Control and Optimisation, vol. 25(1), pp. 206–230, 1987.

A P300–Based Brain–Computer Interface: Towards a Simpler Approach

F. Melinščak*, A. B. Jerbić** and M. Cifrek*
*Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
**University of Zagreb, Zagreb, Croatia
Email: filip.melinscak@fer.hr

***Abstract*—Brain–computer interfaces (BCIs) allow new ways of communication with computers, independent of the normal neuromuscular pathways. One of the neural mechanisms which can be used in a BCI is the P300 wave. In this paper we focused on the realization of an EEG–based P300 speller application. We used the data set IIb from the BCI Competition 2003. The results achieved with simple methods – using only six electrodes and linear classification – surpass those achieved in the competition. The system achieves full accuracy with four repetitions of the stimuli sequence, and 8.4 seconds are needed to select a symbol.**

I. Introduction

Brain–computer interface (BCI) is a communication and control system that is independent of normal neuromuscular pathways [1]. This technology can be especially useful for people suffering from partial or total paralysis that is the result of a brain or a spinal cord injury or some degenerative neuromuscular disease (e.g. locked–in syndrome, amyotrophic lateral sclerosis, cerebral palsy). Due to advancements in medicine, these people's life expectancy is prolonged, emphasizing even more their need for alternative means of communication with the outer world [2].

BCI extracts specific patterns of user's brain activity and uses them to control devices (wheelchair, prosthetic devices) or for communication (mental typewriter). The most popular way of obtaining brain signals for BCIs is electroencephalography (EEG) which non-invasively measures electrical brain activity. There are different neural mechanisms that can be used for controlling BCIs and among the most popular are motor imagery, steady–state visual evoked potentials (SSVEPs) and the P300 wave.

In our study we obtained control signals for the BCI by detecting the P300 wave. P300 is one component of cognitive event-related potentials (ERPs). ERPs are a brain response to an event that can be triggered either externally (visual, auditory, somatosensory stimuli) or internally (mental tasks). The P300 wave is a positive deflection in EEG that peaks approximately 300 ms after the presentation of a rare stimulus that the person perceives as important. Typically, the P300 wave is elicited in an "oddball" paradigm: a person is presented with a sequence of different stimuli in which target stimuli are infrequent. The more distinct and rare this target stimuli are, the more prominent the P300 wave is [3].

A P300 speller is a BCI based on the P300 wave that uses a method first developed by Farwell and Donchin [4], [5]. The basic idea is to present letters of the alphabet, numbers and possibly additional symbols in a form of a matrix. The user focuses attention on the desired symbol. At each moment only one row or column is intensified, eliciting the P300 wave if the row or column contains the attended symbol. Due to a small amplitude of the P300 wave in relation to spontaneous EEG activity, the stimulation by row/column intensifications usually needs to be repeated a number of times in order to extract a distinct P300 response.

BCI development is a fast growing field. In recent years, competitions have been established to address the need for adequate comparison of BCI algorithms. BCI competitions provide a set of data with its description and appropriate class labels, and a set of unlabeled data for testing. The participants' goal is to classify the unlabeled test data with methods of their own choice. After the competition finishes, competition results and true labels for the test data are revealed, allowing other members of BCI community to test their algorithms and compare results.

We chose BCI competition 2003 data set IIb [6] for our analysis. Our goal was to demonstrate that the application of simple methods to the problem of P300 detection can provide acceptable P300 speller performance.

II. Methods

A. The Data Set

The data set IIb, used in this study, was provided by the Wadsworth Center for usage in the BCI Competition 2003 [6]. The data set consists of P300 evoked potentials recorded with BCI2000 software [7].

Data collection followed the aforementioned P300 paradigm introduced by Farwell and Donchin [4]. According to the paradigm, the symbols were arranged into a 6×6 matrix with the target word displayed above the matrix and the current symbol in parentheses (Fig. 1).

The time course of the experiments was as follows: before each symbol a blank matrix (all symbols displayed with the default intensity) was displayed for 2.5 s. Next, the rows and columns of the matrix were intensified consecutively. A single repetition consists of 12 row/column intensifications – one for each row and column – which were presented in a random sequence. For each symbol 15 repetitions of stimuli were used. A single intensification lasted for 100 ms and the matrix was blank for 75 ms before the next stimulus, hence the inter-stimulus interval

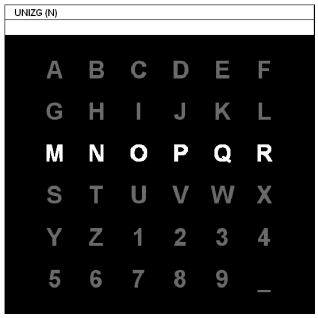


Fig. 1. The 6×6 P300 matrix display with a target word on top

(ISI) was 175 ms, i.e. flashing frequency was 5.7 Hz. After 15 repetitions a blank matrix was displayed for 2.5 s indicating the change of the current target symbol.

The signals were recorded from one subject over the course of three sessions, using 64 EEG channels, standard 10–10 electrode setup, and a sampling rate of 240 Hz. The data from the first two sessions are labeled as being target (P300) or non-target and are designated for the training of the P300 speller. Third session is to be used as the test set, and the true labels were revealed after the end of the BCI Competition 2003.

There are 42 symbols available in the training set, but we used only the first 39 because the signals for the last 3 symbols were recorded with a different ISI (this omission reduced the number of training samples by 7.14 %). With 180 stimuli per symbol (15 repetitions of 12 distinct stimuli), there are 7020 training samples in total: 1170 target and 5850 non-target samples. The test set consists of 31 symbols, and after the competition had ended it was revealed that the symbols formed the following string: "`FOOD MOOT HAM PIE CAKE TUNA ZYGOT 4567`".

B. Preprocessing

One of the goals of data preprocessing is to select a subset of available channels; in practice this corresponds to fewer electrodes, i.e. shorter setup time and lower price. In our study a fixed channel selection was used, consisting of the following 6 scalp locations: Cz, Fz, Pz, PO7, PO8, and Oz. We chose this particular set adhering to the findings of a study by Krusienski et al. that has shown superiority of this configuration to several other common choices [8].

After channel selection, which can be viewed as spatial filtering, a temporal bandpass filter was applied with the goal of removing the DC component and discarding the undiscriminating information. In accordance with suggestions available in literature [9], the signals were bandpass filtered from 0.1 Hz to 10 Hz with an 8^{th} order Butterworth filter.

The signals were divided into 800 ms epochs starting at the stimuli onsets. A single epoch consisted of 1152 data samples (0.8[s] $\times$ 240 [samples/s] $\times$ 6 [# of channels]), which indicated that further dimensionality reduction was needed.

Number of features was further reduced with temporal decimation of samples. Taking into account the sampling rate (240 Hz) and upper cutoff frequency (10 Hz), the signals were decimated with the factor 12: blocks of 12 samples were replaced with their mean (this is equivalent to moving average filtering prior to standard decimation). After decimation, the dimension of the feature vector was 96, which was acceptable for classification.

Before classification, preprocessed training samples $\mathbf{x}^{(i)}$ were organized into an observation matrix $\mathbf{X}_{\text{train}}$ with dimensions $N_{\text{train}} \times D$, where N_{train} denotes the number of training samples and D denotes the length of feature vectors. Corresponding class labels $y^{(i)}$ were organized into a column–vector $\mathbf{y}_{\text{train}}$ of length N_{train}. Analogously, the test samples were organized into observation matrix $\mathbf{X}_{\text{test}}$ with dimensions $N_{\text{test}} \times D$.

C. Classification

From a vast number of choices described in literature [10]–[12], we chose a common, fairly simple and computationally cheap classifier — linear discriminant analysis (LDA). LDA presupposes Gaussian distribution of features within classes and equal covariance matrices of all classes [13]. These presumptions hold approximately for ERP data [14], therefore LDA is a legitimate candidate for P300 applications.

The parameters $\mathbf{w}$ (weights) and b (offset) of the discriminant function $g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ were learned via LDA from the observation matrix $\mathbf{X}_{\text{train}}$ and class labels $\mathbf{y}_{\text{train}}$:

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{1}$$

$$b = \frac{1}{2}\mathbf{w}^\top(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \ln\frac{P(\omega_2)}{P(\omega_1)}, \tag{2}$$

where $\boldsymbol{\mu}_i$ is the mean of the class ω_i, and $\mathbf{\Sigma}$ is the pooled covariance matrix estimate. Additionally, $P(\omega_i)$ is the *a priori* probability of the class ω_i (for the P300 speller *a priori* probabilities are $P_{\text{target}} = 1/6$ and $P_{\text{non-target}} = 5/6$).

After the training, the classifier was applied to the test samples organized into the $\mathbf{X}_{\text{test}}$ observation matrix. Since 180 samples correspond to one symbol, classification was done over blocks of samples. The task of predicting a single symbol was broken down into the prediction of the row and the column. The discriminant function was applied to all samples within a block that were associated with a particular stimulus (particular row/column) and these scores were summed up. The row and column with maximum scores were determined as the targets, and the symbol belonging to both the predicted row and the predicted column was output. The procedure of predicting the target rows and columns is succinctly described with following expressions:

$$\text{predicted row} = \arg\max_{\text{rows}} \sum_{i_{\text{row}}} g(\mathbf{x}_{i_{\text{row}}}), \tag{3}$$

$$\text{predicted column} = \arg\max_{\text{columns}} \sum_{i_{\text{column}}} g(\mathbf{x}_{i_{\text{column}}}). \tag{4}$$

When the system was tested with less than 15 available repetitions of stimuli, the samples from the beginning of the block were used, and other samples were discarded (e.g. if 5 repetitions are used for detection, only the first 60 samples in a block associated with a symbol are used).

TABLE I. P300 DETECTION RESULTS ACHIEVED IN THIS STUDY

Repetitions	Inferred string	Accuracy (%)	ITR (bit/min)
1	FOOD MOOT BCM JIE CAHE ZUNA ZMAOT X5Z7	70.97	80.34
2	FOOD GOOT BAM JIE CAIE TUNA ZMGOT X5Z7	77.42	46.30
3	FOOD MOOT HAM PIE CAKE TCNA ZSGOT 4567	93.55	42.80
4	FOOD MOOT HAM PIE CAKE TUNA ZYGOT 4567	100.00	36.93

III. RESULTS AND DISCUSSION

Most commonly reported metrics in studies regarding P300 BCIs are accuracy and information transfer rate. Accuracy is defined as:

$$\mathrm{Acc} = P = \frac{\#\,\text{of correctly inferred symbols}}{\#\,\text{of symbols}}. \quad (5)$$

Information transfer rate (ITR) is usually measured in bits per minute, and is given by:

$$\mathrm{ITR} = \frac{\#\,\text{of decisions}}{\text{duration}\,[\text{min}]} \times B[\text{bit}], \quad (6)$$

where B denotes bit rate (bits conveyed by one selection). Bit rate B is calculated as [15]:

$$B = \log_2 N + P \log_2 P + (1-P) \log_2 \left(\frac{1-P}{N-1} \right), \quad (7)$$

where P is the accuracy of detection and N is the number of possible choices (i.e. $N = 36$ for the P300 speller).

The described P300 speller achieves full accuracy on the test data by using 4 out of 15 available repetitions (Table I). In this case, the system communicates 7.14 symbols per minute (bit rate of 36.93 bit/min); 8.4 s are needed for a single selection. If a single repetition is used for detection, accuracy drops to 70.97 %, but 28.57 symbols can be transferred per minute (80.34 bit/min); for comparison — expected accuracy of classification by chance is 2.8 %. We also noted that for incorrectly inferred symbols, only the predicted row or column was wrong, but not both.

For comparison, we present the results achieved in the BCI Competition 2003 (Table II). In this competition, only the accuracy of classification was considered for the evaluation of submissions. The minimal number of repetitions needed for perfect classification was self-reported. Algorithm submitted by Matthias Kaper et al. achieved full accuracy with the smallest number of repetitions (five). Kaper et al. used spatio-temporal features and a non-linear Gaussian SVM for classification [17].

In comparison with the winning algorithms of the BCI Competition 2003, the P300 speller described in this paper needs less repetitions to achieve full accuracy, albeit using simpler methods (i.e. linear classification).

TABLE II. BCI COMPETITION 2003 RESULTS (DATA SET IIB) [16]

Rank	Participants	Accuracy (%)	Repetitions
1.	M. Kaper et al.	100.0	5
1.	X. Gao et al.	100.0	5–8
1.	V. Bostanov	100.0	6
1.	B. Blankertz et al.	100.0	6–11
1.	D. Tax et al.	100.0	–
6.	J. Werfel et al.	45.2	–
7.	E. Glassman	35.5	–

IV. CONCLUSION

The P300 speller is a BCI that uses the P300 wave as a control signal. Due to its robustness, the P300 wave is a suitable control signal. However, signal-to-noise ratio is low in case of EEG measurements and detection of the P300 wave usually requires several repetitions of the target stimulus and averaging of the resulting waveforms. This is a challenge for the P300 speller applications because one additional repetition of a symbol implies the repetition of 12 stimuli (6 rows and 6 columns). Therefore, minimizing the number of repetitions is necessary for achieving a high information transfer rate.

We performed our analysis on the publicly available BCI competition 2003 data set IIb. Our results surpass the results obtained by the winners of the competition, while the methods we used are computationally less demanding and are easily implemented. However, the presented results should not be generalized without caution, due to the small size of the analyzed data set.

REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.

[2] A. Kubler and B. Kotchoubey, "Brain-Computer Communication: Unlocking the Locked In," *Psychol. Bull.*, vol. 127, no. 3, pp. 358–375, 2001.

[3] J. Polich, "Updating P300: An integrative theory of P3a and P3b," *Clin. Neurophysiol.*, vol. 118, no. 10, pp. 2128–2148, 2007.

[4] L. A. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials." *Electroen. Clin. Neuro.*, vol. 70, no. 6, pp. 510–523, 1988.

[5] E. Donchin, K. M. Spencer, and R. Wijesinghe, "The Mental Prosthesis: Assessing the Speed of a P300-Based Brain–Computer Interface," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 2, pp. 174–179, 2000.

[6] B. Blankertz *et al.*, "The BCI Competition 2003: Progress and Perspectives in Detection and Discrimination of EEG Single Trials." *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1044–1051, 2004.

[7] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: A General-Purpose Brain-Computer Interface (BCI) System," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1034–1043, 2004.

[8] D. J. Krusienski, E. W. Sellers, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "Toward enhanced P300 speller performance," *J. Neurosci. Meth.*, no. 1, pp. 15–21.

[9] L. Bougrain, C. Saavedra, and R. Ranta, "Finally, what is the best filter for P300 detection?" in *TOBI Workshop III - Tools for Brain-Computer Interaction - 2012*, Mar. 2012.

[10] D. J. Krusienski *et al.*, "A comparison of classification techniques for the P300 Speller," *J. Neural Eng.*, vol. 3, no. 4, pp. 299–305, 2006.

[11] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 4, no. 2, pp. R1–R13, 2007.

[12] A. Bashashati, M. Fatourechi, R. K. Ward, and G. E. Birch, "A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals," *J. Neural Eng.*, vol. 4, no. 2, pp. R32–R57, 2007.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, New York, USA: Wiley-Interscience, 2001.

[14] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components – A tutorial." *NeuroImage*, vol. 56, no. 2, pp. 814–825, 2011.

[15] J. R. Pierce, *An Introduction to Information Theory: Symbols, Signals & Noise*. New York, New York, USA: Dover Publications, 1980.

[16] B. Blankertz. (2003) BCI Competition II: Results. [Online]. Available: http://www.bbci.de/competition/ii/results/index.html

[17] M. Kaper, P. Meinicke, U. Grossekathoefer, T. Lingner, and H. Ritter, "BCI Competition 2003—Data Set IIb: Support Vector Machines for the P300 Speller Paradigm," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1073–1076, 2004.

Utjecaj nedostajućih vrijednosti na rezultate dubinske analize podataka

Mario Somek
Zdravstveno veleučilište/Katedra za informatiku, Zagreb, Hrvatska
mariosomek@gmail.com

Sažetak - Repozitoriji baza podataka sadrže velike količine informacija potencijalno korisnih za generiranje novih znanja. Primjena odgovarajuće informacijske tehnologije u području dubinske analize podataka omogućuje utvrđivanje ovisnosti između informacija što pridonosi stvaranju prediktivnih modela korisnih za potrebe poslovnih procesa. Preduvjet za kvalitetnu provedbu analize odgovarajuća je priprema podataka koja podrazumijeva razmatranje nedostajućih vrijednosti u promatranom skupu podataka te njihov utjecaj na kvalitetu analize. U radu će se na hipotetskim primjerima ispitati utjecaj nedostajućih vrijednosti na kvalitetu rezultata provedene dubinske analize primjenom odgovarajuće programske podrške. Opisat će se odabrani model analize, polazni skupovi podataka, korištena programska podrška, tijek analize, prikazati dobiveni rezultati te izvesti zaključak.

I. Uvod

Razvoj tehnologije i terminologije u području spremanja podataka počevši od jednostavnog prikupljanja i analize, omogućio je danas primjenom složenih algoritama pronalaženje pravilnosti i korelacija unutar velikih skupova podataka. Eksponencijalan trend prikupljanja i pohrane informacija predstavlja izazov u pogledu efikasnog spremanja podataka, upravljanja podacima te transformacije odgovarajućim metodama i alatima [1]. Suvremeno poslovanje zasnovano na informacijskom sustavu u kojem središnji dio čini baza podataka zahtjeva od korisnika vještine koje će temeljem postojećih informacija omogućiti generiranje novih znanja i spoznaja primjenjivih u raznim područjima ljudske djelatnosti [5, 6].

Dubinska analiza podataka (DAP) ili rudarenje podacima predstavlja skup metoda i alata kojima je osnovna namjena izvođenje novih znanja iz postojećih repozitorija baza podataka [3, 4]. To je multidisciplinarno područje u kojem inkorporiraju razne tehnike iz područja strojnog učenja, umjetne inteligencije, statistike i upravljanja bazama podataka [2]. Manipulacija velikim količinama podataka u kratkom vremenu moguća je primjenom višeprocesorskih računala i odgovarajućih znanstvenih metoda koje omogućuju otkrivanje međuzavisnosti podataka korisnih za stvaranje prediktivnih modela, a time i kvalitetne podloge za donošenje poslovnih odluka.

Kvalitetnoj provedbi DAP prethodi slijedno izvođenje pojedinih radnji [7], pri čemu je jedna od važnijih priprema podataka koji su predmet analize. Gotovo uobičajeno pri tome prisutne su nedostajuće (engl. Null) vrijednosti, polja u bazama podataka s nepopunjenim, nepotpunim ili pogrešno navedenim vrijednostima pojedinih atributa. Prema E. Rahm i H. Hai Do [13], rezultat je to pogreške pri unosu podataka, nepridržavanja pravila pri unosu atributa ili kodeksa atributa, sintaksne pogreške ili pogreške uzrokovane korisničkom nepažnjom. U Tablici I. pregled je uzroka pogrešaka koji dovode do nedostajućih vrijednosti. Pretpostavka je da uključivanje takovih uzoraka u analizu može dovesti do pogrešnih zaključaka pri generiranju novih znanja osobito ukoliko je prisutnost velika u odnosu na ukupan broj uzoraka u promatranom skupu podataka. U svom radu K. Pandole i N. Bhargava [15] navode nedostajuće vrijednosti kao važan parametar utjecaja na kvalitetu analize podataka te ističu važnost ispravnog pristupa takovim vrijednostima u promatranom skupu kako bi se osiguralo generiranje kvalitetnog znanja i ispravno funkcioniranje baze podataka.

Razvidno je iz navedenih razmatranja da problematika nedostajućih vrijednosti nije zanemariva, stoga ovim radom na nekoliko skupova ulaznih podataka primjenom odgovarajuće programske podrške ispitat će se utjecaj nepotpunih vrijednosti na rezultate DAP primjenom deskriptivnog modela poznatog pod nazivom pravila pridruživanja (engl. Association rule). Opisat će se ukratko model DAP, postupak pripreme podataka i uzorak koji se koristio pri ispitivanju. Navest će se osobine programske podrške korištene pri ispitivanju te na kraju iznijeti rezultati s pripadajućim tabličnim i grafičkim prikazom.

TABLICA I. UZROCI NEDOSTAJUĆIH VRIJEDNOSTI

Uzrok	Opis pogreške
Neupisana vrijednost	Polje podatka je neispunjeno
Pravopisna pogreška	Grad Krapana umjesto Krapina
Pogreška skraćenice	Osigur. umjesto Osig.
Nepotrebno dodana vrijednost	Glavan ml. umjesto Glavan
Spajanje vrijednosti atributa	Glavan Zagreb umjesto zasebno Glavan i zasebno Zagreb

II. Model analize

Ispitivanje utjecaja nedostajućih vrijednosti obavljeno je modelom analize poznatijom pod nazivom pravila pridruživanja, jednoj od važnijih i često korištenih metoda DAP [8]. U praksi u velikim bazama i drugim skladištima podataka javlja se potreba za ispitivanjem

odnosa između pojedinih atributa pri čemu se odnos iskazuje parametrima potpore (engl. Support) i uvjerenosti (engl. Confidence) ili uvjetne vjerojatnosti. Teoretski, međuzavisnost dva promatrana atributa iskazujemo izrazom: Ako Atribut1 sadrži Vrijednost1- tada Atribut2 sadrži Vrijednost2 pri čemu je kvaliteta tog odnosa definirana vrijednostima parametara potpore i uvjerenosti. Ispitivanje pri radu bilo je usmjereno na parametar uvjerenosti te je u odgovarajućem koraku pri provedbi analize obavljeno postavljanje ograničenja navedenog parametra u svrhu pojednostavljenja čitavog postupka i jednostavnijeg razmatranja rezultata. Kod pravila pridruživanja moguće je primijeniti različite algoritme, a pri ovoj analizi primijenjen je algoritam rasta učestalih skupova (engl. FP Growth) [11].

III. POLAZNI PODACI

Polazna osnova za postupak analize bilo je generiranje četiri hipotetska skupa podataka s obilježjima prikazanim u Tablici II. Svaki skup sadržavao je četiri entiteta s pripadajućim numeričkim podacima, vrijednostima atributima u rasponu od 10 do 90, pri čemu je za svaki entitet naveden kvantitativno jednak broj mjerenja. Skupovi su se sastojali od najmanje 20 i najviše 540 mjerenja.

Primjenom odgovarajućih alata programske podrške obavljeno je generiranje meta podataka iz ulaznog skupa uzoraka. Na Slici I. prikazani su „podaci o podacima“ ili meta podaci za najmanji skup uzoraka. Podaci su cjelobrojni, pri čemu postoje dva posebna atributa i četiri redovna. Generiranje meta podataka omogućuje prikaz statističkih podataka za navedene vrijednosti i njihovo područje raspona.

TABLICA II. OBILJEŽJA POLAZNIH SKUPOVA PODATAKA

Naziv skupa	Broj mjerenja	Broj nedostajućih vrijednosti	Broj atributa
skup1	20	0, 2, 4	6
skup2	60		
skup3	180		
skup4	540		

ExampleSet (20 examples, 2 special attributes, 4 regular attributes)

Role	Name	Type	Statistics	Range
id	id	polynominal	mode = id1 (1), least = i	id1 (1), id2 (1), id:
label	ime	polynominal	mode = ime1 (5), least =	ime1 (5), ime2 (5
regular	mjer1	integer	avg = 52 +/- 25.257	[10.000 ; 90.000]
regular	mjer2	integer	avg = 48 +/- 25.047	[10.000 ; 90.000]
regular	mjer3	integer	avg = 39 +/- 25.110	[10.000 ; 90.000]
regular	mjer4	integer	avg = 36.500 +/- 21.343	[10.000 ; 80.000]

Slika I. Prikaz meta podataka iz najmanjeg skupa uzoraka

IV. PROGRAMSKA PODRŠKA

Pri provedbi DAP za potrebe ovog rada korištena je programska podrška Rapid Miner [12] koja omogućuje analizu primjenom raznih prediktivnih i deskriptivnih postupaka: klasifikacijom, regresijom, stablima odlučivanja, segmentiranjem, utvrđivanjem pravila pridruživanja i drugim metodama dubinske analize. Alat je otvorenog koda zasnovan na Java platformi i sadrži velik broj operatora namijenjenih za provedbu predprocesa nužnih za provedbu cjelokupnog postupka izabrane analize. Prema provedenim ispitivanjima kod pojedinih skupina korisnika Rapid Miner postaje u 2010. god. najčešće korišten alat za provedbu DAP [14].

Za uspješno provođenje cjelokupnog postupka stvara se tijek podataka u dizajn pogledu koji grafički prikazuje odabrane operatore, polja postavki za svaki operator te uspostavljene veze između njih. Alat omogućuje i pogled prikaza rezultata kojim se dobije pregled rezultata po provedenoj analizi. Za navedeni alat postoji dobra korisnička i razvojna podrška, mogućnost nadogradnje raznim dodacima u svrhu prilagodbe rada i daljnjeg razvoja.

Za potrebe ovog rada koristili su se operatori iz skupine namijenjene za provođenje podatkovne transformacije; operatori za diskretizaciju i za pretvorbu u binominalne vrijednosti te operatori iz skupine za provedbu modeliranja; frekvencijski operator i operator za utvrđivanje pravila pridruživanja.

V. POSTUPAK ISPITIVANJA

Uspješna provedba analize zahtijevala je pripremu polaznih skupova koja podrazumijeva tablično strukturiranje podataka. Potom, uslijedio je zasebno uvoz u aplikaciju svakog pojedinog skupa podataka s obveznim definiranjem nazivnog i glavnog atributa te vrste podataka ostalih atributa. U daljnjem tijeku analize primjenom operatora za diskretizaciju [9] obavila se pretvorba numeričke vrste u binominalnu vrstu što podrazumijeva distribuciju podataka svakog atributa u dva raspona. Primjerice, kod atributa Mjer2 numeričke vrijednosti distribuirane su u nominalne u dva raspona: Raspon1 [-∞ - 55] i Raspon2 [55 - ∞]. U daljnjem postupku frekvencijski operator omogućio je primjenu algoritma rasta učestalih skupova (engl. FP-Growth) [11], koji pomoću stabla uzoraka utvrđuje česte uzroke ili odnose između pojedinih atributa. Na kraju korišten je operator za stvaranje pravila pridruživanja pri čemu je kao kriterij definiran parametar uvjerenosti s minimalnim iznosom 0,5 u rasponu od 0 do 1.

Prvotno analiza se provela sa skupovima bez nedostajućih vrijednosti, a zatim kod svakog skupa podataka dodavale su se po dvije, a potom po četiri nedostajuće vrijednosti. Ukupno provedena je analiza podataka iz dvanaest zasebnih skupova. Posebno, promatrao se odnos atributa Mjer3 u Rasponu1 [-∞ - 35] i Mjer2 u Rasponu1 [-∞ - 55]. Nedostajuće vrijednosti dodavale su se kod ispitujućih atributa Mjer3 i Mjer2 pri čemu nije narušen prvotni raspon podataka.

Obzirom na malu količinu polaznih podataka provedena analiza kod svakog od dvanaest skupova primjenom programske podrške obavljena je u kratkom vremenu, manje od jedne sekunde.

VI. REZULTATI I DISKUSIJA

Analizom dobivenih rezultata u odgovarajućem pogledu programske podrške na osnovu grafičkog i tekstualnog prikaza utvrdili su se odnosi pojedinih atributa i uvjerenosti pojave. Na Slici II. KK (Kamada-Kawai) strukturnim prikazom [10] prikazana je ovisnost

atributa Mjer2 u Rasponu2 o Mjer3 u Rasponu2 pri čemu vrijedi pravilo 33 s parametrom uvjerenosti 0,875. Pravilo koje proizlazi po provedenoj analizi iskazujemo tekstualno: „uvjerenost da će za navedeni skup podataka Mjer2 biti u Rasponu2 te da će Mjer3 biti u Rasponu2 iznosi 0,875“. Uvjerenost se izražava numerički u rasponu od 0 do 1.

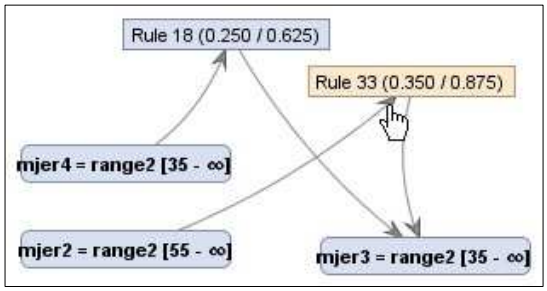
U skladu s iskazanim primjerom na Slici II. rezultati za atribute koji su predmet ispitivanja, Mjer3 u Rasponu1 [-∞ - 35] i Mjer2 u Rasponu1 [-∞ - 55] uz parametar uvjerenosti u iznosu minimalno 0,5, prikazani su u Tablici III te u Grafikonu I.

Kod skupa s najmanjim brojem uzoraka (Uz) vrijednost parametra uvjerenosti najmanja je kod analize s četiri nedostajuće vrijednosti. U odnosu na rezultat analize skupa bez nedostajućih vrijednosti smanjenje je 0,05. Kod skupa sa 60 uzoraka razlika vrijednosti parametra uvjerenosti analize s četiri nedostajuće i analize bez nedostajućih vrijednosti je 0,086. Kod skupa sa 180 uzoraka razlika vrijednosti parametra uvjerenosti analize s četiri nedostajuće i analize bez nedostajućih vrijednosti je 0,027. Kod skupa s najvećim brojem uzoraka razlika vrijednosti parametra uvjerenosti analize s četiri nedostajuće i analize bez nedostajućih vrijednosti je 0,004.

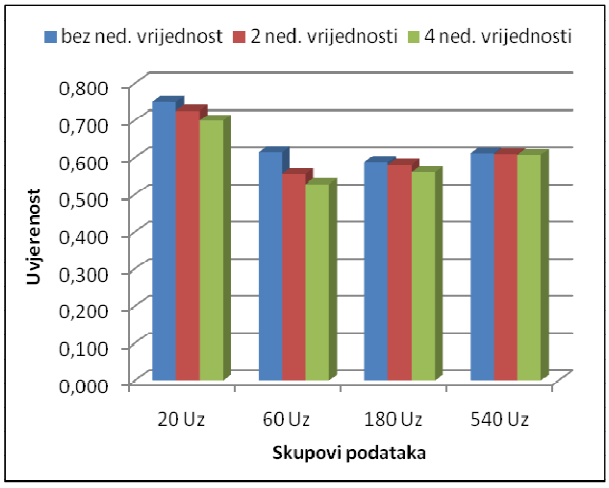
Razmatrajući iskazane promjene parametra uvjerenosti u pojedinim skupovima i prisutnost nedostajućih vrijednosti proizlazi da rezultat konačne analize ovisi o količini uzoraka nad kojim se analiza provodi i broju nedostajućih vrijednost u tom uzorku. Kod skupova s malim brojem uzoraka i velikim brojem nedostajućih vrijednosti rezultat analize biti će nekvalitetan. Kod skupova s velikim brojem uzoraka i relativno malim brojem nedostajućih vrijednosti dobiveni rezultat biti će kvalitetniji. Navedeno potvrđuje rezultat dobiven kod skupa od 540 uzoraka gdje četiri nedostajuće vrijednosti imaju vrlo malen utjecaj na kvalitetu rezultata jer je razlika rezultata analize s četiri i bez nedostajućih vrijednosti minimalna.

TABLICA III. REZULTATI ZA ISPITUJUĆE ATRIBUTE

Nedostajuće (Ned.) vrijednosti	20 Uz	60 Uz	180 Uz	540 Uz
0 ned. vrijednost	0,750	0,615	0,588	0,612
2 ned. vrijednosti	0,727	0,556	0,582	0,610
4 ned. vrijednosti	0,700	0,529	0,561	0,608



Slika II. KK strukturni prikaz parametara pravila pridruživanja za atribute Mjer2 i Mjer3.



Grafikon I. Utjecaj nedostajućih vrijednosti kod polaznih skupova podataka

VII. ZAKLJUČAK

Uspješnost postupka DAP uvjetovana je kvalitetom polaznih podataka koju u mnogome određuju nedostajuće vrijednosti. Provedena ispitivanja primjenom modela Pravila pridruživanja potvrdila su da kvaliteta analize ovisi o kvantitetu nedostajućih vrijednosti u promatranom skupu podataka. Utjecaj određene količine nedostajućih vrijednosti na rezultate analize zanemariv je u velikom skupu i znatno utječe na kvalitetu analize kod skupa s malim brojem podataka. Daljnjim istraživanjem zanimljivo bi bilo ispitati vrijednost graničnog odnosa između kvantitete nedostajućih vrijednosti i podataka promatranog skupa koja dovodi do promjene trenda vrijednosti parametara ispitivanja, u ovom primjeru uvjerenosti, a što bi omogućilo predviđanje utjecaja nedostajućih vrijednosti na DAP.

LITERATURA

[1] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, V. Robles, Machine learning in bioinformatics, Briefings in bioinformatics. Vol 7. No 1., 2005, pp 86-112.

[2] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, D. Steinberg, Top 10 algorithms in datamining, Knowledge and Information Systems, January 2008, Volume 14, Issue 1, pp 1-37.

[3] J. Han and M. Kamber, Data Mining Concepts And Techniques, 2nd Ed., Morgan Kaufmann Publishers, 2006.

[4] I.H. Witten, E. Frank, M.A. Hall, Data Mining, 3rd Ed., Morgan Kaufmann, 2011.

[5] A. Moraru, M. Pesko, M. Porcius, C. Fortuna, D. Mladenic, Using machine learning on sensor data, Journal of computing and information technology – CIT 18, 2010, 4, pp 341-347.

[6] W. Klement, S. Wilk, W. Michalowski, K. J. Farion, M. H. Osmond, V. Verter, Predicting the need for CT imaging in children with minor headinjury using an ensemble of Naive Bayes classifiers, Artificial Intelligence in Medicine 54 (2012), pp 163–170.

[7] The Crows Corporation, Introduction to data mining and knowledge discovery, third edition, ISBN: 1-892095-02-5, 2005.

[8] M. Hoosh Sadat, H. W. Samuel, S. Patel, O. R. Zaïane, Fastest Association Rule Mining Algorithm Predictor, International Canadian Conference on Computer Science & Software Engineering, 2011, pp 43-50.

[9] F. Muhlenbach, R. Rakotomalala, Discretization of continuous attributes, Encyclopedia of Data Warehousing and Mining, JohnWang (Ed.) (2005) pp 397-402.
[10] J. Fagnan, O. R. Zaiane, R. Goebel, Visualizing Community Centric Network Layouts, 16th International Conference on Information Visualisation (IV2012), Montpellier, France, July 2012, pp 10-13.
[11] J. Han, J. Pei, Y. Yin, R. Mao, Mining Frequent Patterns without Candidate Generation: A Frequent-PatternTree Approach, Data Mining and Knowledge Discovery, 8, 2004, pp 53–87.
[12] www.rapid-i.com, (14.1.2013.)
[13] E. Rahm, H. H. Do, Data Cleaning: Problems and Current Approaches, IEEE Bulletin of the Technical Committee on Data Engineering, Vol 23 No. 4, December 2000.
[14] http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html, (14.1.2013.).
[15] K. Pandole, N. Bhargava, Comparison and Evaluation for Grouping of Null Data in Database Based on K-Means and Genetic Algorithm, International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 2, Issue 3, June 2012.

From Digitisation Process to Terminological Digital Resources

S. Seljan*, I. Dunđer ** and A. Gašpar***
* University of Zagreb - Faculty of Humanities and Social Sciences,
Department of Information and Communication Sciences, Zagreb, Croatia
* sanja.seljan@ffzg.hr
** ivandunder@gmail.com
*** ginasplit@yahoo.com

Abstract - Monolingual and multilingual terminology and collocation bases represent valuable additional electronic resources, which can be used in further research, in written communication and in everyday communication. Building of such resources can be supported by terminology extraction tools relying on statistical or language approaches, or on hybrid model, but require considerable human expertise in evaluation and final compilation. The paper describes the whole process: from digitisation of printed material, OCR techniques, sentence alignment and creation of translation memories, up to terminology extraction and evaluation. The performance of tools and applied methodology is assessed through standard statistical measures of precision, recall and F-measure. Experimental results are produced, deficiencies of semi-automatic statistical and linguistic system highlighted and recommendations for further research suggested.

Keywords **– digitization, term and collocation extraction, Multi-Word Unit (MWU), statistical and language approaches, evaluation, English, Croatian**

I. INTRODUCTION

Monolinugal and multilingual terminology and collocation bases represent valuable additional electronic resources, which can be used in further research, in written communication and in everyday communication. The manual compilation of such resources is time consuming, cost-intensive and highly subjective process. Building of such resources can be supported by terminology extraction tools relying on statistical or language approach, as syntactic parsing in [12], or on hybrid model, but require human expertise in evaluation and final compilation.

Translation memory created out of sentenced aligned parallel corpus can be further used for semi-automatic extraction of domain specific terms. Hybrid approach is applied combining: statistical methods based on frequency of term candidates and differences between general and specialised corpora, and linguistic ones based on morphological and syntactic analyses.

Evaluation of extracted terminology and collocation candidates is usually based on precision and recall, but greatly depends on the purpose and usability in practice [9]. In the evaluation process, automatically extracted terms and collocations are evaluated in comparison with manually created reference list based on expert's knowledge and experience and used as „gold standard". The paper describes the whole process: from digitisation of printed material, OCR techniques and its evaluation, through sentence alignment and creation of translation memories.

The lists of automatically produced term candidates are validated by a linguist and compared to the *gold standard* made of manually annotated text. The performance of tools and applied methodology is assessed through standard statistical measures of precision, recall and F-measure. Experimental results are produced, deficiencies of semi-automatic statistical and linguistic system highlighted and recommendations for further research suggested.

II. RELATED WORK

Bases of terminology and collocations can be used independently or integrated, as in [6], [5] with visualizing environment, in machine translation (MT) systems [13], [10] or in information retrieval. Reference [8] has showed that bilingual term entries extracted from domain-specific parallel texts introduce more improvements in MT systems than specialized technical dictionary covering broader domain.

Although multi-word units are composed of two or more orthographic words (linked by dash, conjunction, or blank), they are treated as a single grammatical unit. Multi-word (MW) units can include foreign expressions (e.g., *ad hoc*), prepositions (e.g. *freeze up*, *depend on*), adverbs (e.g., *of course*), idiomatic noun constructions (e.g., *know how*, *per cent*), expressions (e.g., *well being*), as on BNC (British National Corpus) web-page.

Multi-word terminology directly express the concept, as in Sager's lit of requirements [2], does not overlap with other terms, it is lexically systematic, does not contain unnecessary information, and is independent from the context. Terms are congruent to general rules of word-formation. Multi-word terms are composed of two or more words or compound words that are concept-oriented and used in a specific domain.

Collocations are defined a sequence of words or terms that appear more often than would be expected by chance

(e.g. powerful computers/ *strong computers; fast train/ *quick train; fast food/ *quick food; quick meal/ *fast meal; strong tea/ *powerful tea), as in [3]. Collocations represent a high domain of interest in machine translation and considerable effort has been made to perform suitable techniques for collocation extraction from parallel corpora [6], [11]. *Idioms*, i.e. combination of words having figurative meaning, are not included in this base.

As in [3], collocations are characterized by noncompositionality (meaning of the collocation can not be predicted from the meaning of the parts), by non-substitutability (components can not be substituted), and by non-modifiability (not modified through additional lexical material of grammatical transformations). According to [13], collocations are considered to be a subset of *multi-word expressions* that constitute arbitrary conventional associations of words within a particular syntactic configuration.

III. RESEARCH

A. Data Set

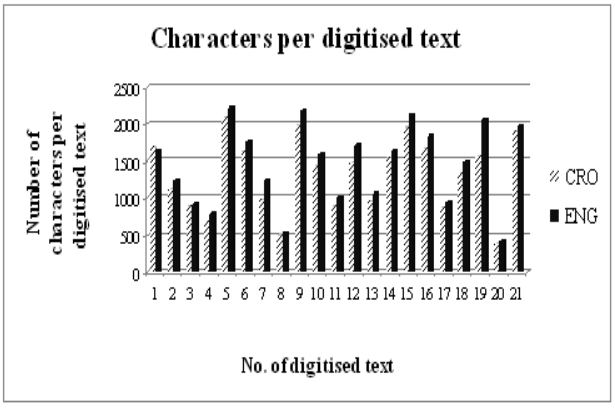
The research is performed on parallel bilingual Croatian-English abstracts relating to the philosophical and sociological topics of society, religion, dignity, freedom, peace, etc. The process includes: digitisation of printed material, OCR techniques and its evaluation; sentence alignment and creation of translation memories; extraction of terminology and collocation using different types of tools based on statistical and language approaches; evaluation and analysis.

This research is made on 21 parallel Croatian-English texts from the philosophical-sociological-religious domain. On average, English texts are composed of 11,97% more characters.

TABLE I. DATA SET

	#characters	***#words***	***#abstracts***	***Arith. mean #characters***
Cro	27 287	3 965	21	1 299,38
Eng	30 354	4 881	21	1 445,43

CHART I. CHARACTERS PER DIGITISED TEXT



B. Methods and Tools

Digitisation and data preparation

For conducting this research, data in form of Croatian and English texts had to be digitised with a scanner. Digitisation means recording, storing and processing content using digital cameras, scanners and computers [1]. It is the process of creating a digital representation of an object, image, document or a signal, enabling them to be stored, displayed, disseminated and manipulated on a computer.

As scanners generate a raster image or a snapshot of an object, Optical character recognition (OCR) software was applied in order to extract, edit, search and repurpose data from scanned objects.

OCR software recognizes text by analysing the structure of the object that needs to be digitised, by dividing it into structural elements and by distinguishing characters through comparison with a set of pattern images stored in a database and built-in dictionaries.

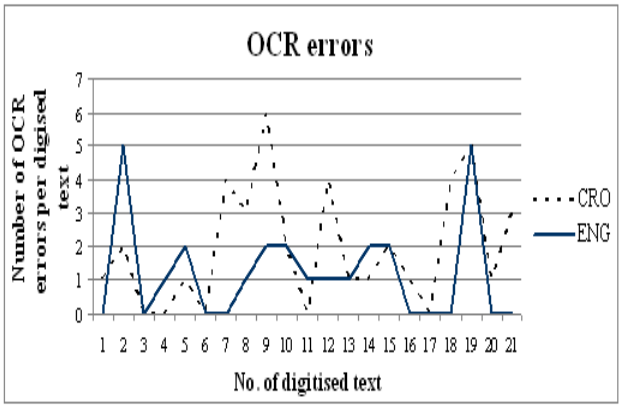
For the purpose of digitisation a HP Scanjet G3110 flatbed scanner was used, set to 300 dpi, grayscale scanning and other default settings. Scanned documents were in A5 format and text was written in Times New Roman font, size 10, standard black font colour on white background.

Optical character recognition was carried out by using Abby Fine Reader 8.0.0.677, which allowed conversion of scanned input texts from bitmap format to encoded text. Errors are unavoidable in advanced computer vision applications such as Optical character recognition, and the noise induced by these errors presents a serious challenge to downstream processes that attempt to make use of such data [7].

Typical errors that occurred during Optical character recognition were misrecognitions of characters, e.g. (,) → (.), (i) → (;), missing whitespace characters ( ) or apostrophes ('), various forms of substitution errors, e.g. (l) → (i), (h) → (l1), (rn) → (m), («)→ ((() as well as space deletion, e.g. (*među crkvenim*) → (*međucrkvenim*) and insertion errors, e.g. (*neučinkovitost*) → (*ne učinkovitost*).

The most frequent OCR errors in Croatian were substitution errors (e.g. (l) → (i)) and space deletion, where two words were unified in one by mistake. The most frequent OCR errors (83%) in Croatian were substitution errors (e.g. (l) → (i)) and space deletion, where two words were unified in one by mistake ((*među crkvenim*) → (*međucrkvenim*)). The most frequent OCR errors (87%) in English were substitution errors (e.g. («)→ ((())) and missing apostrophes (').

CHART II. OCR ERRORS



Arithmetic mean of OCR errors in digitised Croatian texts is 1,95 and in digitised English texts 1,19. OCR errors have an impact on later-stage processing and usability, so all scanned texts were post-edited afterwards.

Extraction Tools

Multi Term Extract - SDL Trados

The tool used for terminology extraction MultiTerm Extract offers a variety of extraction possibilities (set up a statistical threshold, min and max term length, min translation frequency, and max number of translations). It can be used for the extraction from monolingual or bilingual translation memories. For each term candidate the program suggests a number of probable translations presented in a term candidate list on a user-friendly graphic interface.

During the term extraction process it uses stop-list containing functional words, such as articles, conjunctions, prepositions, etc. in order to refine the suggested list of term candidates. After validating terms and their translations it is possible to export them to MultiTerm XML or a tab delimited format. This program is interrelated with other programs in SDL Trados package designed to assist to the translators before and during translation, but can be also used to create independent terminology list.

SDL PhraseFinder

SDL's PhraseFinder uses sophisticated algorithms to understand the structure of a language and identify candidate terms.

Term extraction is based on linguistical and statistical approach and does not support Croatian language. It analyzes the content of monolingual and bilingual files identifying single-word and multiple-word terminology candidates which can be viewed, edited and validated in the PhraseFinder screen. It supports various file formats (HTML, HTM, RTF, SDLX Translation Document (ITD file), SDLX Translation Memory (MDB file), TXT) and languages (English, French, German etc.).

The extracted terminology candidates can be reviewed to identify valid terms and ordered by frequency, rank or alphabetically. Validated terms can be saved in (.phr) format and exported in TXT format for use in other applications. Filter and display settings refer to the following parameters: term length, number of context sentences, ignoring of function words, numbers, first or all uppercases, unfound text, duplicate capitalisation and sort of term candidates.

NooJ

Term extraction was performed by a linguistically-based environment NooJ developed by Silberztein in 2004 at the University Franche-Comté Paris, France. NooJ's multilingual engineering platform provides specific tools for the formalization of linguistic phenomena at the level of orthography, morphology, syntax, semantics, and lexicon. It can parse and process large texts and corpora, lounge sophisticated queries in order to produce various results (concordances, statistical analyses, information extraction, etc.).

A set of local regular grammars is required to perform multi-word units extraction. Based on large-coverage dictionaries and grammars, NooJ linguistic engine can process texts and corpora in real time regardless of file formats (varying from MS Office, HTML and PDF to XML documents).

IV. EVALUATION LISTS

Evaluation of automatically or semi-automatic created lists is performed on the basis of comparison with regard to human-created term list based on human knowledge, experience and intuition. Results are analyzed using standard measures of recall, precision and F-measure.

Reference list

For the purpose of this research a reference list is created in order to assess tools' performance. It is made on the basis of 21 abstracts, out of 42 in total.

The human reference lists consist of 491 terms in English and 506 terms in Croatian. These reference lists contain not only terminology specific for the subdomain, but also terms that more frequently appear but belong to everyday language. The lists contain mainly nominal multi-word units and very rarely collocations.

Structuring the reference list is a demanding task because some expressions in Croatian language are paraphrased in English or translated by anaphoric expressions.

Although, the focus of expert's extraction is to detect bilingual phrase pairs which meet linguistic and terminological criteria and domain-specific coverage, some terms in English language were rejected because of a discrepancy caused by Croatian syntax.

Disregarded terms are for example (*pripadnici društva / members of its society; doseg svjetovnih odgovora / scope of their worldly responses; poprište se njihova sukoba / focus of conflict; svijet i za društvo / world and society; počinjena ili proživljena zla / evil that has been done or experienced; sjećanje na pretrpljenu patnju / memory of suffering that has been experienced*).

The reference list for Croatian language, was supplemented by 89 false positives, i.e. terms not found in the reference list, and consists of 506 terms in total.

The reference list for English language contains 491 terms, including 95 false positives. Table 2 presents number of n-grams and their distribution in the reference list for both languages. In both languages there is the highest number of 2-grams and 3-grams, whiled in English there is higher number of 4-grams and 5-grams.

TABLE II. N-GRAMS OF THE REFERENCE LIST

N-grams total	2-grams	3-grams	4-grams	5-grams and more	# terms & collocations
Cro	254	168	68	16	**506**
Eng	166	167	91	67	**491**

The frequency of terms decreases with their length in both languages. There is bigger proportion of n-grams in Croatian language comparing to English, except for 4-, and 5 grams. Noun phrases prevail in both languages.

NooJ list

Linguistic approach is based on creation of 22 local grammars for Croatian term candidate extraction spanning up to 5-grams, where N stands for noun, A-adjective, S-preposition and C-conjunction, and on 29 local grammars for English term extraction including 7-grams, where POS are as follows: N-noun, A-adjective, PREP-preposition, DET-article and CONJ-conjunction. POS tagging is automatically performed by using NooJ tool for language engineering.

Table 3 presents local grammars for Croatian and English language.

TABLE III. N-GRAMS OF THE REFERENCE LIST

N-grams total (22)	**Local grammars for Croatian language**		
2-gr.	<N><N>;	48	Most common
	<A><N>;	208	
3-gr.	<A><N><N>;	29	
	<A><A><N>;	29	
	<N><S><N>;	12	
	<N><C><N>;	36	
4-gr.	<A><C><A><N>;	17	
3-5-gr.	<N><N><N>; <N><N><A><N>; <A><N><C><A><N>; <N><A><C><A><N>;	Least common	
N-grams total (29)	**Local grammars for English language**		
2-gr.	<N><N>;	10	
	<A><N>;	178	
3-gr	A><A><N>;	21	Most common
	<N><PREP><N>;	97	
4-gr.	<N><PREP><DET><N>;	26	
	<A><N><PREP><N>;	28	
	<N><PREP><A><N>;	23	
	<A><CONJ><A><N>;	14	
5-gr.	<N><PREP><DET ><A><N>;	15	
	<A><N><PREP><DET><N>;	15	
6-gr.	<N><PREP><DET><N><PREP><N>;	5	
3-7-gr.	<N><A><N>;	Least common	
	<N><A><PREP><N>;		
	<N><PREP><DET><N><N;		
	<N><PREP><N><PREP><DET><N>;		
	<N><A><PREP><N><DET><PREP><N>;		

As presented by table 3, 2 and 3-grams are almost equally represented term candidates in Croatian and English. While 2-grams follow the same language pattern of noun phrases possibly proceeded by adjective, 3-grams differ in language patterns. 4-5grams are significantly more represented in English, due to analytic language structure.

Local grammars are created on the basis of the most frequent POS patterns in the reference list. Local grammars differ in number because of the usage of articles and prepositions in English and inflection in Croatian language. Also, a small, high-level priority dictionary for English language is created in order to avoid ambiguities. For example the word *'a'* which always refers to the determiner not the noun (in abstracts) is stored in high-priority dictionary together with other words which correspond to more than one lexical entry, as presented in table 4.

TABLE IV. HIGH- PRIORITY DICTIONARY FOR ENGLISH LANGUAGE

a, DET	for, PREP
an, DET	on, PREP
the, DET	under, PREP
while, CONJ	down, PREP
if, CONJ	above, PREP
in, PREP	after, PREP

Lexical constraint is set for the noun to be in the nominative case at the beginning of the phrase (<N+Case=Nom-Type=Kr>).

V. RESULTS

Statistical analysis is based on statistical measures of *precision* (the proportion between valid automatically extracted terms and all automatically extracted terms), *recall* (the proportion between valid automatically extracted terms and manually extracted terms), and the *F measure* (the ratio between *precision* and *recall*). The

results of automatic term extraction by NooJ are presented in Table 5.

TABLE V. THE RESULTS OF AUTOMATIC TERM EXTRACTION

	NooJ		MultiTerm Extract		PhraseFinder	
	Cro	Eng	Cro	Eng	Cro	Eng
No. of terms	1004	1117	118	118	-	325
Validated terms	316	240	42	42	-	122
Precision (%)	31%	21%	**36%**	**36%**	-	**38%**
Recall (%)	**62%**	**49%**	8%	9%	-	**25%**
F1 (%)	**42%**	30%	13%	14%	-	**30%**

Table 6 presents the most and the least frequent POS-patterns of the reference list and list of validated Croatian and English terms obtained automatically using local grammars. POS patterns which frequency is lower or equal to 5% are: NPAN, ANPDetN, NPDetNPN, NPDetANPN.

TABLE VI. POS-PATTERN STATISTICS OF THE REFERENCE LIST AND LIST CREATED BY NOOJ

Reference list			**List created by NooJ**	
	Cro	Eng	Cro	Eng
AN	186 (37%)	139 (28%)	208 (42%)	178 (37%)
NN	47 (13%)	25 (5.0%)	48 (9.8%)	10 (2.1%)
NPN	83 (17%)	76 (15.4%)	12 (2.4%)	97 (20.6%)
NConjN	44 (8.6%)	43 (9%)	36 (7.3)	-
AAN	27 (5.3%)	22 (4.4%)	29 (5.9%)	21 (4.4%)
ANN	17 (3.3)	2 (0.4%)	29 (5.9%)	4 (0.8%)
AConjAN	2 (0.3%)	3 (0.6%)	17 (3.4%)	14 (2.9%)
ANPN	6 (1.1%)	7 (1.4%)	-	28 (5.9%)

VI. DISCUSSION

Human reference list consists namely of <AN> pattern in Croatian and English followed by <NPN> and <NN> pattern in Croatian and <NPN> and <NConjN> in English. NooJ offer more than 1000 term candidates consisting of indicated local grammars according to which all term candidates consisting of e.g. <AN> patterns for English or Croatian were extracted regardless statistical appearance or relevance. Therefore, it offered high number of candidates which were retrieved by indicated local grammars.

In regard to the number of discarded term candidates NooJ offers better for Croatian language. Language tool NooJ has significantly higher recall comparing to statistically-based tools, denoting that number of suggested term candidates overlap more with reference lists in both languages, as they mostly consist of similar language patterns.

Rule-based candidates highly depend on their identification in the text retrieved by appropriate local grammar. When comparing language-based and statistical-based tools for terminology extraction, language tool offers significantly higher number of candidates mostly composed as noun phrases: <AN> pattern in Croatian 42% and in English 37%, followed by <NN> pattern 9,8% in Croatian and in English by <NPN> pattern (20,6%). Statistically-based extraction tools would offer much better results in bigger domain specific corpora.

Verb-XX collocations represent challenge for statistically-based tools, but since the abstracts belong to various domains of philosophy, sociology and religion and are not voluminous, the collocations are not significantly retrieved as candidates.

As MultiTerm Extract tool is based on statistical frequency, it offered relatively good results consisting of terms which would be suitable for the general language, but not for this domain consisting of several subdomains.

The abstracts collected for this research belong to various subdomains and therefore the characteristic terms do not appear frequently, and are not retrieved by statistical term extraction tool. Term extraction process offered equal results when extracting from two monolingual corpora (Croatian and English) and when extracting from the bilingual translation memory.

The ratio between the extracted term candidates is 12,5:1 indicating that use stop word lists improves significantly the results. Stop word lists eliminates term candidates beginning with commas, prepositions, conjunctions and other functional words appearing at the beginning of tem candidates.

PhraseFinder tool, which is based on hybrid model offered higher score for precision and relatively high value for recall. It offers possibility for term extraction only for English, and not for Croatian language.

VII. CONCLUSION

As this collection of 21 abstract is not voluminous, statistical extraction offered less good results, especially MultiTerm Extract tool which is only statistically-based. Language-based tool NooJ gave the best results, while PhraseFinder which is based on hybrid model gave medium results for English list. PhraseFinder does not offer the possibility for Croatian terminology extraction.

For better results, more extensive corpus in the specific domain should be compiled, in order to achieve significantly better results.

The final compilation of terminology and collocation bases still highly depends on human expertise. Presented tools can serve as assisting tool, but are useful only when using voluminous texts from the specific domain and when reference list is created as intersection of several experts, with predefined purpose of use.

REFERENCES

[1] Lopresti, D. Optical Character Recognition Errors and Their Effects on Natural Language Processing, International Journal on Document Analysis and Recognition, vol. 12, no. 3, 2009, pp. 141-151.

[2] Love, S. Benchmarking the performance of Two Automated Term-extraction systems: LOGOS and ATAO. Mémoire de maîtrise, Univ. de Montréal, 2000.

[3] Manning, Chiristopher. D.; Schütze, H. Foundations of Statistical Natural Language Processing. MIT, 2002.

[4] Seljan, Sanja; Gašpar, Angelina. First Steps in Term and Collocation Extraction from English-Croatian Corpus // Proceedings of 8th International Conference on Terminology and Artificial Intelligence (TIA2009). Toulouse, France, 2009.

[5] Seretan, V; Nerima, l; Wehrli, E. A tool for multi-word collocation extraction and visualization in multilingual corpora. In Proceedings of the Eleventh EURALEX International Congress (EURALEX), 2004, p. 755–766.

[6] Seretan, Violeta. An integrated environment for extracting and translating collocations. Proceedings of 5th Corpus Linguistics Conference, Liverpool, UK, 2009.

[7] Smolčić, J.; Valešić, A. Legal Contexts of Digitization and Preservation of Written Heritage, INFuture2009 – Digital Resources and Knowledge Sharing, Zagreb, pp. 87-94, ISBN: 978-953-175-355-5

[8] Tatsuya, Izuha. Machine Translation Using Bilingual Term Entries Extracted from Parallel Texts. Systems and Computers in Japan, 36(8), 2005, pp. 23-30.

[9] Thurmair, G. Making Term Extraction Tools Usable. Joint Conference of 8th European Association for Machine Translation (EAMT) and the 4th Controlled Language Applications (CLA) Workshop, EAMT- CLAW, 2003.

[10] Turcato, Davide. Automatically Creating Bilingual Lexicons for Machine Translation from Bilingual Text. Proceedings of COLING, 1998. Bilingual Lexicons for Machine

[11] Wanner, Leo; Bohnet, Bernd; Giereth, Mark; Vidal, Vanesa. The first steps towards the automatic compilation of specialized collocation dictionaries. Application-Driven Terminology Engineering: Special issue of Terminology 11:1, 2005, pp. 143–180.

[12] Wehrli, Eric; Seretan, Violeta; Nerima, Luka. Sentence analysis and collocation identification. Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE), Beijing, China, 2010, pp. 27–35.

[13] Wehrli, E.; Seretan, V.; Nerima, L.; Russo, L. Collocations in a Rule-Based MT System: A Case Study Evaluation of Their Translation Adequacy, EMT, 2009.

[14] L'Homme, M—C. Hee S.B. A Methodology for Developing Multilingual Resources for Terminology. Proceeding of LREC - Language Resources and Evaluation, 2006.

[15] Drouin, P. Term extraction using non-technical corpora as a point of leverage. In Terminology, 2003, vol. 9, no 1, p. 99-117

[16] Barlow, M.: Collocate 1.0: Locating collocations and terminology. TX:Athelstan, 2004.

[17] L'Homme, Marie-Claude and Hee Sook Bae. A Methodology for Developing Multilingual Resources for Terminology. In Proceeding of LREC 2006. Language Resources and Evaluation, 2006.

Comparative Study of Color Histograms as Global Feature for Image Retrieval

Vedran Ljubovic, Haris Supic
vljubovic@etf.unsa.ba, hsupic@etf.unsa.ba
University of Sarajevo, Faculty of Electrical Engineering Sarajevo
Zmaja od Bosne bb
71000 Sarajevo, Bosnia and Herzegovina

Abstract- Color histograms are one of the earliest and best known image features used in Content-Based Image Retrieval (CBIR). There is a wealth of scientific work on this topic. However, different papers vary in the specific ways of determining histograms and distance between them. In this paper authors attempt to classify various types of histograms used in literature and compare them using contemporary datasets and metrics for evaluation. Histograms are compared based on their retrieval performance as well as resource usage.

Keywords: **content-based image retrieval, histograms, histogram distance, color spaces, query-by-example**

I. Introduction

Color histograms are one of the oldest and best known global features used in image retrieval [1]. They are often used as a baseline to compare other retrieval methods with. Recent research suggests that global color histograms provide decent retrieval performance when compared to other retrieval techniques [2].

There is a wealth of scientific work on the topic of histograms. However, approaches to dimensionality reduction, color spaces, distance calculation etc. differ. There is a lack of studies with quantitative comparison of various methods for calculating color histograms and distance between them.

Use of common datasets with defined ground truth and standard metrics for evaluating various retrieval methods is an important issue in CBIR [3]. Such datasets and metrics have become available only relatively recently. Therefore, it is of interest to evaluate various methods based on histograms in this new context.

Another often neglected issue is that of resource usage. The progress of multimedia technology did not make that issue any less relevant today [4].

In this paper we will attempt to determine the optimal approach for calculating color histograms and their distance, from perspective of retrieval performance as well as resource usage. We hope to achieve this by performing a large number of query-by-example (QbE) requests on referential annotated datasets and calculate average retrieval performance. Our focus is on the use-cases of personal image collection and Internet image search engines.

II. Previous Work

Earliest work on using color histograms in image retrieval is [1]. It uses histograms in RGB color space. RGB model is convenient for image representation in computers because it matches the way images are displayed on CRT monitors. Other papers on color histograms also use RGB color model (e.g. [2]).

However, RGB model doesn't correspond to the way humans perceive color. HSV color space is explicitly designed to model human color perception, and is therefore used in most papers on histograms as a global feature in CBIR (e.g. [5] [6] [7] [8] [9]). It is also used for the Scalable Color Descriptor (SCD) feature of MPEG-7 standard for multimedia content description [10]. A closely related color model is HSL (also called HSI).

Another model often used in CBIR is Y'UV color model, also known as $Y'C_BC_R$ [11] [12] [13] (not to be confused with CIE YUV model). Primary use of this model is in digital television and popular image formats such as JPEG, where it allows for efficient compression of color information. This suggests that histograms in YUV model could have high tolerance for quantization. Formula for direct conversion from RGB to YUV ($Y'C_BC_R$) is given in ITU-R BT.601 standard and used in JPEG images. Closely related models are YIQ [6] (used in older versions of NTSC format) and $Y'P_BP_R$.

None of the above color spaces is uniform. Papers [14] and [15] suggest that using a uniform color space such as CIE L*a*b* or Luv should deliver superior retrieval performance. However, conversion from RGB to Lab requires calculation of cube root, making it computationally expensive.

Second problem in histograms is that of dimensionality reduction. 3D histograms are impractical and use a lot of memory. Earliest work on histograms [1] uses the following approach: Each of color channels is first quantized by taking a number of most-significant bits. Then these bits are

concatenated to form an index into a single histogram [6] [7]. In this paper we call such method *combined histograms*.

For example, a commonly used quantization scheme is HSV 16*4*4 [10]. Here, value for Hue (H) is quantized as a 4-bit integer (value in range [0,15]), while Saturation (S) and Value (V) are quantized as 2-bit integers (values in range [0,3]). The resulting values are combined into a single histogram index I using the following formula:

$$I = 16 \cdot H + 4 \cdot S + V$$

We see that I is an 8-bit integer (value in range [0,255]) and therefore a histogram with 256 bins is obtained. Some papers (e.g. [8], [9]) use more complex formulas for quantization of each color channel.

Another approach is to calculate separate histogram for each of color channels. Those separate histograms are then concatenated into a single feature vector [5] [11] [16]. For example, if histograms in RGB color space use 32 bins of 1 byte each, total feature vector would have 96 bytes. We call this method *split histograms*.

Typically, normal (non-cumulative) histograms are used, although [5] argues for using cumulative histograms.

Thus obtained histograms will contain values in range [0,N] where N is the total number of pixels in image. In order to efficiently represent it in memory, such histogram needs to be normalized (divide all members with maximum) and quantized. It is typical to quantize each bin to one byte (8 bits) though lower quantizations may be preferable [10].

Also, many of the values in histograms are zeros, suggesting that a simple run-length coding could result in a more efficient storage [5].

Various transforms can be applied to further decrease the size of feature vector. A number of papers (e.g. [10]) use Haar transform. Reference [5] proposes storing only the first three moments of the histogram.

Finally, there must be a method for calculating distance between two feature vectors. Most papers use simple Manhattan distance (L1) or Euclidean distance (L2) [5] [11] [12] [13], although later work suggests that optimal distance metrics for histograms may be Jeffrey divergence (Jensen-Shannon divergence) [2] or histogram intersection [9] [15].

III. Methodology

Use of standardized, annotated datasets is an important issue in CBIR. [3] To this purpose we used four datasets consisting of photograph collections that are often cited in literature: Wang SIMPLYcity [18], UW [19], UCID [20], and MIRFLICKR-25000 [21].

A common metric for comparing retrieval methods is Precision-Recall graph. Two methods are compared by overlaying their graphs [21]. This however is impractical for comparing a large number of retrieval methods in many configurations across several datasets. Preferred approach is to have a simple numerical metric for each query, which is then averaged across a large number of queries. After evaluating literature, we chose the following metrics:

Mean Average Precision (MAP) - mean value for precision is calculated across the 11 usual points in Precision-Recall graph [21] [22]. This is a standard measure used by NIST in its TREC contest.

AP15 - average value for precision in the first 15 retrieved images [21]. This metric simulates a search engine which returns results in pages of 10-20 results. Research suggests that most users don't look beyond the first page [23].

Average Normalized Modified Retrieval Rank (ANMRR) is explained in [6] and [10]. This is the standard metric used in ImageCLEF competition.

Each image in dataset is annotated with a number of tags. A retrieved result is considered relevant if it has at least one tag in common with query image. We wanted to also gauge the ability of tested methods to quickly retrieve results of higher relevance. To this purpose we introduce *relevance score* as number of tags that are common between query image and retrieved image, divided by number of tags in query image:

$$Relevance = \frac{Nr.\,common\ tags}{Nr.\,query\ tags}$$

Relevance is a value in range [0,1] where higher value corresponds to higher relevance of each result. Thus *weighted precision* (WP) can be calculated using the following formula:

$$WP = \frac{\sum_{i=1}^{N} Relevance_i}{N}$$

where N is number of retrieved images. This allows us to calculate *average weighted precision (AWP15)* as a metric comparable to AP15 given above.

A large number of experiments were performed using open-source tool ETFImageSearch which can be downloaded from website: http://f.etf.unsa.ba/imagesearch Experimental results can be found at the same website (section Files).

When forming conclusions we looked for results where all metrics agree to a large extent across most datasets. We've found that there isn't a big difference between AP15 and AWP15. We've also found that results from AP15/AWP15 are more consistent across all four datasets as compared to MAP and ANMRR.

IV. Experimental Results

Firstly, we attempted to determine the optimal **number of bins** for *combined histograms*. For simplicity, we tested

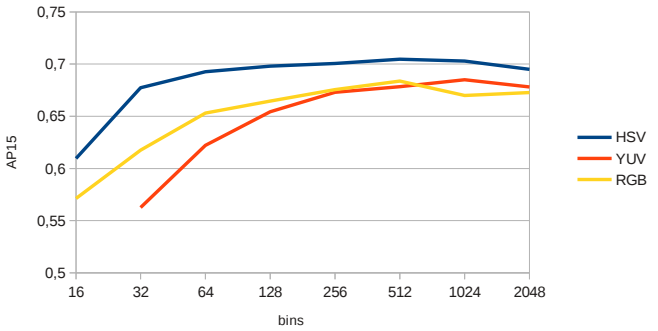


Figure 1: Relationship between number of bins and retrieval performance. Graph shows average precision (AP15) in Wang1000 dataset for histograms in HSV, YUV and RGB color spaces using optimal quantizations for each color space. Very similar results were observed in other datasets as well.

histogram sizes that are a power of two.

Histograms with 256 and 512 bins on average give the best performance across all datasets. Histograms in HSV model also perform well with 128 bins. (Figure 1) Using a larger number of bins (1024 and more) gives inconsistent results. Results for 4096 and more bins are omitted from Figure 1 due to their inconsistent nature, but they can be found on previously mentioned project website. Therefore using such large histograms is not advised and generally there are no benefits above 4096 bins.

Histograms with less than 256 bins continue to give good results and only at below 64 bins there is a considerable decrease in retrieval performance.

Next, we made a comparison of **color models** used. In Wang, UW and UCID datasets the best results are obtained using HSV color model, followed by RGB and YUV color model. HSV also shows better tolerance of quantization and better performance with lower number of bins (Figure 1). Only in MIRFLICKR dataset HSV color model using combined histogram gives inexplicably poor results. This anomaly requires further research.

We've found no statistically relevant difference between YUV, YC_BC_R and YIQ models. For the rest of this paper only YUV color space will be considered. Results for HSV and HSL are similar, and decision on which is better is inconclusive. We've found that CIE Lab model doesn't give a sufficient boost in performance to warrant its much more complex histogram calculation.

Further we tested different **quantization schemes**. RGB space doesn't favor any particular channel when it comes to quantization. With Lab, Luv and YUV spaces greater quantization of the first (luminance) channel is slightly preferred. Thus optimal results in YUV model were obtained with 4*8*8 quantization giving 256 bins. In HSV and HSL color spaces smaller quantization of first channel (hue) is preferred. Optimal schemes were 32*4*4 (512 bins), 16*4*4 (256 bins, also used in [10]), 8*4*4 (128 bins), 4*4*4 (64 bins), 4*2*4 (32 bins).

Custom quantizations described in [8] and [9] give excellent results that sometimes exceed all other types of histograms. However, increase in complexity doesn't justify their use. Histograms where sizes per channel are a power of two can be calculated using binary shifting which is a very cheap operation.

Next issue tested was use of **split vs. combined** as well as **normal vs. cumulative** histograms. Split histograms are of interest because they result in a smaller feature vector. Cumulative split histograms appear to give very good results when Euclidean or Manhattan distance metric is used. However, when optimal distance metric is used for each given type of histogram (as outlined below), combined histograms consistently outperform normal split histograms, which in turn outperform cumulative split histograms.

This proves true even when a very large level of quantization is used. Combined histogram with 32 bins in HSV model (4*2*4) outperforms all kinds of split histograms in every metric, whereas the smallest split histogram with usable performance uses 8+8+8 bins.

Cumulative combined histograms display a sharp decrease in retrieval performance regardless of distance measure used.

An important issue revealed by our testing was that of **optimal distance metric** (Figure 2). 15 distance metrics described in [24] were tested with all described types of histograms. We've found that the optimal distance metric for all types of histograms is Matsushita distance. However this distance metric is fairly computationally intensive, since it requires calculation of square root per each bin. In applications where this is an issue, Bray-Curtis distance (and the very similar Soergel distance) provide excellent results.

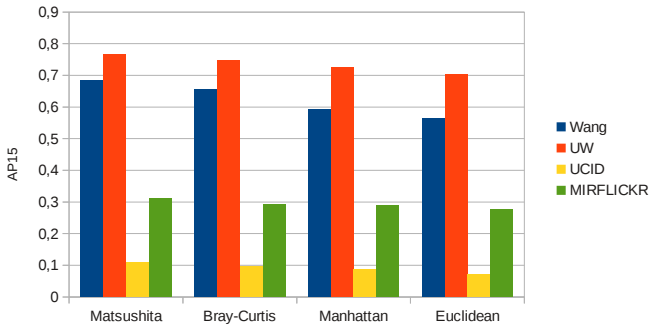


Figure 2: Comparison of distance metrics. Graph shows average precision (AP15) for RGB combined histogram with 512 bins. Each dataset is shown in different color. Similar results were observed with other color spaces.

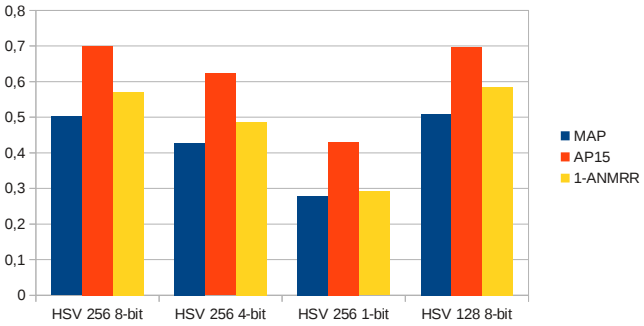


Figure 3: Effect of bin quantization on retrieval performance. Tests were performed on Wang1000 dataset with a combined histogram in HSV model, using histogram with 256 bins quantized to 8, 4 and 1 bit per bin. For comparison, histogram with 128 bins is given. Three retrieval metrics are depicted with different colors (since ANMRR gives scores where less means better retrieval, we display 1-ANMRR). Similar results were observed with other color spaces and datasets.

With cumulative split histograms, best results are usually obtained using Euclidean and Manhattan distance, although results are inconsistent. Bray-Curtis as well as Bhattacharya distance sometimes give better results.

Finally, we explore the issue of **histogram quantization**. In our initial test, each histogram bin was represented as a 64-bit floating point number (double). When bins are quantized as 8-bit integers, there is no detectable loss in retrieval performance. However, further quantization gives a sharp decrease in performance (Figure 3). Histograms with 256 4-bit bins always perform below histograms with 128 8-bit bins. Quantization to 1 bit per bin, as suggested in [10], gives very poor performance.

Optimal distance metrics for 4-bit histograms remain Matsushita and Bray-Curtis/Soergel distance.

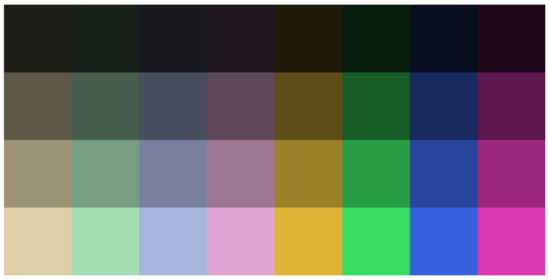


*Figure 4: Palette of colors obtained when HSV color space is quantized to 32 values using 4*2*4 scheme.*

V. Conclusions

On average, the best retrieval score across all datasets is obtained using combined histogram in HSV color space with 256 bins (using 16*4*4 color quantization scheme). Such histogram is also used as an image descriptor in MPEG-7 standard [10]. Each bin is then normalized and quantized to 8 bits (1 byte) per bin. Best distance calculation approach is Matsushita distance.

Described method gives a feature vector of 256 bytes per image. In certain applications this can be considered too large. We've found that the best approach to reduce feature vector size is to reduce the number of bins per histogram. Given feature vector of less than 64 bytes, the best retrieval performance is obtained with HSV combined histogram using 4*2*4 quantization and 8 bits per bin.

The only exception to this is found in MIRFLICKR dataset where histograms in YUV color space perform better than HSV and HSL. This anomaly requires further research.

We've found that using split or cumulative histograms is not beneficial. Most papers arguing for their use are making a comparison based on Euclidean or Manhattan distance, which is a suboptimal distance metric for combined histograms. When best distance metric is used for each type of histogram, combined and non-cumulative histograms always outperform split and cumulative histograms.

This conclusion can be easily understood since combined histograms give a statistic of human-perceptible colors in an image, while split histograms observe each channel separately. For example, HSV combined histogram with 32 bins gives a count of common named colors such as orange, pink, brown etc. (Figure 4) while this information is lost in a split histogram. The problem of histogram "shifting" described in [5] is resolved by using a better distance metric. Further research in the area of distance metrics for histograms could yield even better results and is a promising target for futher research.

References

[1] M.J. Swain, D.H. Ballard, "Color Indexing", *International Journal of Computer Vision*, 1991;7(1), pp. 11-32.

[2] Thomas Deselaers, Daniel Keysers, Hermann Ney. "Features for image retrieval: an experimental comparison", *Information Retrieval*, 11.2 (2008), pp. 77-107.

[3] Henning Müller, Stéphane Marchand-Maillet, Thierry Pun, "The truth about corel-evaluation in image retrieval", *Image and Video Retrieval* (2002), pp. 38-49.

[4] Vedran Ljubovic, Haris Supic, "Issue of resource usage in content-based image retrieval algorithms", *Proceedings of BIHTEL 2012*, Sarajevo, 2012.

[5] Markus Stricker, Markus Orengo, "Similarity of color images", *Proc. SPIE Storage and Retrieval for Image and Video Databases*, Vol. 2420, No. 381-392, 1995.

[6] Savvas A. Chatzichrisofis, Konstantinos Zagoris, Yiannis S. Boutalis, Nikos Papamarkos, "Accurate image retrieval based on compact

composite descriptors and relevance feedback information", *International Journal of Pattern Recognition and Artificial Intelligence*, 24, no. 02 (2010), pp. 207-244.
[7] Shamik Sural, Gang Qian, Sakti Pramanik. "Segmentation and histogram generation using the HSV color space for image retrieval.", *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 2, pp. II-589. IEEE, 2002.
[8] Zhenhua Zhang, Wenhui Li, Bo Li. "An improving technique of color histogram in segmentation-based image retrieval", *Information Assurance and Security, 2009. IAS'09. Fifth International Conference on*, vol. 2, pp. 381-384. IEEE, 2009.
[9] Chong-Wah Ngo, Ting-Chuen Pong, Roland T. Chin, "Exploiting image indexing techniques in DCT domain", *Pattern Recognition 34*, no. 9 (2001), pp. 1841-1851.
[10] Bangalore S. Manjunath, J-R. Ohm, Vinod V. Vasudevan, Akio Yamada, "Color and texture descriptors", *Circuits and Systems for Video Technology, IEEE Transactions on 11*, no. 6 (2001), pp. 703-715.
[11] Zhe Ming Lu, Su-Zhi Li, Hans Burkhardt, "A content-based image retrieval scheme in JPEG compressed domain", *International Journal of Innovative Computing, Information and Control*, Vol 2, No 4, pp. 831-839, August 2006.
[12] Ramadass Sudhir, Lt Dr S. Santhosh Baboo, "An Efficient CBIR Technique with YUV Color Space and Texture Features", *Computer Engineering and Intelligent Systems* 2, no. 6 (2011), pp. 85-95.
[13] Tienwei Tsai, Yo-Ping Huang, Te-Wei Chiang, "Fast image retrieval using low frequency DCT coefficients", *Proceedings of the 10th conference on artificial intelligence and applications*, Kaohsiung, Taiwan. 2005.
[14] K. Konstantinidis, A. Gasteratos, I. Andreadis, "Image retrieval based on fuzzy color histogram processing", *Optics Commun.* 248(4-6) 15 (2005), pp. 375-386.
[16] Gerald Schaefer, "CVPIC colour/shape histograms for compressed domain image retrieval", *Pattern Recognition* (2004), pp. 424-431.
[17] A. Jain, A. Vailaya, "Image Retrieval using Color and Shape", *Pattern Recognition*, vol. 29, no. 8, pp. 1233-1244, 1996.
[18] James Z. Wang, Jia Li, Gio Wiederhold, "SIMPLIcity: Semantics-sensitive Integrated Matching for Picture Libraries", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 23, No 9, pp. 947-963, 2001.
[19] *Annotated groundtruth database*. Department of Computer Science and Engineering, University of Washington. Retrieved on February 13th, 2013, from: http://www.cs.washington.edu/research/imagedatabase/groundtruth/
[20] George Schaefer, M. Stich, "UCID - An Uncompressed Colour Image Database", *Storage and Retrieval Methods and Applications for Multimedia 2004*, *Proceedings of SPIE*, Vol. 5307, pp. 472-480, 2004.
[21] M. J. Huiskes, M. S. Lew, "The MIR Flickr Retrieval Evaluation", *ACM International Conference on Multimedia Information Retrieval (MIR'08)*, Vancouver, Canada, 2008.
[22] Henning Müller, Wolfgang Müller, David McG Squire, Stéphane Marchand-Maillet, Thierry Pun, "Performance evaluation in content-based image retrieval: Overview and proposals", *Pattern Recognition Letters* 22, no. 5 (2001), pp. 593-601.
[23] "Appendix: Common Evaluation Measures", *The Twentieth Text REtrieval Conference (TREC 2011) Proceedings*, 2012.
[24] Daniel Ruby, "The Value of Google Result Positioning", Retrieved on February 13th, 2013, from: http://insights.chitika.com/2010/the-value-of-google-result-positioning/
[25] Maria Hatzigiorgaki, Athanassios N. Skodras. "Compressed domain image retrieval: a comparative study of similarity metrics", *Visual Communications and Image Processing*. 2003.

Intelligent traffic recommender system

Ljiljana Šerić, Mila Jukić, Maja Braović
University of Split, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture
Department for Modeling and Intelligent Systems
e-mail:ljiljana.seric@fesb.hr
web: http://www.fesb.hr/

***Abstract*—In this paper we present our solution of using ICT and ubiquitous computing with the goal of reducing traffic congestion during the peek hours. The solution is in presenting intelligent recommender system that is based on the combination of several machine learning techniques used for traffic density estimation, and recommending the optimal period for accessing the crossroad. We describe the design and implementation of that system, and also perform a simulation that shows its usefulness.**

I. Introduction

As the number of citizens increases and socio-economical status is changing the number of vehicles on the roads is constantly being increased. Every big city suffers from the occasional traffic congestions and with time that situation is only becoming more of an issue. This is true for a number of big cities, but we limit our research on Split, Croatia. Split is the second largest center in Croatia and the gravitational center for almost all of the industry, academy and economy that is active in its narrower and wider vicinity. It is also a popular tourist destination, so the roads that are leading to it and that go through it are often being crowded by vehicles.

The only solution for avoiding these types of congestions in the future is simple but requires financial investment. Building new roads and enhancing the ones that already exist (e.g. by adding new lanes) can improve the state of the traffic, but the government investment in road building and enhancement cannot follow the rate at which the number of vehicles is increasing on the roads. In this paper we propose an alternative solution that can be used as a temporary help before the traffic volume becomes such that new roads are necessary. Our solution consists of using modern technologies and performing better vehicle routing with the aim of minimizing congestions on the roads and optimally distributing the incoming vehicles. This is especially applicable during the so called peek periods of the day when the number of vehicles on the roads is the highest, and when congestions usually occur with higher frequency than at any other time of the year.

After reference to related work, we take a look at the current state of the particular crossroad. Next, we describe the design of the significant parts of the model – Bayesian network, REPTree decision tree and particle system. Then we present the implementation of the recommender service and the user interface. In the last section, we analyse the results obtained from the simulation of situation in which the recommender system is used in practice.

II. Related work

Traffic of big cities is a typical self-organizing system. There is no person or institution responsible for organizing the traffic. However, by subtle influence large changes can be made. In Stockholm [1] the influence of charging a small fee for bridge crossing resulted in drastic changes in vehicle routing. In [2], Barabino et al. proposed a web service for optimal mobility planing by using human and car pooling resources within a company. Our idea was to extend the idea proposed in this paper, but not limit it to one company, but rather use the entire population as our source of information. In the first step of our method, we needed to implement a traffic model that will be used for traffic prediction, and then we needed a recommender system that will enhance traffic optimality.

Probabilistic traffic modelling is not something new. In [3], Sun et al. implemented a Bayesian network for traffic flow forecasting by using the information acquired from adjacent road links. Other attempts of traffic modelling, according to [3], can be divided in simple and complex models. Simple models are based either on the current state of the traffic only (random walk model), or only on the historical average. More complex prediction model is considered to be an Urban Traffic Control System (UTCS). The time series models used in the past include ARIMA model, Kalman Filter, Neural Network, Nonparametric methods, simulation model, local regression and layered models such as ATHENA and KARIMA, fuzzy neural approach and Markov chain model [3]. There were also attempts of modelling the traffic as particle swarms [4], or using particle filters partly updated by cell phone that passes the data [5].

The model presented in this paper is used to predict the number of vehicles per day and distribution throughout the day and has goal of rearranging the vehicles so the distribution contains less peek hours.

III. Analysis of the current state of traffic

The crossroad between the state road D8 and county roads Ž6253 and Ž6139, often called ”Solin-Širina”, is situated in Solin and it represents a central connecting point for all vehicles approaching Split from west (where the towns Trogir, Kaštela and Šibenik are situated), and part of the vehicles coming from the direction of north–north-east (where Klis, Dugopolje and the center of Solin are situated). In this work we have considered modelling the dynamics of the traffic flow only on one lane driving form north to Split. The map

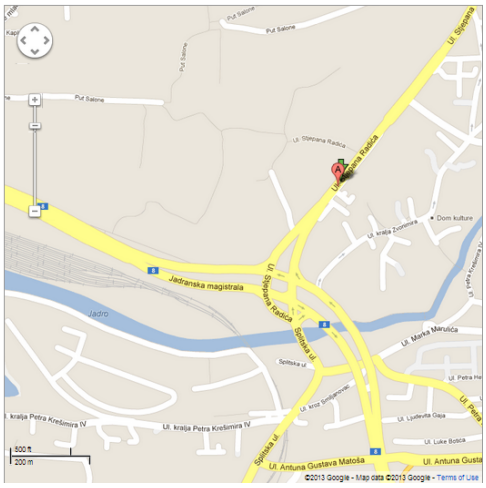


Fig. 1. Google map of crossroad with location of radar counter [6]

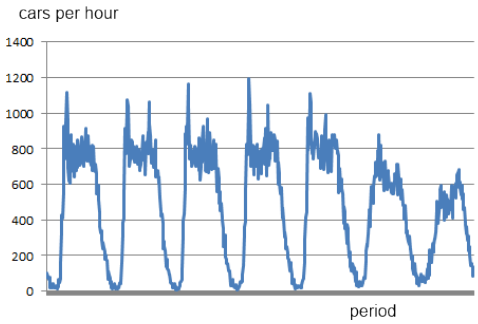


Fig. 2. Typical number of vehicles per hour for a weekin january 2011 starting moday

of crossroad is shown in figure 1. The lane often becomes congested in the morning, when people are approaching their working places in Split. Congestions are also more frequent on Mondays when people are returning to Split from their weekends, during the afternoons when people working in the area are returning to Split from work and people working afternoon shifts going to Split, and additionally during some social events, when people are visiting Split. Figure 2 graphically represent a typical number of cars per hour for 15 minutes period for a week.

In the past, on some occasions various organizations were arranging different working hours for their employees to avoid congestions on the roads. But, since people now are frequently flexible with their working hours, it is hard to establish arrangement between them in order to avoid road congestions. As the result we have situations of crossroad congestions on hours when majority of population is crossing the crossroad, while a slight shift to a prior or later period would improve the permeability of the crossroad.

Our idea is to establish this arrangement by using central recommender system as a web application where individual users would come for advice. The probabilistic traffic model will be established and used for traffic prediction. Acceptance of the advice will be treated as input data for model update.

Fig. 3. Instalation of radar [6]

IV. Model Design

The state of the traffic can be expressed by using various variables, some of them based on the number of cars present on specific places in a period of time, on the speed of vehicles, distances between them or on a number of vehicles in the queue [7].

Data received by radar instrument that we used for training purposes contains information about each vehicle crossing the crossroad, and that information consists of time, velocity and length of the vehicle. Radar instrument installation is shown in figure 3. We used the radar readings to express the data in the format on number of cars per hour and average speed for 15 minutes period. This period includes 10 traffic light cycles, since one traffic light cycle is 90 seconds long.

The number of cars per hour that crosses the particular crossroad is the variable that we used as a model output. This value depends on several variables and there are day, week and year trends. Also, the weather conditions are causing time of peek period to change.

Since the dependencies of the variables are not strict and linear, we decided to build the model using probabilistic reasoning based on the Bayesian network. Additionally, authors have chosen Bayesian network reasoning in the model of traffic flow because it has shown to be rather efficient in modelling meteorological aspect of environment with the goal of sensor data validation described in [8].

Let us first explain the design of a Bayesian network used in this paper. First we have identified two aspects of variables involved in a traffic model:

- chronological aspect and
- meteorological aspect.

Chronological aspect includes variables such as time of the year, time of the week and period of the day. The time of the year variable can be divided in twelve months, or four seasons. We decided to use the former, thus dividing a year into four seasons and knowing that seasons influence the number and distribution of vehicles on the crossroad more than the months do. Time of the week could be used simply as working day or weekend day, but we decided to use seven values for time of the week knowing that different parts of the week

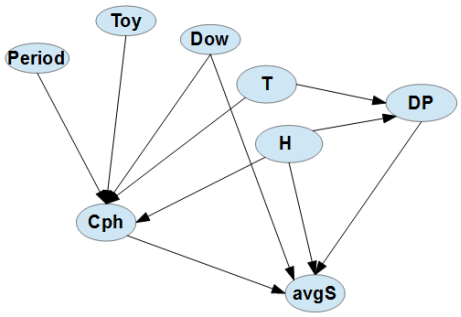


Fig. 4. The design of the Bayesian network

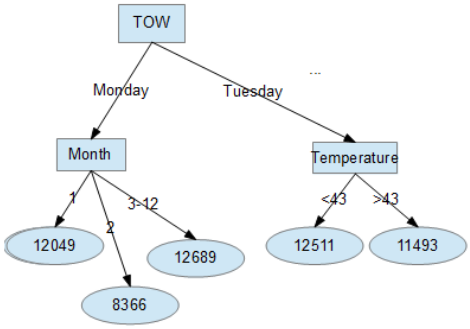


Fig. 5. A smaller part of a decision tree used for determining a number of vehicles per day

influence differently the situations on the crossroads. When considering the effect of meteorological aspect, temperature, relative humidity and dew point are taken into account. Lower temperatures cause shift of peek hours later in the morning. The same is with high humidity, and rainy conditions.

Identified variables used in Bayesian network are:

- Period - period of the day, the day is divided in 96 periods of length 15 minutes, giving total of 24 hours.
- Toy - time of year or season.
- Dow - Day of week.
- T - average temperature of the day.
- H - relative humidity.
- DP - dew point temperature .
- Cph - number of cars passing through crossroad per hour, obrained by counting the number o cars passing during the 15 minutes period ad multiplied by 4 for an hour.
- avgS - average speed of the cars.

Final design of the Bayesian network is shown in figure 4.

The output of the Bayesian network model is a number of cars per hour. The network is trained over a set of data collected from the radar traffic controller over a period of year 2011. The radar data is converted to a set of samples, each describing the state of the crossroad in one period length of 15 minutes. The format of a sample is described in 1 as:

$$D = \{Toy, Dow, Period, T, H, Dp, Cph, AvgS\} \quad (1)$$

Bayesian network conditional probability tables are trained using a set of samples and a simple learning algorithm [9]. Given evidence variables such as the day of the week, the time of the year and predicted meteorological aspect variables, we can perform probabilistic reasoning for determining most probable number of vehicles per hour for each period over a day.

V. Particle System for Probabilistic Reasoning

The core of the algorithm used in this work is a particle system for traffic modelling. The traffic is modelled as a set of particles, each representing a vehicle. The production and distribution of particles is done for each date the model is used on. The more simple way to decide on a number of particles is to take the average daily number of vehicles for fixed number of particles for each day. More accurate approach would be to use average annual daily traffic (AADT) and average summer daily traffic (ASDT) that are included in [10]. Both of these methods would not accurately represent the dynamics of traffic, since the number of vehicles can vary significantly form day to day. Our solution to this problem is to determine the number of vehicles for each day by using the decision tree obtained by REPTree fast learning algorithm often used in Internet traffic classification [11]. The algorithm is trained over a set of samples obtained by combining the radar and meteo data archive. The samples are used as an input for the RepTree algorithm. The resulting decision tree is shown in figure 5.

After determining the number of vehicles, each vehicle is individually assigned to one of the time periods of the day, according to the probability distribution calculated from known variables - the time of the year, the time of the week, and the predicted temperature forecast. Known variables are represented as the input set ζ.

$$\zeta = \{toy, tow, t, h, dp\} \quad (2)$$

The calculation of a probability distribution is flexible because it is not always necessary to know the values of all of the variables in ζ. If some variables are unknown, the summation over all of the values of the variables is then performed.

First, each period is assigned expected number of cars (NOC) by choosing the most probable value according to :

$$NOC(period) = argmax(P(numcars|period, \zeta)) \quad (3)$$

Then, the probability of a vehicle appearing in the specific time period is calculated for each period according to the equation:

$$P(period|\zeta) = \frac{NOC(period)}{\sum_{period=1}^{96} NOC(period)} \quad (4)$$

The result is car appearance probability distribution throughout the day for each period P(period), period={1,2,....96}.

Finally, by knowing the number of vehicles that cross the crossroad in a day, and by knowing the distribution of the vehicles through time periods, we can create the population of particles used for representing the vehicles. Then we can distribute them through time periods in a random manner by following the probability distribution. This is done by calculating the marginal probability values (MPV) according to equation (5).

$$MPV[period] = \sum_{p=1}^{period} P(p) \tag{5}$$

Each particle is assigned a random number (RN) from the set [0,1]. The particle is assigned to a period which satisfies the following equation:

$$MPV[period-1] < RN <= MPV[period] \tag{6}$$

Generated set of particles is now considered as a model of traffic that can be influenced to optimally rearrange the traffic to avoid congestions.

VI. Cloud Application Implementation

The recommender system is developed and deployed to Google App Engine. The deployed application is easy to scale and maintain so there is no reason to be concerned about the number of users and the volume of storage. The user interface of the deployed application has dimensions particularly suited for displaying information on the mobile devices browsers, and requires only the date and eligible time period as an input from the users. The interface for the user input is shown in figure 7. If weather forecast is available via any popular weather forecast service for a specified date, then the forecast is used as a meteorological input parameter. Otherwise, meteorological aspect of model is ignored. When a date is selected for the first time, the number of vehicles for the day is calculated using the decision tree. The Bayesian network data is used for the calculation of the marginal values for each time period according to equation (5). Then, the particles are generated. Each particle is individually assigned to one of the 96 periods of the day by generating random numbers and equation (6). Generating random number adds noise to the probability distribution generated by the Bayesian network, and thus is making the model more dynamic. The generated particle distribution is then stored permanently in a server DataStore. The period with the minimum number of particles from the selected range is recommended to the user. After the acceptance of the period, one randomly selected particle from the selected range is reassigned to the recommended period. In this way the model is updated each time it makes a recommendation.

VII. Simulation Results

Simulation is performed to prove the efficiency of the system. According to the decision tree, on September 14th 2012., 12688 vehicles are expected to cross the crossroad. The peak hours are predicted between 6:00 and 10:00 o'clock in the morning, and in that period 3723 vehicles would cross the crossroad. The simulation is done by performing the recommendations repeatedly. The resulting vehicle distribution after 1, 100, 400, 500 and 600 recomendations is shown in figure 8.

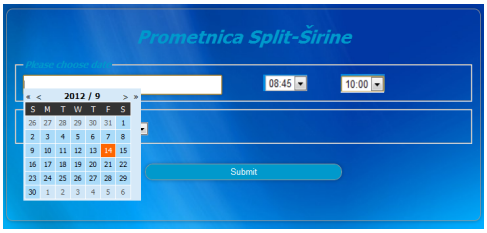


Fig. 6. Input interface

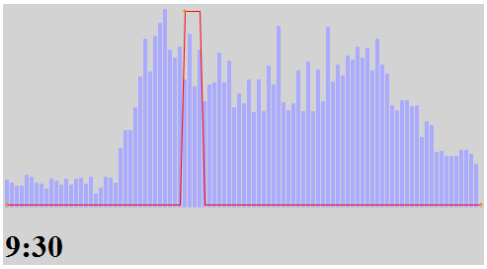


Fig. 7. Recommendation interface

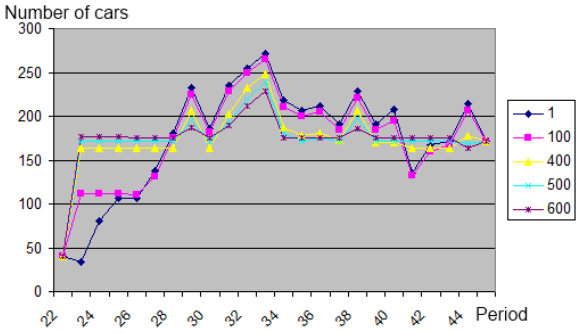


Fig. 8. Simulation results for the peek period after 1, 100, 400, 500 and 600 recommendations

It is visible that after 600 recommendations the distribution of vehicles between 6:00 and 10:00 is more planar, with peaks lowered. It can be concluded that if only 5% of the total population that is using the specific crossroad would use the recommender system the situation in peek periods would significantly improve and the traffic congestions would be less likely to occur.

VIII. Conclusion

In this paper it is shown how the Internet and the technology can be used to solve everyday problems like traffic congestions. By combining several machine learning technologies and implementing a particle system to mimic the traffic, we can model the state of the traffic and recommend optimal period for crossing the specific crossroad. Recommendations can be used as update information so the model is not static.

The strength of this system relays on its users. The more users are using it in the right way, the better it will reflect the real state of the traffic. In this model, only the most obvious aspects (eg. weather, the day of the week) are taken into account. By identifying and introducing other less obvious variables the model could be more accurate. By introducing real time data from traffic counters in the form of number of cars that already crossed the crossroad for the model update the system could be significantly improved, but this requires infrastructure investments.

ACKNOWLEDGMENT

The authors would like to thank the people from "Hrvatske ceste" and "Županijska uprava za ceste", especially Željko Sikirica and Ivan Danolic for sharing the radar traffic controller data, as well as for other useful information.

REFERENCES

[1] J. Eliasson, "Lessons from the stockholm congestion charging trial," *Transport Policy*, vol. 15, no. 6, pp. 395 – 404, 2008, ¡xocs:full-name¿Decision-support for sustainable urban transport strategies¡/xocs:full-name¿. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967070X0800053X

[2] B. Barabino, S. Salis, and A. Assorgia, "Application of mobility management: a web structure for the optimisation of the mobility of working staff of big companies," *Intelligent Transport Systems, IET*, vol. 6, no. 1, pp. 87–95, March 2012.

[3] S. Sun, C. Zhang, and G. Yu, "A bayesian network approach to traffic flow forecasting," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 7, no. 1, pp. 124–132, March 2006.

[4] W. Peng, W. Jiang-Ping, and X. Jing, "The application of particle swarm optimization on intelligent transport system," in *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on*, vol. 4, August 2009, pp. 389–391.

[5] L. Mihaylova and R. Boel, "A particle filter for freeway traffic estimation," in *Decision and Control, 2004. CDC. 43rd IEEE Conference on*, vol. 2, December 2004, pp. 2106–2111.

[6] "Google maps, solin, 43.537707,16.487432," accessed on 2.2.2013.

[7] E. Curtis, "Traffic control systems handbook," http://ops.fhwa.dot.gov/publications/fhwahop06006/index.htm, accessed on 21.09.2012.

[8] L. Šerić, M. Štula, D. Stipaničev, and M. Braović, "Bayesian proprioceptor for forest fire observer network," in *Proceedings of the 6th KES international conference on Agent and Multi-Agent Systems: technologies and applications*, ser. KES-AMSTA'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 14–23.

[9] R. Bouckaert, "Bayesian network classifiers in weka," Department of Computer Science, Waikato University, Hamilton, NZ, Tech. Rep., 2005.

[10] M. Božić, D. Kopić, and F. Mihoci, "TRAFFIC COUNTING ON THE ROADWAYS OF CROATIA IN 2011 - digest," May 2012, http://www.hrvatske-ceste.hr/, accessed on 21.09.2012.

[11] J. Park, H.-R. Tyan, and C.-C. Kuo, "Internet traffic classification for scalable QOS provision," in *Multimedia and Expo, 2006 IEEE International Conference on*, July 2006, pp. 1221–1224.

Cloning Fuzzy Inference System Component Behavior with an Artificial Neural Network Component

Dušanka Bošković, Zikrija Avdagić
Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina
dusanka.boskovic@etf.unsa.ba, zikrija.avdagic@etf.unsa.ba

Abstract - This paper presents results in cloning Fuzzy Inference System (FIS) component behavior using an Artificial Neural Network (ANN) component. Both FIS and ANN components are part of the intelligent framework VIMAS developed in Java programming language. VIMAS objective is to enable integration of the computational methods in a holistic manner. The FIS component generates training set for the ANN component containing a sequence of representative input combinations accompanied with the respective output. The FIS component sends necessary configuration data and training set data to the ANN component. After the training, the ANN component is presented with input data, and results are compared.

I. INTRODUCTION

Artificial Intelligence (AI) methods proved successful in solving complex problems, especially when integrating these methods to complement each other deficiencies. The results presented in the paper are part of the work aimed to bring together different AI methods and provide a comprehensive framework for their co-operation and to facilitate their interaction. The VIMAS framework is developed using agent technology and by utilizing the best principles of software engineering as object orientation and component approach. The VIMAS framework provides only basic infrastructure services to the agents. An VIMAS agent primary role is to provide means of communication and hence co-operation between different AI computational components as ANN and FIS.

There are many examples of hybrid systems based on utilizing the co-operation between the fuzzy systems and the neural networks, cited in the literature [1]. In addition to neuro-fuzzy hybridization there are other possibilities for neuro-fuzzy synergy. In [2], the authors describe how to use the information in a fuzzy rule system for the adaptation of a neural network. The fuzzy knowledge is exploited to generate hints, additional data used to speed up the learning process of the ANN.

The main objective of this paper is to present results related to an implementation of the Neural Network as a clone to a Fuzzy Inference System (FIS) solution. This approach helps to transfer knowledge about the system modeled in fuzzy logic into a suitable ANN. It is expected that the operation of the FIS component can be reproduced by an ANN. The motivation for the cloning is to combine the advantages of the ANN as a capability to learn, and the advantages of the FIS, capability to form a linguistic rule base. This feature of the FIS is important when the system knowledge is vague and having more linguistic then numerical nature, and when it is not easy to describe a system in terms of input and output values for the ANN training.

The approach described in the paper enables transfer of expert domain knowledge integrated in the FIS knowledge and rule base into the numeric form embedded in the ANN connection weights. The linguistic FIS model is used to produce input-output data pairs for the neural network training. The ANN ability to generalize from training data shall compensate for the incomplete FIS model. For the purpose of reproducing the FIS model with the ANN component, the FIS model can focus on the most significant and representative features of the problem, instead of producing rules for all input combinations.

Sections II and III provide description of the FIS and ANN component, respectively. The FIS and ANN components are implemented in the Java language and their design is illustrated by representative UML diagrams. The section V describes method used for knowledge transfer from the FIS component into the form of a training set data for the ANN component. To validate implementation of the proposed ANN cloning, the experimental results of a case study are presented in the section V. The last section presents conclusions and gives comments for future development.

II. FIS COMPONENT

The FIS component implementation is based on work described in [3]. The core of the Fuzzy Inference System implementation is a fuzzy rule base with fuzzy rules and fuzzy variables. The FIS component includes different implementations for methods of fuzzification, inferencing and defuzzification. The Fuzzy Inference System class diagram is presented in Fig. 1.

`The FuzzyRuleBase` class contains collection of the fuzzy rules, implemented as an instance of a `FuzzyRule` class. The `FuzzyRule` class contains a list of antecedent fuzzy clauses and a single consequent fuzzy clause. The `FuzzyClause` class is composed of a linked pair: a `FuzzyVariable` and its fuzzy value represented with a `FuzzySet` object.

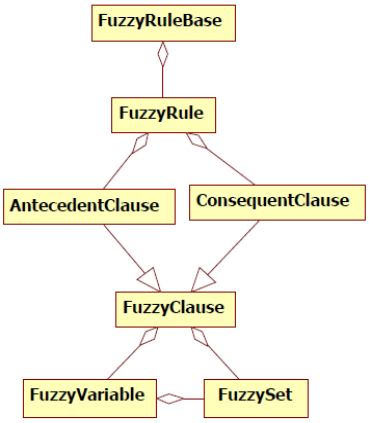


Figure 1. Fuzzy system class diagram

The FIS input and output fuzzy variables are all implemented using the `FuzzyVariable` class. `FuzzyVariable` is defined with a name, a range of possible numeric values and a set of possible linguistic values, implemented by a `FuzzySet` class. The `FuzzySet` class is an abstract class, and a base class for specific classes containing all relevant data to define characteristic membership function. The hierarchy of implemented fuzzy set classes is presented in the class diagram in Fig. 2.

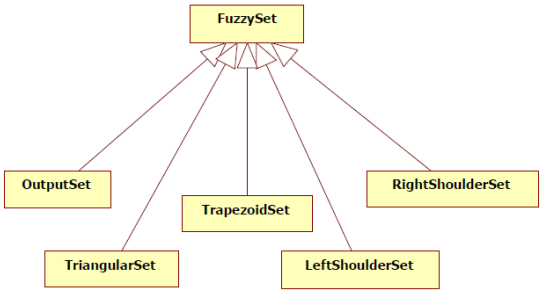


Figure 2. Fuzzy set classes hierarchy

After the fuzzification of the input crisp values, obtained fuzzy values are combined with the rule base to conduct the fuzzy reasoning process. The reasoning process produces a new fuzzy set – `OutputSet`, describing output variable, and after the defuzzification the output is converted into a crisp value.

III. ANN Component

The implementation of a neural network is defined with the network architecture and a learning algorithm. The ANN component implements a feed forward backpropagation neural network [4] using a single class - `NeuralNetwork`. The network architecture is defined with the number of neurons in the input layer, the hidden layer and the output layer. The number of input and output neurons relates to the number of input and output variables. Variables are implemented in a separate class Variable.

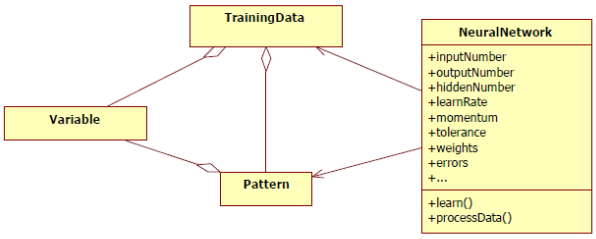


Figure 3. Neural network class diagram

The ANN component is equipped with the methods to read configuration data from a specified source, and afterwards to read training data. The class `TrainingSet` is responsible to hold the training data set and to manage the loop through the training process. The class Pattern is responsible to present normalized input data to the `NeuralNetwork`. The ANN class diagram is presented in Fig. 3.

The learning algorithm is configured with the following parameters: learning rate, tolerance, and momentum. During training, the `NeuralNetwork` state is defined with the following arrays of data: connection weights, errors and thresholds. The trained network is saved and can be easily evoked when needed to process the data.

In the processing mode the `NeuralNetwork` is presented with the input data. During the processing for network validation or pruning, all relevant data, such as produced output value, desired output value, and errors are captured in the `TrainingSet` object. The validation data, as produced-desired value pairs, are written in the text file for further processing and analysis.

IV. Training Set Data Generation

The generated training data set includes the representative features of domain knowledge modeled in the FIS component. The number of different training patterns depends on the model complexity, specifically the number of input variables and the number of membership functions (MF) describing them. The representative values are selected as having the highest membership degree for each membership function.

The algorithm for generating training data set is:

1. Create a list of all fuzzy input variables, forming an ordered sequence according to a designated unique numeric identifier;
2. \Define an array `numberOfMFs`, where each element of the array corresponds to a specific input variable and contains a number of MF describing the variable
3. Define a matrix `testValues` with dimension: the number of input variables by the number of MFs for each variable;
4. Populate the matrix `testValues` with the representative crisp values, calculated as having the maximum value of the corresponding MF;

5. Generate all possible combinations of the input variables based on the values from the matrix `testValues` – and generate a new matrix `testInputs`;
6. Loop through the matrix `testInputs` and run the FIS component to evaluate the input values sequentially; Save input data with defuzzified output in the matrix `trainingSet`;
7. Final step is to delete the rows in the matrix `trainingSet` for the input values that did not fire any rule.

The training set data is written in a text file, and description of the input and output variables are described in the separate configuration file. These files are needed to configure network topology. The FIS component sends a signal message to the ANN component, indicating that configuration data and training data set are available.

V. Experimental Results

This section presents experimental results of the transfer of knowledge from a FIS model to the ANN component used to estimate the quality of the software method. The estimation is based on five software quality metrics and the result is used to decide if code refactoring is needed. The metrics used for code evaluation are: *duplicated code*, *long method*, *long parameter list*, *temporary field* and *cyclomatic complexity*.

These code attributes are considered as code smells, and are considered as indicating design faults. The list of common bad smells can be found in [5]. The output of the FIS model is the estimation of the quality of the software method in relation to a certain measure of code smells.

TABLE I. Bad Smells Fuzzy Model

Fuzzy Variable	Fuzzy Value	MF Type with Characteristic Points
longMethod	small	leftShoulder [15, 40]
	medium	trapezoid [0, 30, 70, 100]
	large	rightShoulder [60, 85]
parameterList	veryShort	leftShoulder [3, 5]
	short	triangle [3, 7, 11]
	medium	trapezoid [7, 9, 13, 15]
	long	triangle [13, 16, 19]
	veryLong	rightShoulder [18, 22]
cyclomatic Complexity	simple	leftShoulder [3, 10]
	complex	triangle [7, 13, 20]
	veryComplex	triangle [16, 33, 50]
	untestable	rightShoulder [40, 60]
duplicatedCode	verySmall	leftShoulder [20, 30]
	small	triangle [0, 25, 50]
	medium	trapezoid [25,40, 60, 75]
	large	triangle [50, 75, 100]
	veryLarge	rightShoulder [75, 100]
temporaryField	verySmall	leftShoulder [5, 10]
	small	triangle [5, 22, 40]
	medium	trapezoid [20, 30, 70, 80]
	large	rightShoulder [60, 80]
codeQuality	excellent	leftShoulder [10, 20]
	very good	triangle [10, 25, 40]
	good	triangle [30, 45, 60]
	bad	triangle [50, 65, 80]
	veryBad	rightShoulder [70, 80]

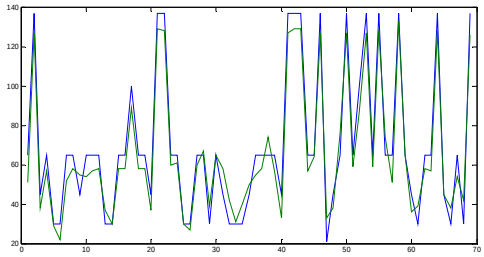


Figure 4. Validation of the ANN results (green line) against the FIS results (blue line)

The fuzzy model developed is named BadSmells and it is a system with five input fuzzy variables, based on the previously listed software metrics, and one fuzzy output: code quality. The domain knowledge of the FIS is presented in the Table I. The domain expert knowledge is based on the empirical values for bad smells described in [6]. This was a basis to define fuzzy variables and fuzzy rules.

The FIS component is used to generate training data set for the ANN component as described in the section V.

The network topology used for training is 5-5-1, five neurons in the input layer corresponding to the number of input variables and one neuron in the output layer matching the one output variable: *code quality*. The parameters of the network that can be defined during the runtime are: order of the training data, sorted or shuffled; and activation functions in the hidden and the output layer, sigmoid or tanH.

After training, the ANN component performance is validated. The quality of a software method is modelled as discrete value and the mean-square error is used for the quantitative evaluation of the ANN component. The results of the evaluation are presented in the Table II. The results produced by the ANN component are compared against the results of the FIS component for the same input values. The outputs of the FIS and the ANN component are presented in the Fig. 4. where the ANN type with sigmoid activation function and shuffled data is used.

The quality estimation based on the metrics data combined with a set of heuristic rules substitutes subjective decisions made by software developers. The usage of the trained ANN component to replace the FIS component facilitates domain modelling process. The FIS component prior to cloning can be implemented with only a small set of the most relevant fuzzy rules, instead using the complete set of 1200 rules.

The presented graph indicates that the ANN component can be used as a successful FIS clone for the purpose of the code quality evaluation.

TABLE II. Average MS errors computed over 4 ANN types

NN Parameters	Training	Validation
Sigmoid, sorted data	0.0651	0.0656
Sigmoid, shuffle data	0.0546	0.0557
TanH, sorted data	0.0868	0.1602
TanH, shuffle data	0.1175	0.0981

VI. CONCLUSION

The results presented indicate that the Artificial Neural Network component can be trained with the data generated from the model designed for the Fuzzy Inference System, and after the training is used in a successful manner to perform the same task as the FIS component.

This is a simple way to transfer more qualitative and linguistic domain knowledge into the numeric form suitable for the ANN training. The approach presented in the paper can facilitate automatic selection of the software methods that need refactoring.

Implementation of the knowledge transfer between the ANN and FIS component provides a solid framework for the future work addressing also the other ways of ANN and FIS integration and co-operation.

REFERENCES

[1] J. Fulcher, and L.C. Jain, Eds. *Computational Intelligence: A Compendium*, ser. Studies in Computational Intelligence. Berlin, Germany: Springer, 2008, Vol. 115

[2] Y. Jin, and B. Sendho, "Knowledge Incorporation into Neural Networks From Fuzzy Rules," *Neural Processing Letters*, vol. 10, no. 3, pp. 231-242, 1999.

[3] Z. Avdagic, D. Boskovic, and A. Delic, "Code evaluation using fuzzy logic," in *Proc. of the 9th WSEAS International Conference on Fuzzy Systems FS'08*, 2008, pp 20-25

[4] S. Russel, and P. Norvig, *Artificial Intelligence: A Modern Approach*. 2nd ed., New Jersey, USA, Pearson Education, 2003.

[5] M. Fowler, *Refactoring, Improving the Design of Existing Code*, Boston, MA, USA: Addison-Wesley, 1999.

[6] M. Mäntylä, C. Lassenius, "Subjective evaluation of software evolvability using code smells: An empirical study," *Journal of Empirical Software Engineering*, vol. 11, no. 3, pp 395-431, 2006.

Principal Component Analysis of Fatigue Strength

Lj. Šerić*, K. Korun Curić ** and M. Šušnjar***
University of Split, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Split, Croatia
∗ Department for Modelling and Intelligent Systems
∗∗ Department of Structural Mechanics and Design
∗ ∗ ∗ Bit-Art Informatika, Split, Croatia
email:ljiljana.seric@fesb.hr

***Abstract*—In this paper authors examine the possibility of using modeling technique based on Principal Component Analysis (PCA) on fatigue strength modeling of selected specimens of steel. The PCA technique will be used for data analysis and model development with goal of identification of variables with most influence on fatigue strength. The model is trained upon data available in literature and application of described model is tested on series of laboratory experiments.**

I. Introduction

We are living in times when data is easily available but structure of those data is not explicitly available. Structuring large amount of data by hand is expensive and time consuming. Instead of structuring the data by hand, we are preferably developing techniques and algorithms for automatic data processing and eventually data understanding. A great number of machine learning techniques is developed and introduced as a part of well-known data mining and data processing techniques.

Machine learning techniques are often used on data describing phenomenon that we are not completely aware of its inner structure. Fatigue strength of material is not completely theoretically described and is mostly described in the form of experiment data. Our hypothesis is that it is possible to find structure in such data by using some of machine learning techniques.

In this paper we will first present related work in applying machine learning techniques in fatigue strength modelling as well as usefulness of machine learning techniques for material characteristics modelling. In the following chapter we will describe the data set used as input data for principal component analysis and training data for modelling. Next, we will show results of principal component analysis of the data and discuss the outcome of analysis. Finally modelling technique used in this paper will be described and results will be shown.

II. Related work

Significant number of scientific researches deals with methods aiming to predict lifetime of structures. The authors mostly used neural network as a method or predicting the lifetime of various materials. . Past researches could be divided into basic groups of lifecycle prediction through simulation of S-N curve [9],[10],[13],[14],[6],[15] and the groups of material behaviour prediction through the propagation of cracks and fracture mechanics [12]. The authors have worked with a variety of materials but in the present researches composite materials and aluminium alloys are most interesting being the dominant materials in modern construction. Several other authors have successfully applied the neural networks in other areas of lifetime predictions: in dynamic fracture mechanics [8], in temperature changes of nickel alloys [17], in modelling the propagation of cracks [11], etc. Previous studies focused on the modelling of problems with varying relatively small number of parameters in a very narrow field of application. Their characteristic is focus on the newer materials such as composites and lightweight materials like aluminium. It is observed that researchers are focusing on one type of material for individual study. Method selection for fatigue strength modelling is based on a characteristic of input data and the way they are prepared. Statistical analysis of input and predictive variables provides guidance on the selection method. If the strongest predictive variables have a good distribution of the target variables most methods will be able to give fairly good results. Modelling of fatigue strength problems required quality data preparation because data in its original form cannot provide good predictive models. The reason is scatter of experimental results, which is usual problem of fatigue strength. In modelling of Wöhler curve or other already statistically processed data is used. Such modelling solves the problem of noise on which predictive methods are sensitive.

Predictive methods for modelling fatigue strength that can be found in literature are following [16]:

- Neural networks
- Classification and regression trees (C&RT) - originally presented the 1984. Based on the partitioning of the data set in the two subset in such a way that the data subsets in more homogeneous than before partitioning. The process is recursive, then each of these divides again and the process continues until it reaches homogeneity or some other criterion branching [3].
- QUEST method - represents a statistical binary tree. The principle of operation is such that observed branching points in several ways and chooses the best. Provides fewer features than C&RT method and seeks discretized target variable [18].
- CHAID method - Chi-squared Automatic Interaction Detector - an exploratory method used to study the relationship between a dependent variable and a series of predictor variables. CHAID modeling selects a set of predictors and their interactions that optimally predict the dependent measure. The developed model is a classification tree that shows how major ”types”

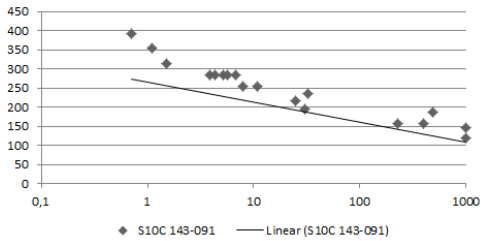


Fig. 1. S-N curve for S10C series 143-091

formed from the independent variables differentially predict a criterion or dependent variable [7].

- Regression methods
- SVM method (Support Vector Machine) - method for mapping input data on the output is used for classification and regression functions [4].
- Clustering methods - Nearest Neighbour Analysis

III. Fatigue strength

Fatigue is the progressive and localized structural damage that occurs when a material is subjected to repeated loading and unloading. If the loads are above a certain threshold, microscopic cracks will begin to form at the stress concentrators such as the surface, persistent slip bands, and grain interfaces. Eventually a crack will reach a critical size, and the structure will suddenly fracture. The shape of the structure will significantly affect the fatigue life; square holes or sharp corners will lead to elevated local stresses where fatigue cracks can initiate.

Fatigue strength is expression used to describe a property of materials, the amplitude of cyclic stress (S) that can be applied to the material without causing fatigue failure. It is represented by fatigue life, N_f, which is the number of stress cycles which a material can withstand before breaking. Fatigue strength, when viewed as a branch of science, is a sophisticated research and theoretical considerations geared to finding optimal dimension of each and individual part of the structure. Under dimensioning of any part leads to failure and endangering of its allocation. Structural failure often has catastrophic consequences. Over dimensioning, on the other hand, raises the price of the structure and causes additional problems in terms of its dimensions and weight. Fatigue strength seeks and finds the optimal design parameters.

Fatigue strength of material can be expressed in the form of S-N curves or Wöhler curves. An example of S-N curve, constructed for material S10C series 143-091 is shown in figure 1. The curve is constructed based on experimental data by approximating linear curve through experimental data. S-N curves are derived from tests on samples of the material to be characterized where a regular sinusoidal stress is applied by a testing machine which also counts the number of cycles to failure. Each test generates a point on the plot though in some cases there is a runout where the time to failure exceeds that available for the test. Analysis of fatigue data requires techniques from statistics, especially survival analysis and linear regression.

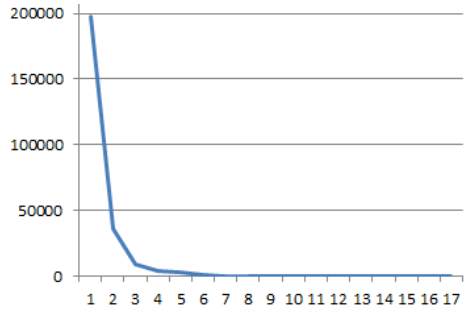


Fig. 2. All sorted eigenvalues for training data covariance matrix

Constructing S-N curve for a material requires many laboratory experiments. Each experiment results in one point in a plot and curve is approximated as linear prediction through experimental data. Predicting S-N curve of a material based on some features of the material would be of much help because it would save much time and money invested in experiments. The correctness of predicted S-N curve could be tested by smaller number of experiments.

A. Description of used data

Experiments in the fatigue strength must follow the standard procedure so that a certain pattern with a particular treatment under particular conditions from a particular material and varying load conditions give results in the form of lifetime (number of loading cycles to failure). All input attributes are set of test data that is mapped to the results. In modelling experiments by predictive methods all these attributes are potential variables. Since the parameters are heterogeneous in terms of how they are defined, they can be descriptive (mechanical processing, thermochemical processing ...) or numerical (lifetime, temperature ...). List of attributes used in training set with their description is given in the table I. Numerical attributes are normalized using mean value and standard deviation. Descriptive attributes are quantified by constructing descriptive table. For example, possible specimen surface are from set {circular groove, sandblasted sample, smooth sample, double groove, cropped sample} are translated to numerical values $\{1, 2, 3, 4, 5\}$.

Training set used in this work was obtained from catalogue [19], [20], [21]. The curves and tables were digitalized by hand.

IV. Principal components identification

Principal component analysis (PCA) [1] is generally used for dimensionality reduction, either for data visualization or data preprocessing before using some other machine learning algorithm. The method consists of presenting the set of training data in the form of matrix, calculating covariance matrix and calculation of eigenvector and eigenvalues of training set covariance matrix. Calculated eigenvectors and eigenvalues can help us in identification of features with most influence on training set. Figure 2 shows sorted values of all eigenvalues of training set. The image shows that first 8 eigenvectors have most significant influence on training set.

Figure 3 shows values of first 6 eigenvectors. Each component of a vector corresponds with one feature of the training

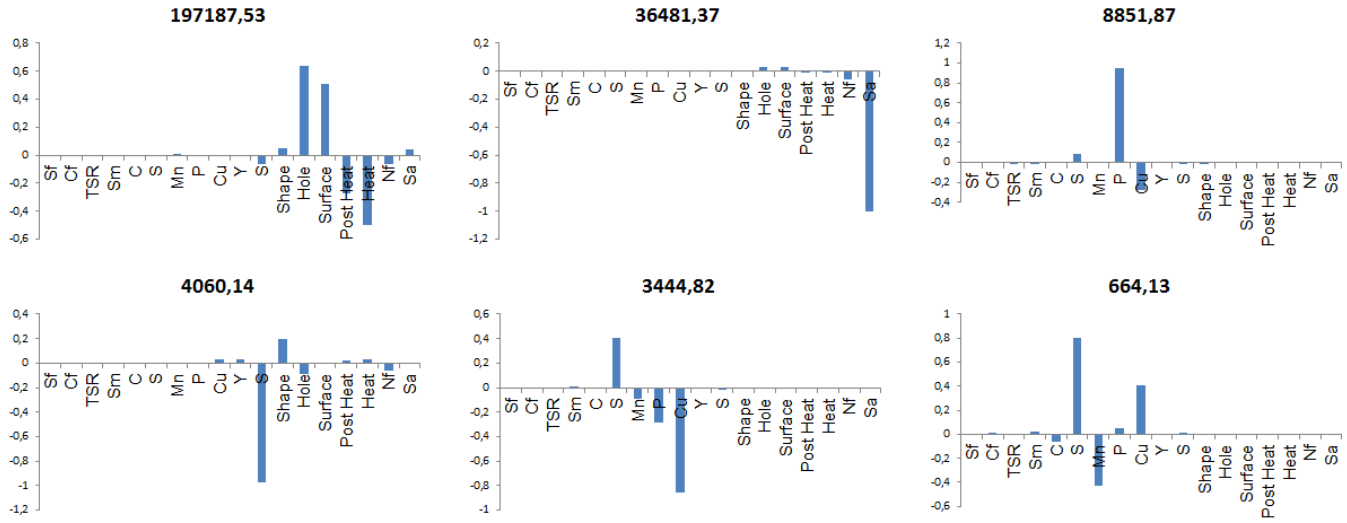


Fig. 3. Eigenvectors with highest eigenvalues

set. The features are sorted as in table I. The eigenvalues of vectors help us choose most significant vectors, while eigenvectors are used in analysis of influence of specific features on specific vectors. The Figure 3 is used in analysis of this influence. Vectors with largest eigenvalues are mostly consisted of features describing the treatment of the material. Among chemical characteristics, most influence has the proportion of phosphorus which is considered as impurities of the material.

V. PCA MODEL FOR DATA PREDICTION

Besides its usefulness for analysis of a set of data, PCA can also be used as a tool for building models for outcome prediction on a set of data about phenomenon, especially when data are presenting phenomenon where we don't have enough theoretical knowledge about. In [2] the PCA was successfully implemented to construct a model or human shape completion and animation so this method can cope with highly dimensional data. Resulting model will be linear. Although usage of non-linear models provides more flexible fitting, as explained in [5] linear models can be quite robust, especialy in the domains where theoretical knowledge is insufficient.

The training set consists of 18 features. The training set digitalized from the catalogue [19] is used as input data for PCA algorithm.

First the matrix of means μ is calculated and data is normalized to values betwean -1 and 1. Then the covariance matrix is calculated:

$$\Sigma = \frac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^T \tag{1}$$

The covariance matrix is treated with singular value decomposition and result is matrix of eigenvectors Φ . Two matrices are used as material model:

Φ - feature matrix

μ - means vector used for normalization of data

These two matrices are parameters of the resulting model and are used on new data.

A specimen data (D) is multiplied by feature matrix and thus transformed into PCA space. We will denote this transformed specimen of data as D_{pca}.

$$D_{pca} = \phi^T \times (D - \mu) \tag{2}$$

Transformed data looses some of information during transformation, which can be evident by transforming data back into original space. Let us denote the transformed data with D_{back}.

$$D_{back} = \phi \times D_{pca} + \mu \tag{3}$$

We can now calculate the difference matrix :

$$\Delta = D - D_{back} \tag{4}$$

The difference between original data and data transformed to PCA and back can be used as a measure of similarity between data and training set. The loss of data can be explained as this: features of PCA transformation are fitted to training set so that minimum information is lost by transformation to PCA space. If most of information is preserved data is more similar to training set.

The similarity of data is measured as measure of distance (MOD) as a matrix F-norm calculated as :

$$MOD = Fnorm = \sum \Delta_{ij}^2 \tag{5}$$

This gives us the measure of similarity of the data D and training set. The more similar D is to training set, less data would be lost in conversion and the MOD would be lower. Data that is not well adjusted to the training set would have higher value of the MOD.

TABLE I. DESCRIPTION OF ATTRIBUTES INVOLVED IN FATIGUE STRENGTH OF MATERIAL

Attribute	Description
Fatigue strength, Test stress conditions	
Specimen Fatigue At 1E7 Cycles MPa	Fatigue strength at 10^7 number of cycles
Test Stress Ratio	Constant stress ratio used in test, $\sigma_{min}/\sigma_{max}$
Sigma m	Constant mean stress, σ_m , algebraic mean of the maximum and minimum stress
Chemical composition of specimen material	
C	Mass percentage of carbon in material
Si	Mass percentage of silicon in material
Mn	Mass percentage of manganese in material
P	Mass percentage of phosphorus in material
S	Mass percentage of sulfur in material
Cu	Mass percentage of copper in material
Tensile test	
Tensile Yield MPa	Stress at which a material begins to deform plastically
Tensile Strenght MPa	Maximum amount of tensile stress that a material can take before failure
Specimen shape, Mechanical and Heat Treatment	
Specimen Stress Concentration Factor	Ratio of the highest stress σ_{max} to a reference stress σ, geometrical.
Specimen Shape	Specific shape of fatigue specimen
Specimen Hole Drilled	Existence of hole in a specimen
Speciman Surface	Condition of specimen surface after mechanical treatment
Post Heat Treatment	Post heat treatment, secondary
Heat Treatment	Heat treatment
Test results	
Test N_f in 10E4	Number of cycles specimen endure before failure occurs
Test Sa MPa	Stress amplitude of each test

A. Predicting stress amplitude (S)

The model can be used for predicting stress amplitude S of a specimen given N_f value of fatigue life. The specimen is described by known chemical and manufacturing features and number of cycles while S value is unknown. Knowing the extend of value in provided training set, we are searching for the value of S in range between 0 and 500. The predicted value is used which fits best to constructed model. The algorithm for choosing S is following:

For each value of S from the range 0-500
Generate data sample
Transform sample data to PCA space with matrices Φ and μ (S) value is known. The algorithm for predicting N_f value is

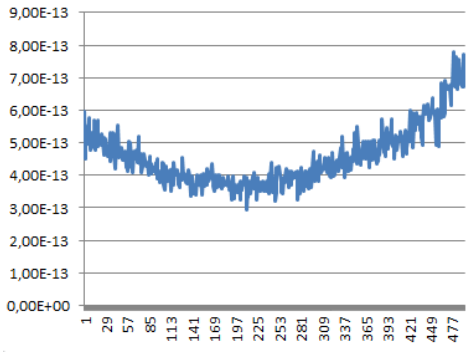


Fig. 4. Measure of distance for various values of S for specimen S10C 173-000 given N =100 $\times 10^4$

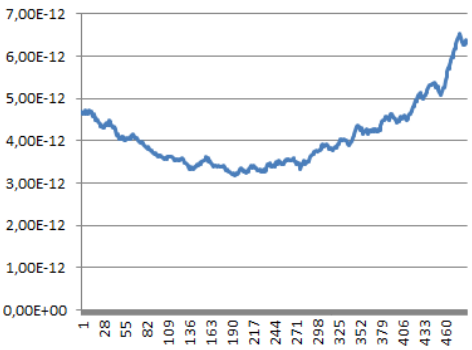


Fig. 5. Measure of distance for various values of S for specimen S10C 173-000 given N =100 $\times 10^4$

Transform data back into original space
Compare original data and transformed data
Choose S where data is most similar to original data.

Similaritiy between original and transformed data is calculated as F-measure, or measure of distance between data. Typical distribution of measure of distance for a specimen for values of S 1 to 500 is given in figure 4. The graph in Figure 4 has minimum value but also has a lot of noise. The noise is result of fitting model to experimental data. Experiments for S-N curves are performed by repeating measurements under same conditions so the results are statistically processed. Small variations in predicted S cause choosing different points in PCA space that have different measure of distance to known data. To reduce noise, instead of MOD, we are introducing density of F-measure calculated by sum of 10 consecutive measures. This is possible because we assume that values numerically close to the expected best fitting value also have low MOD. The resulting graph is shown in Figure 5.

Finally, the value of S with smaller density of measure of distance is used as predicted value. The noise found in measurements can also be considered as strength of the model, since laboratory measurements always exhibit similar noise in measurements.

B. Predicting fatigue life (N_f)

Predicting N_f is used for specimens where stress amplitude

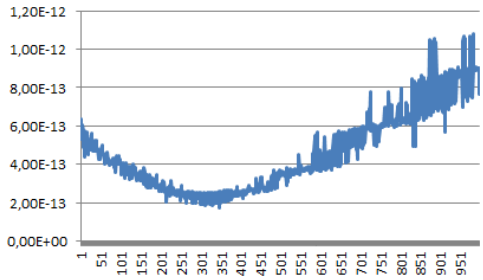


Fig. 6. Measure of distance for values of N_f between 1 $\times 10^4$ and 1000$\times 10^4$ for specimen S10C 173-000 given S =260.68 MPa

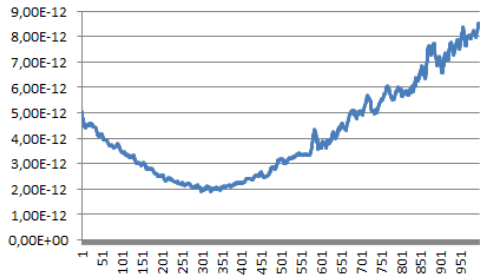


Fig. 7. Density of measure of distance values of N_f between 1 $\times 10^4$ and 1000 $\times 10^4$ for specimen S10C 173-000 given S =260.68 MPa

identical to algorithm predicting S, only different range is used. The value of N_f is expected in the range between 1 $\times 10^4$ and 1000$\times 10^4$. The typical distribution of distance measure for values of N_f is shown in figure 6.

Since this graph also shows great influence of noise, we are using density of distance measure calculated by 10 consecutive measures and resulting graph is shown in figure 7.

The same procedure for selecting the predicted value is used as in predicting S value, the value having minimal density distance measure is used as predicted value of N_f.

C. Predicting S-N curve

Predicting Wöhler curve of a material given its chemical and manufacturing characteristic can lead us to theoretical knowledge about inner structure of fatigue strength of material. Predicting how slight variations in manufacturing can lead to variations in specimen fatigue characteristics can be used for manufacturing purposes for finding optimal combination of characteristics for desired purpose. Predicting Wöhler curve is done by predicting S for values of N_f from the set $\{20, 40, 60, 80, 100, 120, 150, 200, 400, 600, 800, 1000\}$ all expressed in 10^4 cycles. Prediction for each value of N_f is done individually, unrelated to previous data. Resulting points are treated the same as experiment results for creating S-N curve. The curve is approximated as linear curve obtained by linear regression performed on predicted points. The results are presented on a graph with logarithmic x axes.

VI. Results

Series of experiments with various specimens were performed before to validate the modelling technique described in [16] . Here we present comparison of experimentally measured

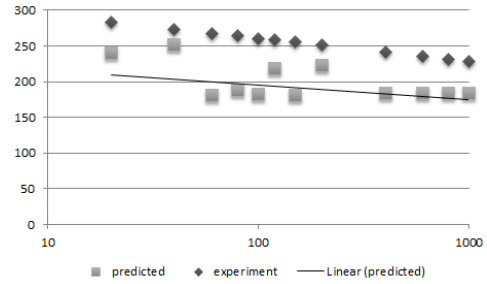


Fig. 8. Results of prediction and experiment for S for S10C 173-000

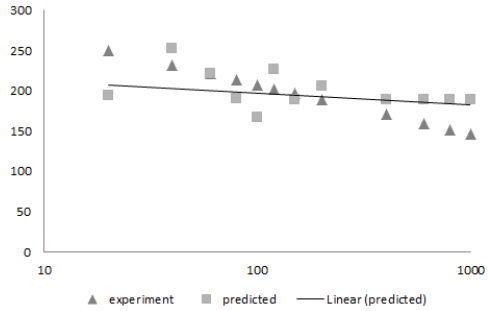


Fig. 9. Predicted and experimentally measured S-N curve for S15C 292-002

data and predicted data for specimens S10C 173-000 and S15C 292-002. Figures 8 and 9 show comparison of experimentally measured S and S predicted by method for variations of N_f from the set. Data is presented on a plot area with logarithmic x axes, as usual for presenting S-N curve. Predicted S-N curve is obtained by approximating linear curve through data points. The difference between experimental results and predicted curve exists in both cases, but this can not be a reason to reject this modelling technique because repeated experiments would also result in different values. Both predicted S-N curves follow the expected trend of S-N curve, predicting lower S values for higher N_f.

VII. Conclusion

Many techniques used in machine learning can be used for describing influence of various variables on outcomes that are not completely theoretically described. PCA can be used for identification of variables with most influence on fatigue strength of material. Using training data set with even higher dimension would be desirable because this method would successfully reject unimportant dimensions or features. Predicting fatigue strength descriptors, namely the S-N curve, of the material is ambitious task but with sufficient amount of data can be done. The correctness of predicted S-N curve is not adequate to replace experimental data but can be surely used as aid in some situations when certain amount of error is allowed. Predicting S-N curves for a specific kind of material, based only on smaller variation of features, but with sufficient data set would give better results and our future research would be in that direction.

References

[1] H. Abdi , L. J. Williams , Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics, Volume 2, Issue 4, pages 433-459, July/August 2010.

[2] D. Anguelov, P.Srinivasan, D.Koller, S.Thrun, J. Rodgers, J.Davis. SCAPE: Shape Completion and Animation of People. Proceedings of the SIGGRAPH Conference, 2005.

[3] L.Breiman, J.H.Friedman, R.A.Olshen, C.J.Stone, Classification and Regression Trees, Chapman Hall/CRC, 1984

[4] C.Cortes, V.N.Vapnik, Support-Vector Networks, Machine Learning, 20, 1995

[5] R. M. Dawes,The robust beauty of improper linear models in decision making,American Psychologist, pages 571–582, 1979.

[6] R.C.S.Freire, A.D.D. Neto, E.M.Freire, Building of constants life diagrams of fatigue using artificial neural networks, International Journal of Fatigue, 27 (2005) 746-751

[7] G.V.Kass, An Exploratory Technique for Investigating Large Quantities of Categorical Data, Applied Statistics, Vol. 29, No. 2 (1980), p.119127

[8] Cheng-Jung Li, Asok Ray, Neural network representation of fatigue damage dynamics, Smart Material Structure, vol. 4 (1995) 126-133

[9] S.Mathur, P.C.Gope, J.K.Sharma, Prediction of Fatigue Lives of Composites Material by Artificial Neural Network, Proceedings Of the Society for Experimental Mechanic 2007 Annual Conference and Exposition, Springfield, Massachusetts, USA, June 4-6, 2007. Paper 260.

[10] J.R.Mohanty, B.B.Verma, P.K.Ray, D.R.K.Parhi, Application of artificial neural network for fatigue life prediction under interspersed mode-I spike overload, Journal of Testing and Evaluation, Vol. 38, No.2, Paper ID JTE101907

[11] J.R.Mohanty, B.B.Verma, D.R.K.Parhi, P.K.Ray, Application of artificial neural network for predicting fatigue crack propagation life of aluminum alloys, Archives of Computation Material Science and Surface Engineering, 2009. Volume I, Issue 3. 133-138

[12] J.A.Newman, N.E.Dowling, A crack growth approach to life prediction of spot-welded lap joints, Fatigue and Fracture of Engineering Materials and Structures, 1998, 21, 1123-1132

[13] P.Orbanić, M.Fajdiga, A neural network approach to describing the fretting fatigue in aluminum-steel couplings, International Journal of Fatigue, 25 (2003) 201-207

[14] T.T.Pleune, O.K.Chopra, Artificial Neural Networks and Effects of Loading Conditions on Fatigue Life of Carbon and Low-Alloy Steels, ASME Pressure Vessels and Piping Conference, July 27-31 1997, Orlando, USA

[15] I.Sohn, D.Bae, A study on the Fatigue Life Prediction Method of the Spot-welded Lap Joint, Seoul 2000 FISTIA World Automotive Congers, June 12-15, 2000, Seoul, Korea

[16] M.Šušnjar, Development of predictive method for determination of material fatigue life, doctoral dissertation (in Croatian), FESB, 2012.

[17] V.Venkatesh, H.J.Rack, A neural network approach to elevated temperature creep-fatigue life prediction, International Journal of Fatigue, 21 (1999) 225-234

[18] Wei-Yin Loh, Yu-Shan Shih, Split Selection Methods for Classification Trees, Statistica Sinica 7 (1997) 815-840

[19] Databook on fatigue strength of metallic materials, The Society of Materials Science, Japan, Volume 1, 1996.

[20] Databook on fatigue strength of metallic materials, The Society of Materials Science, Japan, Volume 2, 1996.

[21] Databook on fatigue strength of metallic materials, The Society of Materials Science, Japan, Volume 3, 1996.

Agile Methods for Cloud Computing

S. Kalem, D. Donko and D. Boskovic
Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina
samra167@gmail.com, ddonko@etf.unsa.ba, dboskovic@etf.unsa.ba

Abstract – Since agile development was invented in the mid-1990s, it changed the way of doing software development. The emphasis is placed on a short development cycles based on a feedback from customer. But, agile development was missing important component – development platform for supporting fast development cycles. That missing linking component can be recognized in a cloud computing, which eliminates major distribution requests which can downgrade agile development. The goal of this paper is to describe the connection of agile methods for software development with cloud computing platform. By describing that connection, benefits and improvements in whole software development process are being pointed out. As a practical part of this paper, the application for warehouse management is developed applying agile method called Dynamic Systems Development Method (DSDM) on Google App Engine platform as a service. This development is compared with the development of the application for warehouse management by agile methods but in a traditional way.

I. INTRODUCTION

Agile software development methods are attempting to offer an answer to the eager business community asking for less extensive software development methodologies with faster software development processes [1].

Agile methods are focused on disclosure of requirements through collaborative process of design where software and features are often being released and that allows end users and developers to experiment. Developers can implement and deliver user requirements in a short period of time; users can review implemented functionalities and give their remarks and suggestions. After that, developers can take into consideration those remarks and suggestions and improve the implemented part of the system. This creates highly interactive environment between developer and user of the application. Linking agile methods with cloud computing, the higher level of service is being offered while eliminating loss and lowering the expenses.

Cloud services are responsible for fast fulfillment of user requirements. Agile methods tend to separate project requirements in smaller, achievable segments. This approach guarantees user's feedback information on every task of the project. Segments can be planned, developed and tested individually to maintain high quality standards and to avoid different problems that may occur.

The main benefits when using agile methods in conjunction with cloud computing include:

- Increased quality of application
- Effective use of resources
- Lowered time-to-market
- Cost savings

The main goal of this paper is to demonstrate the mentioned benefits of linking agile methods with cloud computing. Firstly, a short description of agile methods and cloud services is given and it is emphasized what are the gains of this connection. The goal is to demonstrate these gains throughout the rest of the paper based on practical work. Next section gives a short introduction of a selected organization and presents suggested cloud models based on analysis of infrastructure and operation of the organization. In the third section chosen agile method for development is being shortly described through warehouse management application. Also, main functionalities of this application are given in this part. After describing functionalities, the agile development is compared with agile development in conjunction with cloud computing. Comparison is made upon two types of development. The one is using only agile methods and the second one is using agile methods and cloud computing. Based on this comparison, the conclusion is made and presented in final section of this paper.

II. SPRINT CLOUD COMPUTING

In this section, an introduction to a Sprint company is made and the cloud models that this organization adopted are presented.

A. Sprint company - introduction

Sprint is a company that provides services of developing and implementing software solutions, making complete IT infrastructure and giving software support for companies of different sizes. The organization is consisted from three departments: software development department, finance department and ICT department. Business development of the organization is focused towards development and implementation of software solutions for efficient execution of tasks, managing resources and planning.

Sprint made a transition to cloud computing from traditional computing methods and businesses processes to increase the effectiveness of IT initiatives, reduce cost of in-house operations, increase operational flexibility and generate a competitive advantage. As initial point that served as the motivator for continuing to consider the

decision of migrating to cloud computing was the identification of Sprint suitability for adopting cloud computing based on mathematical model [2]. This mathematical model serves as initial guide when questioning the decision of making a transition to cloud computing. Several factors including industry drivers, size of the organization, business needs, security policies and the existing base of the organization were considered during identification of cloud model that best fits an organization's needs. Based on carried analysis and business needs, the suggested cloud model was implemented and it is presented in the sequent part of this section.

B. Sprint adopted cloud models

Fig. 1 illustrates Sprint hybrid implementation model. Hybrid cloud consists from organization's private cloud and public cloud from which the organization uses certain services. Inside private cloud, users can use available server, storage, network and application services, while during software development; the department for software development can use the solution that is offered by Google public cloud – platform as a service. Google App Engine is used during application development and testing. Hybrid implementation model as suitable implementation model was proposed and adapted based on business and infrastructure analysis of Sprint.

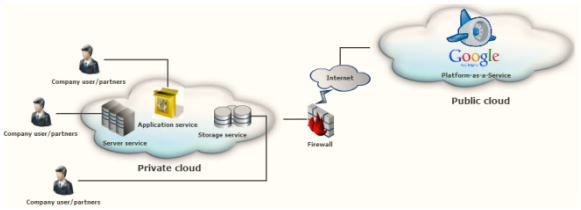


Figure 1. Spring hybrid implementation model

Since hybrid model is composed of private and public cloud, it was relevant to analyze the characteristics of both of these models and to examine their applicability to Sprint. Some of the characteristics of private cloud that were analyzed and that are applicable to Sprint are [4]:

- Organization is offering IT services to a large number of partners as a part of core business.
- Company's data are relevant. To maintain control, it is important to keep information behind firewall.
- Organization has big data center that works well with a lot of spare capacity. It would be much expensive to use public cloud even if it is necessary to add new software to transform that data center into cloud.

Some of the characteristics of public cloud that were analyzed and that are applicable to Sprint are [4]:

- When large number of users uses applications with standardized workload.
- When testing and developing application code is done.
- When collaboration projects and ad-hoc software projects using platform as a service are done.

Besides implementation model, the organization offers infrastructure as a service and software as a service, while platform is rented from Google cloud provider. Google App Engine platform as a service is used for application development. An example of software as a service that will be offered to clients as a cloud service is warehouse management application that is mentioned in next section.

III. CASE STUDY

As an introduction to this section, commonalities for two ways of carried development of the application are described. Later in this section, it is described the way the application was developed before Sprint migrated to cloud computing and the way the application was developed using adopted cloud model and available cloud services.

A. Commonalities of the development of the application

The chosen agile method for developing warehouse management application in both cases was DSDM.

DSDM model is consisted from three phases. First phase is called pre-project phase where project goals and priorities of feasibility study were defined. During this phase, participation of all necessary professional services in analysis were ensured, which took place in the next phase. Second phase was a phase of a project life cycle, which was realized through following activities and sub-activities [6]:

- Study – feasibility and business study
- Functional Model Iteration
- Design and Build Iteration
- Implementation

Third phase, which is also called post-project phase ensures efficiency and effectiveness of functioning. It is realized through maintenance – functional upgrading and error correction according to DSDM principles [6]. Because these projects weren't used in professional settings, this phase was not done thoroughly.

The project was carried out by the team of four students while attending a subject – Advanced Software Engineering, at the Faculty of Electrical Engineering in Sarajevo. During development, team members switched roles in iterations so each team member contributed to design, programming and testing. Also, additional roles were assigned according to the DSDM roles. The end-users of the application were the professor on the subject and his assistant.

Sprint organization that is described in second section is organization that was selected as a first part of work on the project on the mentioned subject. Each team had to select separate organization for which they would develop one application and then consider migrating to

cloud computing and re-develop the same application to point out the main benefits of using cloud computing with agile methods.

Project was realized through three iterations in both cases. During development, iterations took fixed period of time that had to be according to the duration of attending the mentioned subject.

The implemented software is supposed to follow the warehousing processes, to record them and to contribute to better understanding of the operations of the company's warehouse via reports and other forms of information presentation.

Fig. 2 represents detailed look of the software functionalities, illustrates inputs and outputs from modules and interaction between some components.

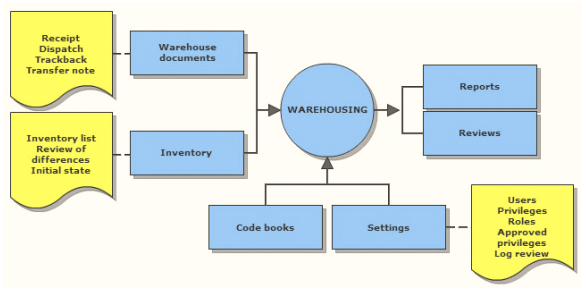


Figure 2. The functionalities of warehouse management application

B. *Application development without use of cloud computing*

Before migrating to cloud computing, the team developed the application as web application using ASP.NET, Microsoft Visual Studio as IDE and Microsoft SQL Server as the database for the application. Team worked together during design phases on iterations so that the best possible solution for the application could be made.

Also, each team member had to install on their working machines the right version of MS Visual Studio and MS SQL Server.

After completion of each iteration, the working part of software had to be copied on a CD and taken to professor for examination and marking. After completion of all three iterations, the integrated version of application was copied multiple times so that the final work could be evaluated.

C. *Cloud based application development*

Application was developed using cloud platform as a service – Google App Engine.

The application was developed as web application. The development environment was Eclipse. Google App Engine has plugin for developing application in Eclipse IDE. During implementation, the developers created design prototype which was then tested and revised. During testing, tests were created and run parallel through cloud machines.

After creating first increment of the application, that increment was then uploaded on Google App Engine Platform so the professor accessed it, performed validation and eventually added some new requirements. Application was offered to professor as service that can be accessed through web browser. This procedure was applied on finished features of second and third iteration and then on integration of finished features of these iterations. After finishing application, it was delivered as a whole, working solution to the professor for examination and marking.

Although, the same development environments weren't used in cases, the concepts and the way of development were the same. Same tasks, same roles were applied to same team members during both ways the application development.

IV. AGILE METHODSAND CLOUD COMPUTING

Agile development is based on iterative and incremental approach. Iteration is expressed through time frame with usual duration of 2-4 weeks, and increment is functional product that is delivered at the end of iteration [7]. Agile development enables continuous development of software product, and cloud computing continuously delivers products to buyers. Cloud computing, in a very short period of time, provides tools and infrastructure for teams doing agile projects, while team provides continuous value to the buyer using the cloud.

Agile manifesto says: "Individuals and interactions over processes and tools" [5]. This means that tools don't initiate or limit the interaction. The goal is to minimize focus on tools, what is one of the many benefits of cloud computing. Developing applications for cloud represents production environment that many teams can use almost immediately. While most of the products require on-premise installation, using cloud product, the need of the administrators to run rigorous and often long procedures when installing new versions of software products is reduced. In traditional, on-premise world, this requires coordinated efforts of preparing user installation process, then persuading certain number of administrators who have local installations of product to upgrade their environment and after that, transition to often rigorous process of upgrading. Connecting agile methods with cloud computing is combination which lowers development expenses, time to market and improves productivity.

A. *Gains during agile development with cloud computing*

Gains and benefits during agile development with cloud computing can be seen in many phases of agile development. These gains affect different steps and characteristics connected to agile development, so that developed application is built faster and the quality of the application is better.

The gains can be seen through:

- Infrastructure
- Frequent communication
- Model sharing
- Application development lifecycle
- Testing and integration
- Prototyping and demo versions

Infrastructure is needed before the development begins, and software development requires infrastructure flexibility. When using agile methods without cloud computing, high and continuous capital expenses are required in advance which include software licenses and current upgrades of that software, purchasing hardware, improvements of network infrastructure, tools for monitoring/testing, security products, maintenance, IT expenses and etc. Connecting agile methods with cloud computing, the need for risky infrastructure investment plan is removed. Hardware and software infrastructure is supported, where developers can implement and run applications directly without exceeding hardware and software for application development using one of the properties of cloud computing – elasticity. This property means that resource allocation can get bigger or smaller depending on demand.

Since one of the main reasons of project failure is lack of communication between all participants caused by inadequate or non-existing tools, cloud can solve that problem. There are many software as a service solutions to help collaboration between people when exchanging and sharing resources.

In early development stages, different project participants can have different requirement conception what leads to communication gaps. Cloud based environment enables group sharing, commenting and improving different models what significantly leads to lowering gaps between iterations [9].

When cloud computing is used, the tools for application development lifecycle management are integrated with cloud environment. Thus, they can be run on the cloud or can be used for developing applications on the cloud, all within cloud environment.

Traditionally, clients are called to the development team location for testing. Client arrival to the location and providing application instances for a large number of clients becomes demanding. But, using cloud computing, clients can access the application from any location where Internet connection exists. Using cloud computing for application testing, changes in current code can be made for further testing and incremental implementation can be done in cloud environment; the only thing that is necessary is that the tester/buyer have the access to the URL of the cloud environment through which they can access the application for testing. Clients can review the increment and give valuable feedback.

Thus, long feedback cycle can be eliminated, reducing the probability of wrong said or wrong interpreted requirements. This reduces the time and efforts for the organization that develops the software, thereby increasing the user's satisfaction [9]. Delivering software as cloud service, the needs of the end user are integrated on more efficient way. This access stimulates higher degree of innovation, disclosure of requirements and validation in cloud computing.

Working prototypes built by the developers locally and implemented on cloud platform can immediately be available to buyers to obtain faster feedback information. Additional risks connected with unapproved prototypes are also minimized versus traditional on-premise implementation model where significant investments in setting up the infrastructure have to be made so the prototype can be accessible.

B. Comparing the development of application when using agile methods with and without cloud computing

Agile development flow for whole project (all the iterations) is shown on figure 3. Figure 3 shows phases of agile development with assigned duration time. Mentioned benefits in previous part of this section in large scale affect the duration time of phases shown on Fig. 3.

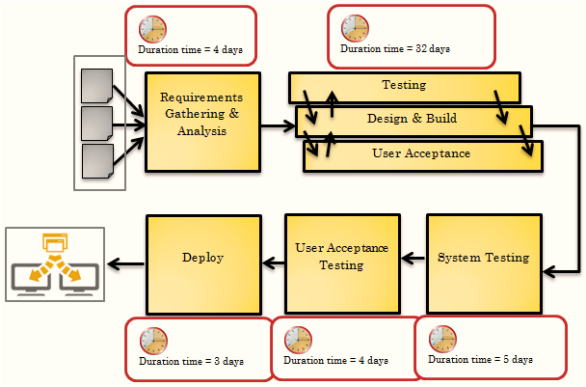


Figure 3: Development time in days combining agile methods with cloud computing

Graph on Fig. 4 shows comparison of duration of application development when using agile methods without and with cloud computing.

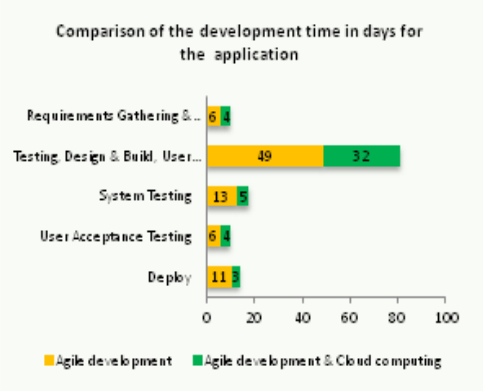


Figure 4: Development time in days combining agile methods with cloud computing

Shorter duration time in phases of agile development with cloud computing are a result of exploitation of all possible benefits of this conjunction. The exploited benefits of mentioned conjunction are:

- Improved communication using cloud tools like Skype when holding meetings for determining details about some functionalities of the application, since client wasn't able to attend the meeting. As a way of document sharing and when performing different questionnaires, Google Docs and Google Forms were used.
- Using Eclipse IDE as working environment, development of the application is made easier. As an addition, other Eclipse plugins can be used like plugin for calculating the metrics of produced code or plugin for easier storing the code on repository.
- The need for installing certain tools and their version is removed. This caused a lot of problems when developing the application without the use of cloud computing. Due to tools incompatibility, the beginning of the project was stooped until the problem was solved. This caused delay of whole project and increased the project expenses since the iteration had to be prolonged two more weeks than planned.
- Since the implementation on the cloud is done automatically, the expenses, time and effort is lowered as opposed to the traditional system implementation on the location of the client.
- Validation is performed after uploading the increment or whole application on the platform – in this case, on Google App Engine platform using Google Eclipse Plugin. User acceptance testing was done much faster and easier since the clients accessed the application through the web browser.
- During second project iteration when using cloud computing, new requirement had to be included in the implementation. Therefore, the resources had to be increased. It didn't affect the duration time of the phase because cloud computing enables fast allocation of resources and easier change management.

V. CONCLUSION

The goal of this paper was to show all the benefits of connecting agile software development methods with cloud computing as a new concept and a way of improvement. Different resource and other characteristics of selected company are examined to shown which factors contribute to decision to migrate to cloud computing solution. The proposed adopted cloud computing solution is presented and DSDM method is shortly described through case study of warehouse management application. Software development with agile method is compared with software development with agile method that utilize cloud computing. All advantages of second approach are pointed out.

It can be said that the future of IT will be cloud computing, or software applications which are delivered and used through the Web browser. By delivering and developing applications in this way, agile software development methods enable gaining superior quality of software and make easier frequent changes to software requirements.

REFERENCES

[1] P. Abrahamsson, O. Salo and J. Ronkainen, "Agile software development methods: Review and analysis", 2002.

[2] S.C. Misra and A.Mondal, "Identification of a company's suitability for the adoption of cloud computing and modeling its corresponding Return on Investment", Mathematical and Computer Modelling, pp. 1-10, 2010.

[3] E. Burke, "Categorizing Data Sensitivity for Computer Security", Interface, nb. 222, 2002.

[4] J. Hurwitz, R. Bloor, M. Kaufman and D. Halper, "Cloud Computing For Dummies", Indianapolis: Wiley Publishing, Inc., 2010.

[5] Manifesto for Agile Software Development, Agile Alliance, 2001. http://agilemanifesto.org/

[6] "wikidot", 20. August 2011. http://dsdmofagilemethodology.wikidot.com/

[7] M. Granville G., "The Characteristics of Agile Software Processes", IEEE Computer Society, pp. 1-3, 2001.

[8] "Google Developers", Google, 4. July 2012. https://developers.google.com/appengine/

[9] R. Shriver, "Agile Cloud Development", The Virtualization Practice, LLC, 4. June 2012 http://www.virtualizationpractice.com/agile-cloud-development-the-future-of-software-16226/.

Generation Technique for Django MVC Web Framework Using the Stratego Transformation Language

D.V. Annenkov*, E.A. Cherkashin**
* National Research Irkutsk State Technical University, Irkutsk, Russia
** Institute of System Dynamics and Control Theory SB RAS, Irkutsk, Russia
annenkov@ib-soft.ru, eugeneai@icc.ru

Abstract - Domain Specific Languages (DSL) allows one to raise level of abstraction, improve development productivity, and establish an equitable communication between domain experts and developers. Language-oriented programming (LOP) is a new paradigm based on DSL construction, allowing separating domain-specific and technology-specific aspects of a system under development. LOP shares some ideas with model-driven architecture and model-driven development. Spoofax language workbench is used as a primary tool for DSL design, and based on Stratego, a transformation language with programmable rewriting strategies, and Syntax Definition Formalism as language for grammar definition. As an example of DSL a simple textual language for domain modeling is considered. Rewriting rules and strategies are used as an uniform approach to generate, validate DSL code, and make arbitrary abstract syntax tree transformations. Rules for code generation implemented using so called "string interpolation" technique. Source DSL code translated to python code that can be deployed within Django web framework, resulting to a web-application with create/update/delete functionality on a corresponding database. Developed DSL is an example of "definition by transformation" approach. To get real benefits from DSL we need to add more domain specific features in DSL.

I. INTRODUCTION

Language-oriented programming (LOP) can be considered as style of software development which involves use of a set of domain-specific languages (DSL) instead of general-purpose languages (GPL) [1]. DSL allows one to capture requirements in the user's terms. DSL is a programming language designed for a particular domain. Because of focusing on specific class of problems it allows to express the domain in more precise terms as compared to GPL and other modeling tools such as UML.

There is interconnection between model-driven development (MDD) and LOP: both of them tends to reduce gap between problem domain and its implementation. It is a step in rising level of abstraction after GPL such as Java, C#, Python etc (Fig.1). LOP can be considered as new programming paradigm that tends to unite such approaches as generative programming, model-driven approaches (Model Driven Development, MDD, Model Driven Architecture, MDA), intentional programming [2]. Main advantages of LOP:

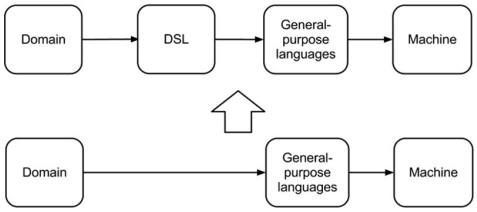


Figure 1. Rising level of abstraction from GPL to DSL

- improved developers productivity;
- communication with domain experts;
- declarative approach to programming (define *what* one is going to obtain, not *how* it is to be obtained).

To support language-oriented programming one need a development tool that known as *language workbench* [3]. Language Workbench provides tools for DSL definition (parsing, transformation, code generation), integration between DSLs, rich editing environment (code highlighting, static analysis, code completion and other modern IDE features).

There are number of different approaches how to develop and use a DSL. Comparison of some approaches as internal DSL, compile-time metaprogramming and strategic term rewriting represented in [4]. The xText framework is one of relatively wide-used tools for textual external DSL development. It works on Eclipse platform and uses EBNF for syntax definition with ANTLR as parser generator [5]. Another approach is *projectional editors (PE)* implemented in JetBrains Metaprogramming System (MPS) [6]. Projectional editors is an alternative to source editors. In MPS developer deals with Abstract Syntactic Tree (AST) directly using PE to edit the tree.

In this article we consider an approach for DSL development based on strategic term rewriting. Spoofax language workbench is used as a primary tool for DSL building. Spoofax is based on Stratego, which is a transformation language with programmable rewriting strategies and Syntax Definition Formalism (SDF) [7], as language for grammar definition. In Stratego, DSLs are implemented through term rewriting, where a source DSL program is transformed to a target program (Python, Java etc.) using a set of transformation rules and strategies.

Spoofax language workbench allows covering all main aspects of DSL construction [8]:

- grammar definition and parsing;
- semantic analysis (DSL program validation) ;
- DSL code transformations;
- target code generation;
- integration of the DSL and its tools into an IDE.

One of the best practices of DSL development is to *semantic model* is a part of DSL. Semantic model used in this work in same sense as in [1]: "Semantic Model is the library or framework that the DSL populates". Semantic model provides a runtime context of the code generated from DSL.

We use Django web framework as the semantic model. Source DSL code is translated to a python code that can be deployed within this web framework, resulting in a web-application administration subsystem with create/update/delete functionality on a corresponding database.

II. DEFINING THE DSL

We consider a simple textual DSL which describes system under development as a set of related entities. DSL may be used as a programming or as a modeling language. The difference between the modeling and the programming properties of the languages is somewhat blurred and not precisely defined. Using criteria proposed in [9] we can define DSL under development to be closer to modeling than programming language.

Customer Relationship Management is considered as a problem domain. We model primary data describing customers and its contacts using the DSL. Let's start with the example.

```
entity Customer
  name            : String
  description     : String
  website         : URL
  repr name
end
```

This fragment of code represents a `Customer` entity with attributes `name, description` and `website`. The `repr` keyword used to define string representation of the entity. In this example the value of the `name` field used as a string representation of the Customer entity.

Entities can be related using many-to-one associations. For example:

```
entity Contact
  name   : String
  phone  : String(11)
  email  : Email
  customer -> Customer
end
```

Contact entity associated with `Customer` entity using `customer` property.

Using Stratego we can define grammar of the DSL using SDF notation:

```
context-free start-symbols
        Start
  context-free syntax
    "module" ID Definition* -> Start {cons("Module")}
    "entity" ID Property* Repr? "end" -> Definition
                                    {cons("Entity")}
    "repr" ID           -> Repr {cons("Repr")}
    ID ":" Type         -> Property {cons("Property")}
    ID "->" EntityAssoc -> Property {cons("Property")}
    ID                  -> Type {cons("Type")}
    ID "(" PINT ")"     -> Type {cons("Type")}
    ID                  -> EntityAssoc
                           {cons("EntityAssoc")}
```

Productions are annotated with a constructor name `n` to uniquely identify them in the abstract syntax tree using the `{cons(n)}`annotation. ID represents an identifier consisting of chars, digits and underscore symbols:

```
[a-zA-Z][a-zA-Z0-9\_]* -> ID
```

PINT represents positive integer with no leading zeros:

```
[1-9][0-9]*  -> PINT
```

Stratego generates corresponding algebraic signatures that describe abstract syntax of DSL.

```
signature
  constructors
    EntityAssoc : ID -> EntityAssoc
    Type        : ID * INT -> Type
    Type        : ID -> Type
    Property    : ID * EntityAssoc -> Property
    Property    : ID * Type -> Property
    Repr        : ID -> Repr
    Entity      : ID * List(Property) * Option(Repr) →
                       Definition
    Module      : ID * List(Definition) -> Start
                : String -> ID
                : String -> PINT
```

Now we can define rewriting rules for transformation and code generation. A rewrite rule has the form

`R: p1→p2`, where `R` is the name of the rule, `p1` the left-hand side pattern of the rule and `p2` the right-hand side pattern. A pattern is a term with variables.

Let's write rewriting rule for top-level form of the DSL, for a module. A module has name and contains one or more entities.

```
to-django-model:
        Module(x, d*) ->
        $[ # -*- coding: utf-8 -*-

            from django.db import models

            [d'*]
        ]
        with
            d'* := <map(to-django-model)> d*
```

Left side of *`to-django-model`* is a pattern `Module(x, d*)`. Pattern matches against AST nodes and if successful, `x` binds with the module name and `d*` - with the list of entities. Body of a rule represents code template using a *string interpolation* technique. Text within `$[...]` block remains unchanged, except block in square brackets like "`[d'*]`", that interpreted as a variable. In `to-django-model` rule above variable `d'*` is assigned in a "`with`" clause . The expression `<map(to-django-model)> d*` is similar to `map` function in *functional languages*: `to-django-model` rule applied to every item in list d*.

Let's define next rule for `Entity` translation.

```
to-django-model:
    Entity(x, p*, r) ->
        $[ class [x](models.Model):
        [p'*]
        def __unicode__(self):
                return "[x]: {0}".format([to-string])
        ]
    with
        to-string := <to-string-repr> r;
        p'* := <map(to-django-model)> p*
```

Here we use name `to-django-model` for the rewriting rule, but this rule has a different pattern to match. In this case a successful match binds `x` with an entity name, `p*` with property list and `r` with `repr` field. The `r` value is optional and can be either `None()` or `Some(Repr(name))`. And we have two rules for these cases:

```
to-string-repr:
        Some(Repr(name)) ->
        $[self.[name]]

to-string-repr:
        None() ->
        $[self.pk]
```

The rules can be read as follows: use field with specified `name` or use `pk` field if no `repr` provided. Variable p* is used to obtain a transformed code for each property using <map(`to-django-model`)> strategy with the following rule:

```
to-django-model:
        Property(x, t) ->
       $[    [x] = [field_type]
        ]
        with
            field_type := <to-django-model> t
```

And at the last step is to define rules for every type used in our DSL. Every rule will translate DSL type to corresponding type of Django web framework model field type.

```
to-django-model:
    Type("String") ->
        $[models.CharField(max_length=256)]
to-django-model:
    Type("String", len) ->
        $[models.CharField(max_length=[len])]
to-django-model:
    Type("Int") ->
        $[models.IntegerField()]
…
```

Other type-translating rules have the same form and we omit them here for brevity.

Association field is handled by a separate rule:

```
  to-django-model:
       EntityAssoc(e)  -> $[models.ForeignKey([e])]
```

Applying `to-django-model` rule to the top-level form of DSL (to the module) a complete module for Django web framework is obtained.

Similarly we can define rules for translation to Django admin settings. Also, we need an additional rule to write generated code to python source code files:

```
generate-django-app:
    (selected, position, ast, path, project-path) ->
        None()
   with
     module_name := <get-module-name> selected;
     models := <to-django-model> selected;
     models_file :=
   $[[<project-path>]
[module_name]/models_generated.py];
     <debug> $[Writing [models_file]];
     mf_handle := <fopen> (models_file, "w");
     <fclose> <fputs> (models, mf_handle);
     amdin_file :=
   $[[<project-path>]
[module_name]/admin_generated.py];
     <debug> $[Writing [amdin_file]];
     af_handle := <fopen> (amdin_file, "w");
     <fclose> <fputs>
       (<to-django-admin> selected, af_handle)
```

Stratego language allows one to define side-effects in rewriting rules. In the rule above we use <`fputs`> strategy to write the result of a code generation and <`debug`> strategy to provide some information to console.

One can assign `project-path` variable to the path to the Django project and set this rule on "save" action using Spoofax, which is activated on saving changes in DSL source code. As a result, all changes to the DSL code will be reflected in the files of Django project.

III. CODE CHECKING AND COMPLETION

One of the great advantages of the usage of the strategic term rewriting techniques is an ability to express aspects of DSL such as code transformation, code validation and context completion using the same notation. Let's consider support of code checking and completion using Stratego language.

Spoofax workbench generates sample rules for code analysis then creating the project. Main rule is as follows:

```
editor-analyze:
    (ast, path, project-path) ->
        (ast, errors, warnings, notes)
    with
        editor-init;
        analyze;
        errors   :=
           <collect-all(constraint-error, conc)> ast;
        warnings :=
           <collect-all(constraint-warning, conc)> ast;
        notes    :=
           <collect-all(constraint-note, conc)> ast
```

One can add rules such as constraint-error, constraint-warning, constraint-note to define custom error checking, warning and notes. The `collect-all(s, un)` strategy collects all subterms where s strategy succeed with user-defined union operator `un`. In this case the union operator is a list concatenation.

Consider a rule that checks whether the property specified in `repr` clause belongs to the entity.

```
constraint-error:
    Entity(x, p*, Some(Repr(prop))) ->
          (prop, $[[prop] is not a [x] property])
    where
         not(<some(?Property(prop, _))> p*)
```

This is just an ordinary rewriting rule that rewrites Entity node to the tuple consisting of property name and the error message. This rewriting succeeds only if condition in where succeeded. The rule `not(<some(?Property(prop, _))> p*)` used as condition can be read as follows: succeed only if there are no property of name prop that belongs to the entity. `?Property(prop, _)` is a pattern to match. The `some(s)` strategy applies the parameter strategy `s` to as many direct subterms as possible and at least one. Application of `some(s)` strategy fails if no successful application of parameter strategy `s` occurs.

Code completion or autocomplete allows predictions of a code fragment without the user actually typing it completely. It simplifies and speeds up development in text-oriented environments. Spoofax provides ability to define custom code completion rules for DSL.

Consider a simple rule that defines autocompletion for built-in types:

```
editor-complete:
    (Type(COMPLETION(prefix)), position, ast, path,
project-path) ->
       ["String", "Int", "Email", "URL", "Date"]
```

Or another rule for completion of association entities:

```
editor-complete:
    (EntityAssoc(COMPLETION(prefix)), position, ast,
path, project-path) -> proposals
    where
      proposals :=
        <collect-all(?Entity(<id>, _, _), conc)> ast
```

The `?Entity(<id>, _, _)` pattern is a *term projection* that used to an extract entity name. The `collect-all` strategy used to get all entity names and put them into a list which is assigned to `proposals` variable.

Stratego provides a powerful pattern matching engine, combined with special strategies it allows to make queries to AST. Consider some examples.

```
get-all-string-props =
   collect-all(?Property(_,Type("String")), conc)
get-all-entities-with-2-props =
  collect-all(
    where(?Entity(_,<length => 2>,_)), conc)
get-all-entities-assoc-with-customer =
  collect-all(
     where(?Entity(_,
       <some(?
Property(_,EntityAssoc("Customer")))>,_)),
      conc)
```

The `get-all-string-props` collects all properties that have `String` type. Query `get-all-entities-with-2-props` returns all entities having exactly two properties. We use *term projection* technique to impose constraint to properties count. To test the condition (but not to rewrite a node) `where` strategy is used. The `get-all-entities-assoc-with-customer` query is slightly more complicated. It returns all the entities associated with Customer entity. Here we use a strategy with pattern `?Property(_, EntityAssoc("Customer"))` to find out association with Customer if any. One can use similar querying strategies to perform analysis and validation of the source DSL code.

IV. CONLUSION

Rewriting language fits well to DSL development activity using "language definition by transformation" approach. With good support of the IDE one can get a productive environment for DSL development.

Developed DSL is an example of "definition by transformation" approach. It can be used for rapid prototyping of Django applications. To get real benefits from the DSL one need to add more domain specific features to it. For example DSL can be extended with ability to add constraints imposed by the domain. Constraints can be translated to runtime validation rules for input data using a Django form-handling library. Some constraints can be translated not only to a python code. One can perform client-side validation by a JavaScript code and server-side validation by a python code generated from the same constraint. The DSL allows defining business rules in a single place. Since in a general case the rules implemented or generated as a source code are located in various modules within Django project it is not easy to figure out overall picture of defined business rules from that code, that's why the business rules represented in a DSL is a better way of system modeling.

Other domain specific features also can be added easily using rewriting rules with full support of IDE-like features from Spoofax workbench. The approach allows to use the DSL as fully-featured language for domain-driven design approach.

REFERENCES

[1] Martin Fowler, "Domain Specific Languages". Addison-Wesley Professional, 2010.

[2] Sergey Dmitriev. "Language Oriented Programming: The Next Programming Paradigm". URL: [2] http://www.onboard.jetbrains.com/is1/articles/04/10/lop/ (access date: 30.01.2013).

[3] Martin Fowler, "Language Workbenches: The Killer-App for Domain Specific Languages?" 2005. URL:http://www.martinfowler.com/articles/languageWorkbench.html (access date: 30.01.2013).

[4] Naveneetha Vasudevan, Laurence Tratt. "Comparative Study of DSL Tools", Electronic Notes in Theoretical Computer Science (ENTCS), Volume 264, Issue 5, July, 2011, 103-121p.

[5] "xText page on Eclipse platform" web-site. URL: http://www.eclipse.org/Xtext/ (access date: 30.01.2013).

[6] "MPS page on JetBrains" web-site. URL: [6] http://www.jetbrains.com/mps/ (access date: 30.01.2013).

[7] Mark van den Brand, Paul Klint, Jurgen Vinju. "The Syntax Definition Formalism SDF". CWI, 2007. URL: http://homepages.cwi.nl/~daybuild/daily-books/learning-about/sdf/sdf.pdf (access date: 30.01.2013)

[8] Lennart C. L. Kats, Eelco Visser. "The Spoofax Language Workbench. Rules for Declarative Specification of Languages and IDEs". OOPSLA '10 Proceedings of the ACM international conference on Object oriented programming systems languages and applications, 2010, pp. 444-463.

[9] Yu Sun, Zekai Demirezen, Marjan Mernik, Jeff Gray, Barrett Bryant. "Is My DSL a Modeling or Programming Language?" University of Alabama at Birmingham, USA, 2008.

ISS

International Conference on

INFORMATION SYSTEMS SECURITY

Steering Committee

Chairs: Željko Hutinski, University of Zagreb, Varaždin, Croatia
Marin Golub, University of Zagreb, Croatia
Stjepan Groš, University of Zagreb, Croatia

PAPERS

Security Evaluation in Cloud

Orlan Mirković
Ericsson Nikola Tesla, Zagreb, Croatia
orlan.mirkovic@ericsson.com

Abstract - While creating the cloud security architecture the opportunity arises to apply IT best practices and the principles of security for a particular domain and to solve a specific set of issues related to security, measurability being one of the hardest. The article explains a measurable model for a cloud by providing a set of controls and defining metrics principles. Further this article enlists the set of tools that can be used to evaluate the security of private, public, community or hybrid cloud.

I. INTRODUCTION

The cloud computing represents a one step further in evolution of providing a cheap, scalable, elastic and secure environment to be used by the user. It is significant that cloud was driven and motivated from business. Unlike other constructs like Virtual Machine, backup to disk, self-provisioning, etc. that focus on a particular aspect of functionality, with cloud computing technology pulled the "synergistic trick" and integrated various disciplines together to meet the business requirements such as:

- Shared resources
- Hardware on demand
- Automated provisioning and relinquishment
- Scalability and elasticity
- Security
- Technology independence

To meet the above requirements and yet to keep the price down the cloud has formed by pooling computing resources in huge aggregate quantities. Also additional hardware was added to the infrastructure on demand, virtualization layer was introduced and the governance of it was automated.

No new technology or paradigm has been introduced. IT departments already have all the same tools, but two key factors make all the difference:

- A convergence between technologies, from their synergies, and from complementary approaches for managing IT resources.
- A critical mass of value added that can be operated and delivered at an acceptable cost.

II. SCALABILITY, INFRASTRUCTURE AND PLATFORM ABSTRACTION AND ITS INFLUENCE ON SECURITY

The need for the cloud infrastructure to scale, to be polymorphic and after all safe has created the need for the state-of-art governance including security. For instance when user requires more servers or more database space, he is not interested in equipment and work load involved in the process- Such requirement calls for additional layer of secure and simple, robust management. When new protocols, virtual machines, servers and other nodes appear and reappear dynamically in the cloud the security becomes intangible. Logging and monitoring rise to another level of complexity and requires a very profound risk assessment in order to introduce meaningful controls. Alerting and vulnerability monitoring become ephemeral.

Such a dynamic and scalable environment requires sophisticated governance. Such governance should continuously display the gap between predefined KPI values given by respective SLAs on one side, and its current state.

To obtain such level of quality and integrity of data the methodologies and controls are of crucial importance. For instance Ericsson Nikola Tesla (ETK) uses ITIL Configuration Management Database (CMDB) for the cloud's Asset and Configuration Management. Also, as we already mention, to secure the automation of monitoring and health of the cloud, it is necessary to add the set of control outside the scope of ITIL. ETK here uses the subset of ISO27001 controls [5].

To determine the Statement of Applicability (SOA) the subset of ISO27001 controls applicable for use the domain and scope must be constrained first through the exercise given in the next section.

III. DOMAIN AND SCOPE

A. Types of Clouds

Typically there are two categories used to define cloud types. The first category defines the deployment model of the cloud. In respect to model the cloud inherits advantages but also the risk from usual systems with the same presence. The cloud can be set as:

- Public
- Community
- Hybrid
- Private

The second category determines the type of the service the cloud is providing. Typical types are:

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)

- Software as a Service (SaaS)

This article presumes the reader is already acquainted with the basic cloud terminology so the above terms will not be explained in this article.

In ETK current projects deal with two types of cloud as shown in Table I.

B. *Architecture*

Controls in general, consequently security controls as well, must be used on measurement points inside the cloud. After measurement acquisition by using bottom-up principles, the acquired data are aggregated and translated into the business-related KPIs. Arguably the following categories represent the essence of the cloud technology:

- Infrastructure
- IP Network
- Virtualization
- Software
- Service Interfaces

Typical architecture is given in Fig. 1 with one important difference. Whilst usual cloud design has separate interface for tenants and for end users, and thus creating a bigger set of vulnerable points, ETK suggest layering the cloud interface as well and thus lowering the number of API, and with that consequently, lowers the risk that is inherently directly proportional to the complexity of the interface(s).

C. *Functional Constraints*

We have already started to set constraints with only private and community cloud type, without Paas and implementing same interface for tenants and users. Further in the ETK Health model the service provider has complete control over all layers of technology so we can basically introduce well know REST concept as a prerequisite for out SaaS cloud type. Of course this assumption cannot be used in the IaaS case used in our labs.

The REST - six constraints applied to the architecture:

- Client–server
- Stateless
- Cacheable
- Layered system

TABLE I. THE DOMAIN AND SCOPE OF THE CLOUD TYPES CURRENTLY USED IN ETK

Type \ Domain	Public	Community	Hybrid	Private
IaaS	Amazon EC2			***ETK Lab***
PaaS				
SaaS	Google	***ETK EMH***		

Non-ETK related clouds are for comparison only

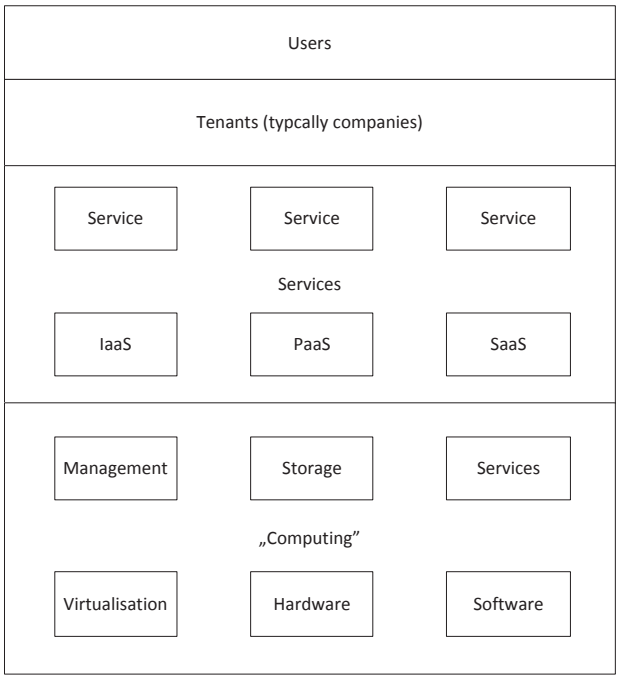


Figure 1. Cloud Architecture Model

- Code on demand (Servers are able temporarily to extend or customize the functionality of a client by the transfer of executable code.)
- Uniform interface

And of course our non-functional requirement: performance, scalability, simplicity, modifiability, visibility, portability, and reliability.

With this the additional constraint we can safely assume we succeeded to significantly lower the number of controls that are required to be implemented

IV. RISK ANALYSIS

By simplification of Whitman's Risk Formula the components defining a risk are given in Fig. 2.

While elaborating all the components would bloat this article, it is essential to clarify the definition of Asset and how to define it in the cloud.

As previously mentioned in this paper, all assets must be listed in CMDB. While CMDB contains Configuration Items (CI) so the asset and their states (i. e. configuration), the business value is usually kept in the Service Portfolio, where the SLA, underpinning contracts and all relevant

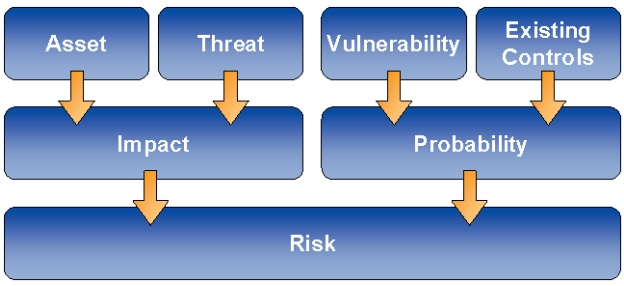


Figure 2. Components defining the risk

KPIs are defined.

Now, with classic systems: dedicated hosting, application servers, etc. we can have many different granulations of all respective CIs belonging to particular systems. The cloud always contains assets that are simply valued differently, depending on the business. For instance, the systems with let's say hundreds of servers, the granularity will definitely be a server, or maybe even rack with a set of servers. With smaller systems CI could be limited on such a small component as a memory or a network card.

Of course the management of such a CMDB would be extremely tedious and time consuming. In order to resolve the contradictory requirements another property called "Storage" has been added with many-to-many relationship to "Asset". Here "Asset" becomes the business value of the asset, and its details are captured in the Service Portfolio, wile "Storage" is becoming our new CI. With this the Variable Granularity of CI is introduced but without cluttering the risk assessment with noise that countless interdependencies introduce.

Of course such a vast and complex system must be kept current all the time otherwise it immediately looses integrity. The Plan, Do, Check, Act (PDCA) model coming from ISO27001, ITIL and many other methodologies and standards is a must.

To recapitulate, four factors determine the success in a cloud Risk Analysis

- CMDB
- Service Portfolio
- Variable granulation of the CI
- PDCA

A. Statement of Applicability

With our domain and scope defined in Table I we can now create a SOA.

It can be argued the SOA should envelope the controls that are relevant to either the Operating Systems or services. The platforms like databases are also inherently a part of the scope of SOA, but with security requirements their interfaces greatly reduce in size and are strictly used by the service (i.e. application) interfaces. After the analysis scope of test is reduced to:

- Communications and operations management controls related to Operational procedures and responsibilities, System planning and acceptance and Monitoring (ISO27001 A.10)
- Access controls related to Operating Systems and Applications (ISO27001 A.11)
- Security requirements controls (ISO27001 A.12)
- Compliance (ISO27001 A.12)

Other controls are considered to be a part of other activities such as development, testing, design etc. In Table II a full set of controls in scope is given.

TABLE II. CONTROLS IN SCOPE

Control	Description
A.12.1: A.12.1.1	Security requirements of information systems Security requirements analysis and specifications
A.12.5: A.12.5.1 A.12.5.2 A.12.5.3 A.12.5.5	Security in development and support processes Change control procedures Technical review of applications after OS changes Restrictions on changes to software packages Outsourced software development
A.10.1: A.10.1.3 A.10.1.4	Operational procedures and responsibilities Segregation of duties Separation of development, test and operational facilities
A.10.3: A.10.3.2	System planning and acceptance System acceptance
A.12.4: A.12.4.2 A.12.4.3	Security of system files Protection of system test data Access control to program source code
A.12.2: A.12.2.1 A.12.2.2 A.12.2.3 A.12.2.4	Correct processing in applications Input validation Control of internal processing Message integrity Output data validation
A.11.6: A.11.6.1 A.11.6.2	Application and information access control Information access restriction Sensitive system isolation
A.11.2: A.11.2.1 A.11.2.2 A.11.2.3 A.11.2.4	User access management User registration Privilege management Password management Review of user access rights
A.11.5: A.11.5.3 A.11.5.4 A.11.5.5 A.11.5.6	Operating system access control Password management systems Use of system utilities Session time out Limitation of connection time
A.10.10: A.10.10.1 A.10.10.2 A.10.10.3 A.10.10.4	Monitoring Audit logging Monitoring system use Protection of log information Administrator and operator logs
A.15.2: A.15.2.2	Compliance with security policies and standards, and technical compliance Technical compliance checking

Table III represents the controls to be added in the future.

V. SECURITY METRICS

Once we have defined SOA, or areas in which we want to measure the effectiveness of our controls it is necessary to identify a meaningful set of metrics. ISO 27001 only defines a set of measurements to be aligned with the organization. Further, metrics can focus on individual control or group of controls as in this article, but we could also start with ISMS processes and then start

TABLE III. THE CONTROLS TO BE IMPLEMENTED

Control	Description
A.10.9	Electronic commerce services
A.10.9.1	Electronic commerce
A.10.9.2	Online transactions
A.10.9.3	Publicly available informationSecurity requirements of information systems

E-commerce and financial transactions

to include relevant sets of controls.

Metrics must be achievable, relevant and timely. SMART criteria are used to create metrics:

- Specific
- Measurable
- Attainable
- Relevant
- Timely

In addition ISO 27001 requires the use of metrics and implementation of the measurements obtained in Plan-Do-Check-Act (PDCA) cycle. In the planning phase, cloud policy should dictate the measurements and define objectives. At this stage, the measurements are to be defined and monitoring implemented. Do phase it to measure and analyze performance. Action ohase is used to improve and fix defects.

It is crucial to automate measurements as it is the only way to collect data in timely manner. Metrics must be expressed as a number or percentage to minimize subjectivity. Frequency of measurement should balance the effort to collect data and their relevance. The best approach is to start with a very small number of measurements and fully automate them, then gradually expand the number.

Table IV illustrates the possible set of metrics. Same metrics can not be applied to all cases. The goal here is to illustrate the use of metrics.

ISO27001 requires that metrics are captured, analysed and reported using well-defined formats and processes.

TABLE IV. EXAMPLE OF SECURITY METRICS

Control	Measurable Description
A.12.1.1	% of applications with the Security requirement Specification (SRS) % Of 3rd party software with security risks assessment done before purchasing
A.12.5.1	% of change with a formal risk assessment report The number of incidents (downtime) due to uncontrolled changes
A.12.5.2	Technical review of applications after OS changes
A.12.5.3	% of software development customization and configuration changes, clearly documented and approved by management
A.12.5.5	% of 3rd party software development contracts addressing the security responsibilities of the seller % of critical software applications with the contractual obligations (SLA, BC) % of applications assessed for compliance to security requirements
A.10.1.4	% of critical applications with a separate test environment
A.10.3.2	% of applications with security baseline % of applications with business continuity plans
A.12.4.2	% depersonified test data used in test harness % of applications with production data in test environments
A.12.4.3	% of applicatin in production with source code
A.12.2:	Correct processing in applications
A.12.2.1	Number of input data validation vulnerabilities detected by the application. The average number of input data validation vulnerabilities across critical applications in scope.
A.12.2.2	Number of processing errors detected by the application. The average number of processing errors detected by critical applications
A.11.6.1	% of applications with documented roles and privileges
A.11.6.2	Number of applications using the same hardware and OS instance % of critical applications in a shared environment
A.11.2.1 A.11.2.2	% of unauthorized users in applications % of applications with more than three admini IDs % of users with unauthorized privileges
A.11.2.3	% of applications with default passwords % of applications with the capabilities to support passwords
A.11.2.4	% of critical applications with periodic review of user access rights
A.11.5.5 A.11.5.6	% Of applications with session/connection timeout
A.10.10.1 A.10.10.2	% of critical applications monitored 24/7 % of the applications covered by log analysis
A.10.10.3	Protection of log information
A.15.2.2	% of applications with security audit reports % of internet web applications with periodic penetration tests Average cycle time to fix critical vulnerability found during the assessment

VI. VULNERABILITY TEST

Once when the exercise around the scope, domain, and functionalities is completed one can start to draw some conclusions out of it. The ISO 27001 standard specifies the risk assessment process, and ISO27002: 2005 provides additional guidance, but either does not provide detailed instructions on how the estimate can actually be implemented. So every organization must choose the approach that is most applicable to the industry scope, complexity and risk types.

Vulnerability test plan can be made up once the assets are in the CMDB, and the risk of their confidentiality, availability and integrity, identified, analyzed and evaluated. The risk assessment is based on the collection of data and processes, as well as all individual input parameters. The analysis will reflect the individual prejudice, so that the process of gathering information to determine what is actually known - and what is not.

Luckily ISO 27001 defines six steps that must be followed in conducting a risk assessment:

- Identify assets in the ISMS.
- Identify threats to the confidentiality, availability and integrity of the defined asset.
- Identify vulnerabilities those threats could exploit.
- Assess the potential impact of these threats.
- Assess the likelihood of these events.
- Risk assessment

The assets have been indentified and controls are chosen only to build the SOA reflecting controls suited for the type of clouds we are interested in.

What is left now is to build a measurement suite. We do that with two simple questions: what to measure and how to measure. Also, it is crucial at this point to lower the level of abstraction and start to recognize the technology in the cloud and build the harnesses around it.

A. *Types of Test*

It is necessary to provide the secure environment given the types of cloud in the scope. Now when there are security requirements defined in Table IV the four types of tests are emerging in the closest alignment with chosen SOA given in Table III:

- Ports and services Enumeration
  - Focus on web applications
- Vulnerability test
  - Sys info and logs
  - "Shared" directories, open connections
- Privacy
  - "Left over" data, analysis of file system
- Security
  - vulnerable applications
  - Rootkits
  - malware
  - zombies

B. *Test Context*

With the context determined the technology dependent areas to build the test harness around emerge.

1) Source of the threat

a) External attacks

b) Systems or applications within the cloud

c) Protecting the privacy of the user

2) Cloud type

a) Infrastructure (IaaS)

- Windows
- Linux

b) Platfrm (Paas)

- MSSQL
- IIS
- Apache
- MySql

c) Service (SaaS)

- VOIP Service
- IMS

3) Test scope

a) Prehardening host OS

b) Checking security from outside the cloud

c) Checking security within cloud

C. *Tools*

The testing tools took a significant part of the time spent in designing this cloud vulnerability tests as their quality enable the visibility, easy enumeration and ease the understanding of the testing results. The list has been created with usability and inter-correlation in mind.

Each candidate has been tried and the following functionalities have been cross-checked:

- Tool name
- Version
- Eavesdropping
- Scanning Tools
- Banner grabbing / Information Leakage
- Vulnerability Scanning
- Password Quality Check
- Denial of Service
- Domain Search

The tools being used are:

- Dsniff
- Ettercap
- Hydra
- IKE-Scan
- IP Fragment Tools
- IPv6 test tools
- John the Ripper
- LdapMiner
- Metasploit
- Nexpose
- Nikto
- Nmap/nmapFE
- Oracle-tools
- Sqlninja
- WebScarab
- Wireshark
- W3af

VII. TEST HARNESS DESIGN

Fig. 3 is an example of how a cloud with defense in depth in mind is implemented. The security domains corresponding to zones in classical design where the security functions in the figure are logical entities and should not be viewed as separate physical security devices. In practice, each security domain is not served by dedicated virtual or physical routers as it was in before

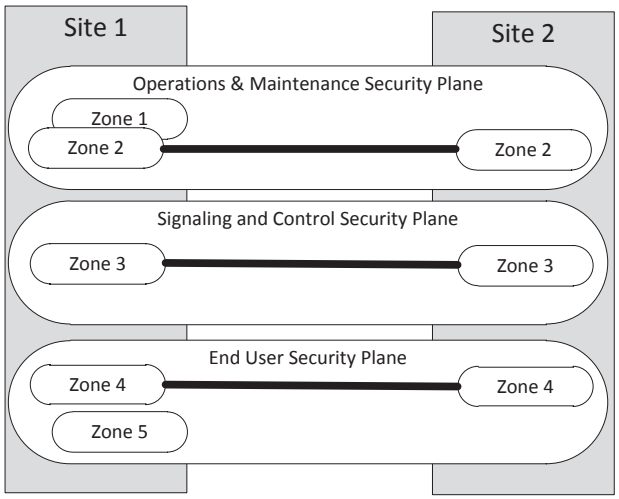


Figure 3. Relation between security planes and security zones

cloud design but rather a logical separation takes place what should be taken into account. It is not possible to traverse from one security zone to another without being inspected by the firewall policy.

To be able to fully test all the cloud components the measurement points were put inside as well as outside the cloud in order to gain full visibility. This leads us to the design given in Fig. 4.

VIII. TEST RESULTS

After setting the testing environment the 3 tests have been run in different times of day.

The audit was performed on 23 systems, 21 of which were found to be active and were scanned. The prehardening was completed on 3 systems. There were 91 vulnerabilities found during the scan. 22 were critical vulnerabilities. 55 vulnerabilities were severe. There were 14 moderate vulnerabilities discovered.

Critical vulnerabilities were found to exist on 12 of the systems, making them most susceptible to attack. 21 systems were found to have severe vulnerabilities. Moderate vulnerabilities were found on 18 systems. No systems were free of vulnerabilities.

IX. CONCLUSION

The results are an eye-opener and they somehow shake our trust in cloud. On the other hand after the analysis of such thorough result showed that almost all vulnerabilities are caused by either weaknesses in the OSs that had not been prehardened, or lack of isolation of sensitive traffic. Hence the controls were drawn and set in the updated security requirement from the test results. After analyzing them actually three groups of controls have been articulated.

- Sites perform the business driven services and they cannot be trusted.
- Sites must perform required prehardening control before entering the production cloud.

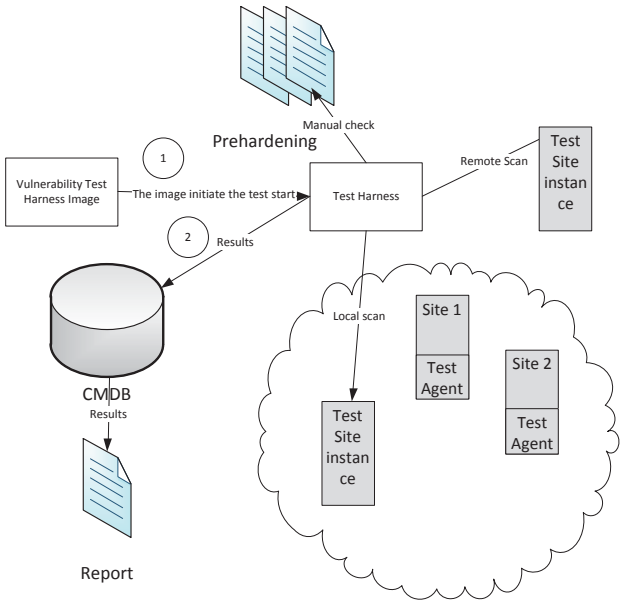


Figure 4. Test harness Architecture

- Defense-in-depth design is crucial. Zone, site and traffic separation with fully enforced set of policies among domains and sites is mandatory.
- As almost all sites are providing services by waiting for incoming traffic the outbound firewall group policy is mandatory.

This approach brings confidence that highly critical service can be brought to the cloud with all the benefits such environment facilitates without downgrading the enterprise level of security. Unfortunately, still being a mostly manual process, the deployment time is all but satisfactory (days). Fully automated harness is a must. Also reporting of aggregate values (% of systems with audit, number of input data validation,…) helps with determining the average "health" of the cloud, but finer granulation and focus on individual vulnerabilities is a must. Further work will focus on not just reporting the individual vulnerabilities but also providing the solution for them [9].

X. REFERENCES

[1] Ericsson White Paper, 284 23-3075 Uen, "Managing Network Security", http://www.ericsson.com/technology/whitepapers/

[2] IETF Request for Comments: 2828, "Internet Security Glossary", http://www.ietf.org/rfc/rfc2828.txt?number=2828

[3] Vic (J.R.) Winkler, "Securing the Cloud, Cloud Computer Security Techniques and Tactics", Syngress, ISBN: 978-1-59749-592-9

[4] L. M. Kaufman, Data security in the world of cloud computing, IEEE, Security & Privacy, vol. 7, no. 4, 2009

[5] Orlan Mirković, Lecture, "Cloud Security Evaluation", KOM 2012

[6] ISO/IEC 27001:2005, www.iso.org

[7] Antonio Regalado, "Who Coined Cloud Computing?", http://www.technologyreview.com/news/425970/who-coined-cloud-computing/, 2011

[8] Wikipedia, "Cloud computing", http://en.wikipedia.org/wiki/Cloud_computing

[9] GIAC Enterprises, Chris I. Cain, Erik Couture, "Establishing a Security Metrics Program", Final Project Report, 2011

Security Assessment of Virtual Machines in Open Source Clouds

Aleksandar Donevski, Sasko Ristov and Marjan Gusev
Ss. Cyril and Methodius University,
Faculty of Information Sciences and Computer Engineering,
Skopje, Macedonia
Email: aleksandar.donevski@outlook.com, sashko.ristov@finki.ukim.mk, marjan.gushev@finki.ukim.mk

***Abstract*—A company security perimeter is impaired if virtual machines are migrated from on-premise to the cloud. Although cloud service providers (CSP) offer some security level to migrated virtual machines, mostly from outside the cloud, the cloud customer is challenged with newly raised security challenges due to multi-tenancy, i.e. the threat from other co-tenant virtual machines hosted inside the cloud. CSP also faces threats from all tenants in the cloud. Both CSP and the tenants can be compromised in case of vulnerability. In this paper we analyze virtual machine security threats that might occur by other tenants or outside the cloud. Also, we compare the security level that the most common open source clouds provide for a particular virtual machine. The security assessment is realized for virtual machines with different operating systems.**

***Keywords*-Cloud Computing, OpenStack Cloud, Security Assessment, Virtualization.**

I. Introduction

Cloud computing market is constantly growing and there are various offers for the customers. For example, the customers can outsource their data, virtual machines or backup to the cloud infinite storage space [1]. Whatever cloud type and deployment model they are using, the access control and assured deletion must be guaranteed by the CSP [2]. A possible solution is to use the security as a service in order to implement and maintain the protection to meet customer's security level requirements [3].

The virtual machine instances in the cloud should be protected against the attacks generated not only by external adversaries, but also by the other co-tenants (virtual machine instances), regardless they are on the same or different physical machine in the same cloud [4]. CSPs should also include the security requirements specified by the customers in virtual machine allocation, besides the traditional performance / cost-effective resource consumption [5].

Selecting of a private cloud offers a possibility to control the security, but it lacks scalability and elasticity features [6]. Usage of public clouds from one side offers high scalability and elasticity, but from the other side it is constrained by the security since most of the control is now transferred to CSP. The problems mainly arise due to multi-tenancy, virtualization, data and application transfer to third party. Therefore, a proper security evaluation is a necessity and highest concern of the customers.

In this paper we focus on OpenStack [7], as an open source scalable solution where more than 60 leading companies participate in its development. Security vulnerabilities for OpenStack clouds are analyzed for possible attacks that may be generated on virtual machine instances both from inside and outside. The objectives are to check the validity of the following hypotheses:

- H1 Segregating the networks will increase the security vulnerability risks from inside the cloud for the OpenStack cloud provider;
- H2 Segregating the networks will not have an impact on the security vulnerability risks from inside the cloud for the tenants; and
- H3 The security vulnerability risks will remain the same from outside the cloud while segregating the networks, both for the tenants and the OpenStack CSP.

We have set the hypothesis H1 since we expect that network segregation will create new cloud services and thus increase security vulnerability risks; the hypothesis H2 since virtual machine instances are on the same network (private or public); and the hypothesis H3 since both virtual machine instances and cloud services have the same public IP addresses. The experimental research uses an OpenStack cloud default installation, where the virtual machine instances are installed with different operating systems.

The paper is organized in the following sections. Section II presents the related work and Section III briefly describes the Openstack cloud architecture and its components. The security assessment methodology is explained in Section IV and the results of the assessment for OpenStack services for inside and outside vulnerabilities are discussed in Section V. Section VI presents the results for the same security assessments for virtual machine instances. Finally, Section VII presents conclusion and future work.

II. Related Work

Additional security challenges and vulnerabilities exist in the cloud when compared to the on-premise solutions. Grobauer et al. [8] define indicators if a particular vulnerability is cloud specific to distinguish general security issues from cloud-specific security issues. Shaikh et al. [9] define data protection, network security, virtualization security, application integrity and identity management are the major types of

security threats. Kaufman [10] addresses the unique security of public cloud compared to traditional information security.

Cloud multi-tenant environment introduces new security risks. Brown et al. [11] summarize the risks of cloud multi-tenancy and propose countermeasures. Zhang et al. [12] propose information risk management framework to identify threats and vulnerabilities in the cloud and to mitigate the security risks.

Several papers propose methodologies and evaluate the security of cloud solutions and service layers. Ristov and Gusev [13] propose a methodology for security evaluation of on-premise systems and the cloud. Chen and Yoon [14] propose a framework of checklist of IT auditing to secure the cloud. However, third party auditor should not bring new vulnerabilities towards data privacy [15]. Ristov et al. [16] evaluate ISO 27001 [17] control objective importance for on-premise and the cloud service layers IaaS, PaaS and SaaS.

Open source cloud security is also assessed. The maximum compliance with ISO 27001 security standard is applied by Eucalyptus and CloudStack [18] in front of OpenNebula and OpenStack [7] open source cloud solutions [19]. Gusev et al. [20] analyze the security vulnerabilities of Eucalyptus cloud from inside and outside and determine that Eucalyptus server nodes are more vulnerable from inside, rather than from outside of the cloud, while virtual machine instances have the same security vulnerability risks both from inside and outside.

Ristov et al. [21] assessed the security vulnerabilities of OpenStack using one network configuration. They proved that OpenStack cloud multi-tenant environment rises new security vulnerabilities risks from inside the cloud, both for the tenants and the OpenStack cloud provider. In this paper we extend the research assessing the OpenStack vulnerabilities both from inside and outside, but using two different OpenStack network configurations.

III. Software Architecture of OpenStack

In this section we describe the software architecture of the newest "Folsom" release of the OpenStack cloud, focusing on the *networking features* that OpenStack cloud has to offer.

A. Software components

The cloud controller is the component that controls the system, network and VM instances. Each cloud node hosts VM instances and communicates with the controller. The OpenStack cloud node consists of seven core components that make this cloud a complete IaaS cloud:

- *Compute (Nova)* - is the core component in the OpenStack cloud which provides virtual machine instances upon demand.
- *Image (Glance)* - a repository for the images in the cloud which provides discovery, storage and retrieval of virtual machine images for the Nova service.
- *Object Store (Swift)* - allows storing or retrieving files from the cloud.

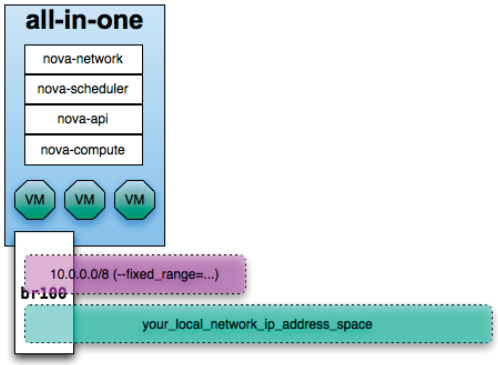


Fig. 1. OpenStack networking example [7]

- *Network (Quantum)* - enables users to create their own networks and attach them to interfaces (Network as a Service).
- *Block Storage (Cinder)* - provides persistent block storage for the virtual machine images in the cloud.
- *Identity (Keystone)* - provides authentication and authorization for all the components in the cloud.
- *Dashboard (Horison)* - provides a modular web-based tool for managing the cloud.

Our testing environment is deployed with all these seven components. The Quantum is a new component in the newest OpenStack release and we focus on its impact to the security of virtual machine instances and OpenStack cloud services.

B. Software aspects of deployment

OpenStack cloud can be deployed only on Linux-based operating systems, i.e Ubuntu, CentOS and RedHat. The components can be deployed independently, i.e they can be deployed on one physical server or each service can be deployed on separate physical server. OpenStack cloud supports all most common hypervisors: KVM [22], Xen [23], UML [24], and Hyper-V [25].

All components of the OpenStack cloud can be deployed as: 1) *Single node* on one physical server; 2) *Dual nodes* on two physical servers; and 3) *Multiple nodes* on separate servers and theoretically unlimited number of compute nodes. The dual and multiple node deployments use an idea to separate instantiation of virtual machine instances from the cloud controller for performances improvements.

Since the goal of this research is the security, in the experimental research we use OpenStack cloud deployment on the Single Node mode, as presented in Fig. 1. We deployed two different network configurations described in more details in the following Section III-C.

C. Software view on networking

The virtual machine instances and OpenStack cloud servers can be configured mainly with two types of IP addresses:

- *Floating (Public) IP addresses*. The main purpose of the floating IP's in the cloud is the need for accessing the

running virtual machines instances in the cloud from the public networks outside of the cloud (Internet).
- *Fixed (Private) IP addresses*. The main purpose of the private network (fixed IP addresses) in the cloud is the requirement for isolating the cloud components and running virtual machine instances from public networks (Internet).

Both the private and public networks are depicted in Fig. 1. The OpenStack cloud allows its customers to deploy the cloud with only one network, i.e. both the virtual machine instances and cloud servers will obtain the floating public IP addresses as a fixed, as well. In this case, the OpenStack cloud services use the public network (floating) for intra cloud service communications.

The security vulnerabilities that arise from isolating the cloud services from the public network into separate private network are the main objective of this research.

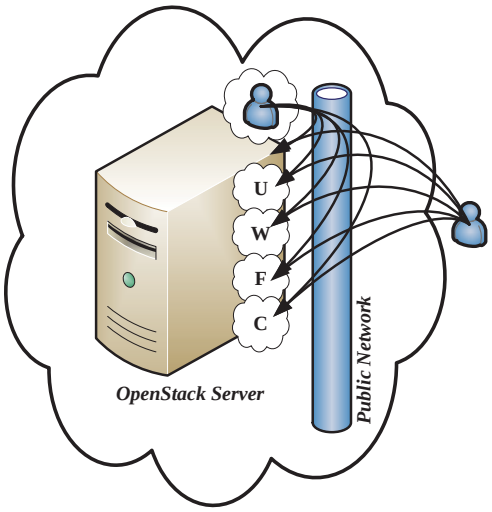


Fig. 2. Security assessments on OpenStack cloud deployed with one network

IV. Security Assessment Methodology

The methodology for security assessment is based on two different network deployments of OpenStack cloud deployed in Single Node:

- *One network* - the same public IP address is used both for floating and fixed IP address for a certain virtual machine instance or cloud server; and
- *Two networks* - two different IP addresses are used for floating and fixed IP address for a certain virtual machine instance or cloud server. That is, the internal services are segregated in separate private network.

The security is assessed both from inside and outside of the physical OpenStack server and virtual machine instances. The main goal is to determine the security vulnerabilities of OpenStack server and several virtual machine instances for different network configurations.

A. Assessment domain

The research domain is based on security assessment of vulnerabilities from two different aspects. The first addresses the assessment of OpenStack node components and the second the operating system instanced on the VM.

The vulnerabilities that might appear in architectural components on physical OpenStack server will be assessed for both the Compute and Controller modules, installed with Ubuntu Server 12.04 (64-bit) operating system.

Assessment of operating system based vulnerabilities will be performed on four virtual machine instances hosted on the cloud, each with one of the following operating systems: 1) Windows 2008 R2 Standard (64 bit); 2) CentOS 6.3 (64 bit); 3) Fedora 17 (64 bit); and 4) Ubuntu 10.04 Server Edition (64 bit). All virtual machine instances are installed with default configuration in order to detect all possible vulnerabilities and propose a measures for their improvements.

B. Test data

The security will be assessed from inside and outside the OpenStack cloud with segregated/non segregated cloud services from the public network for vulnerabilities that might appear from inside the cloud. To perform the two different assessments, we use the Nessus 5 vulnerability and configuration assessment scanner [26] with External Network Scan policy. Each vulnerability is rated according to the associated Common Vulnerability Scoring System (CVSS) [27] score:

- *Info* if CVSS score is 0;
- *Low* for CVSS score $\in \{1, 2, 3\}$;
- *Medium* for CVSS score $\in \{4, 5, 6\}$;
- *High* for CVSS score $\in \{7, 8, 9\}$; and
- *Critical* if CVSS score is 10.

C. Test cases

1) Security Assessment with One Network (Public) Deployed: Fig. 2 depicts the testing environment for the security assessment of the OpenStack cloud deployed only on public network. W denotes the Windows operating system, while C, F and U denote Linux-based operating systems CentOS, Fedora, and Ubuntu, respectively.

Test case 1 is based on security assessment of the vulnerabilities from inside of the cloud. The Nessus client is deployed on one virtual machine instance with Ubuntu 12.04 server operating system for inside security assessment and it's purpose is to scan all four virtual machine instances hosted on the cloud, as well as the OpenStack services deployed on the physical OpenStack server. The test case 1 goal is to assess the vulnerabilities that arise from the cloud multitenancy. OpenStack services, the clients and virtual machine instances communicate among each other using the only one public network.

Test case 2 is based on security assessment of the vulnerabilities from outside of the cloud. The Nessus client is

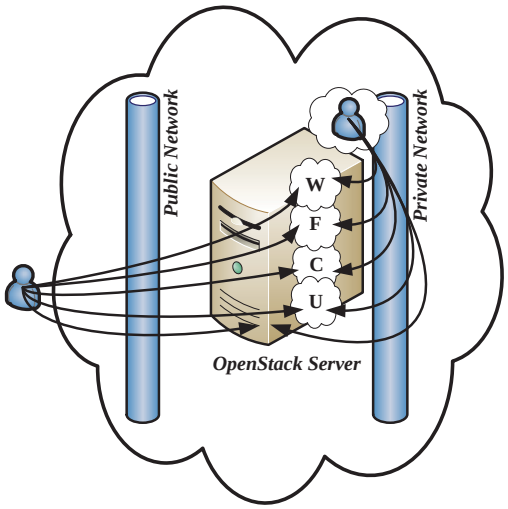


Fig. 3. Security assessments on OpenStack cloud with two networks

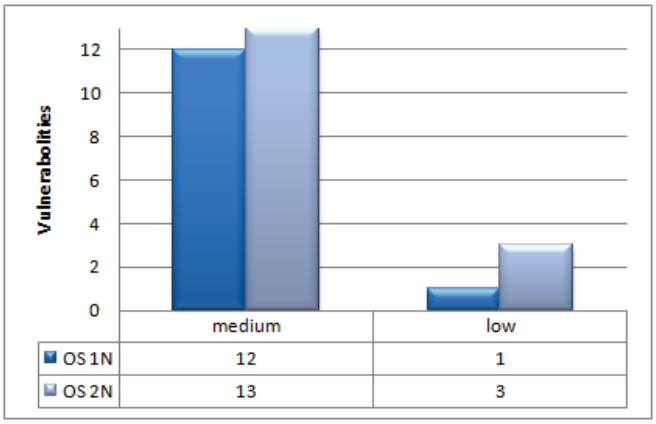


Fig. 4. Summary results of OpenStack security assessment from inside

deployed on a workstation outside of the OpenStack cloud and it's purpose is to scan the same four virtual machine instances which are hosted on the cloud, as well as the OpenStack server.

2) Security Assessment with Two Segregated Networks (Public and Private) Deployed: This section describes the other testing environment of OpenStack cloud, i.e. while deployed with two segregated networks, depicted in Fig. 3.

Test case 3 is similar to Test case 1, i.e. it is based on security assessment of the vulnerabilities from inside of the cloud, but OpenStack services, the clients and virtual machine instances communicate among each other using the private network.

Test case 4 is similar to Test case 2, i.e. it is based on security assessment of the vulnerabilities from outside of the cloud, but when floating network is used only for communication with the outside world.

V. Vulnerabilities of the OpenStack Components

In this section we present the results of security assessments of all four test cases made on the physical OpenStack server. We present only medium and low vulnerabilities since we have not found neither critical nor high vulnerability during both assessments, and the vulnerabilities with CVSS score 0 are informative, rather than real vulnerabilities.

A. OpenStack Server Vulnerabilities from Inside

Fig. 4 depicts the summary results of the security assessment of the OpenStack services from inside the cloud with two different network deployments, i.e. the test cases 1 and 3 for OpenStack services.

Both test cases detected the same one low and 12 medium vulnerabilities. 6 "Web Server Generic XSS", and 6 "Web Server Generic Cookie Injection" on TCP ports: 5000, 8773, 8774, 8775, 8776, and 35357 are medium vulnerabilities, while "Web Server Uses Plain Text Authentication Forms" is low vulnerability on TCP port 80.

These common vulnerabilities can not be secured with reconfigurations. The medium vulnerabilities means that the web server which hosts cloud web services is prone to cross-site scripting attacks and cookie injection attacks. OpenStack Cloud Software [7] (Vendor) should provide a patch for these vulnerabilities. The vendor must redevelop the cloud in order to transmit the user sensitive data using secured HTTPS protocol.

The results prove the hypothesis H1, i.e. additional 1 medium and 2 low vulnerabilities are detected for test case 3 compared to test case 1 for OpenStack cloud services. "DNS Server Cache Snooping Remote Information Disclosure" on UDP port 12217 is the additional medium vulnerability, i.e. DNS server is vulnerable to cache snooping attacks. The DNS software vendor should fix this vulnerability. "DHCP Server Detection" on UDP port 67 and "Web Server HTTP Header Internal IP Disclosure" on TCP port 9191 are the additional low vulnerabilities. Applying filtering will keep the information off the network and mitigate the risk of the first additional low vilnerability, while the second one is not a real vulnerability since public ip addresses will be used in real world cloud.

B. OpenStack Server Vulnerabilities from Outside

Fig. 4 depicts the summary results of the security assessment of the OpenStack services from outside the cloud with two different network deployments, i.e. the test cases 2 and 4 for OpenStack services.

We observe that both test cases detect the same medium and low vulnerabilities, which proves the hypothesis H3 for OpenStack cloud services. Even more, these are the same vulnerabilities as those detected in test case 1 for OpenStack cloud services.

VI. Instance Vulnerabilities in OpenStack

In this section we present the results of security assessments of all four test cases made on the virtual machine instances hosted on OpenStack cloud. Since we have not detected any security vulnerability on Linux-based virtual

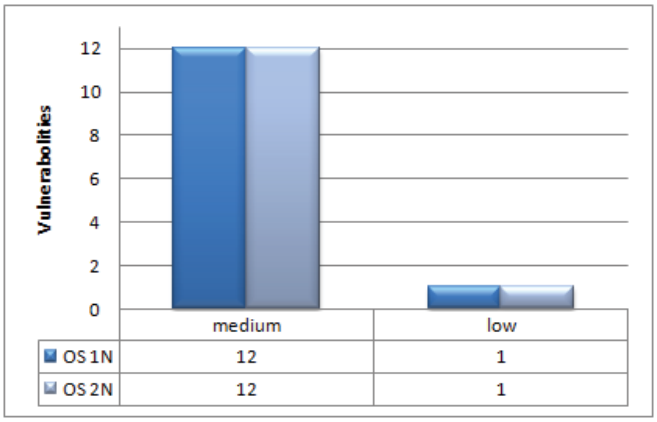


Fig. 5. Summary results of security assessment from outside

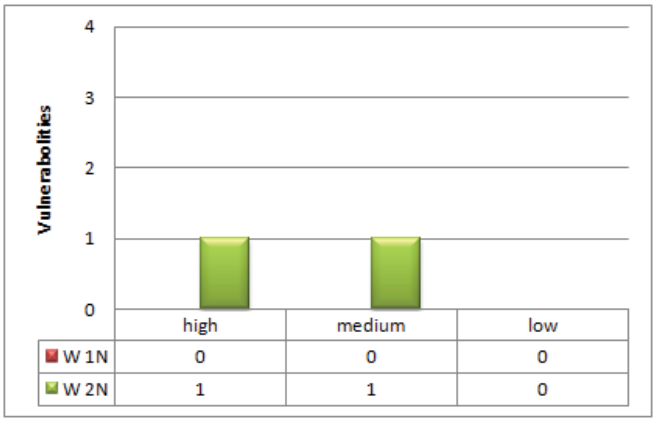


Fig. 7. Summary results of outside security assessment on Windows

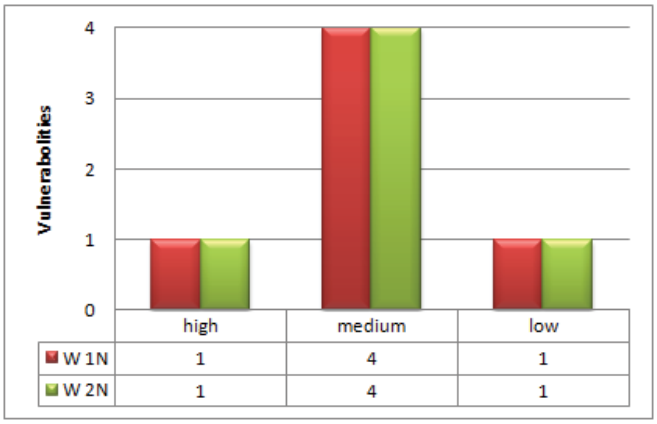


Fig. 6. Summary results of inside security assessment on Windows

machine instances, we present and analyze the result of the assessment of Windows 2008 R2 virtual machine instance. We omit the results of the assessments with CVSS score 0 since they are informative, rather than real vulnerabilities. Critical vulnerabilities are not detected during the assessments and are also omitted in the presented results.

A. Virtual Machine Instance Vulnerabilities from Inside

Fig. 6 depicts the summary results of the inside security assessment on Windows virtual machine instance hosted on OpenStack with two different networks deployment, i.e. test cases 1 and 3 for Windows virtual machine instance.

The same 1 high, 1 low and 4 medium vulnerabilities are detected in both test cases, which proves the hypothesis H2. Let's explain them in details and propose their improvement.

The high vulnerability "Vulnerabilities in Remote Desktop Could Allow Remote Code Execution" on TCP port 3389 allows execution of arbitrary code. This vulnerability can be resolved by patching the Windows.

The first medium vulnerability "Terminal Services Encryption Level is Medium or Low" on TCP port 3389 allows the attacker to easily eavesdrop the communication. This vulnerability can be resolved by changing the RDP encryption level to "High" or "FIPS Compliant". The second medium vulnerability "Microsoft Windows Remote Desktop Protocol Server Man-in-the-Middle Weakness" on port 3389 allows unauthorized access to the remote Windows. It can be resolved by reconfiguration of Windows to use SSL as a transport layer. The third medium vulnerability "Terminal Services Doesn't Use Network Level Authentication (NLA)" on port 3389 allows man-in-the-middle attacks and protects the remote computer from malicious users and software. It can be resolved by enabling Network Level Authentication (NLA) on the Remote Desktop Protocol (RDP) server. The last medium vulnerability "Server Message Block (SMB) Signing Disabled" on TCP port 445 allows man-in-the-middle attacks against the SMB server, and can be resolved by enforcing message signing in the Windows host configuration.

The low vulnerability "Terminal Services Encryption Level is not FIPS-140 Compliant" on port 3389 can be resolved by changing the RDP encryption level to "FIPS Compliant".

We can conclude that although many high, medium and low vulnerabilities are detected on Windows virtual machine instance, still their risk can be mitigated with reconfiguration and installing the latest patches.

B. Virtual Machine Instance Vulnerabilities from Outside

Fig. 7 depicts the summary results of the outside security assessment on Windows virtual machine instance hosted on OpenStack with two different networks deployment, i.e. test cases 2 and 4 for Windows virtual machine instances.

We have not detected any security vulnerability from the outside the cloud while deployed with only one network, but we detected 1 high and 1 medium vulnerabilities from the outside the cloud while deployed with two networks.

The high vulnerability "Vulnerabilities in Remote Desktop Could Allow Remote Code Execution" and the medium vulnerability "Terminal Services Doesn't Use Network Level Authentication (NLA)" on port 3389 are explained previously.

Although the results negate the hypothesis H3 for cloud tenants (Windows virtual machine instances), the same vulner-

abilities exist on the Windows virtual machine instance, since they are detected from inside for both network deployments.

VII. Conclusion and Future Work

We have realized security assessments of OpenStack cloud with two different network deployments. The experiments addressed the security vulnerabilities of OpenStack cloud services and four virtual machine instances with different operating systems CentOS, Fedora, Ubuntu and Windows, both from inside and outside the OpenStack cloud.

The results of the assessments proved hypotheses H1 and H2 that segregating the networks in OpenStack rises new security vulnerability risks only for CSP from inside, but remain the same for the tenants. The security vulnerability risks are also the same for the CSP outside from the OpenStack cloud, which partially proves the hypothesis H3. However, segregating the networks in OpenStack can increase the security vulnerability risks for the tenants using Windows virtual machine instances.

Vulnerabilities on Linux based operating systems are not detected for both network deployments, neither from outside, nor inside. But although Windows based virtual machine instances are detected with security vulnerability risks, they can be secured by implementing existing patches or reconfiguration.

OpenStack services are also vulnerable for both network deployments. 2 medium vulnerabilities on 6 different ports must be secured with new patches that should be developed.

We realized the security assessment of OpenStack with two different network deployment and virtual machine instances hosted on different operating systems. We will continue the security assessment on the other open source clouds with different deployment in order to help the customers to select not only the cloud solution with the highest level of security, but also the most appropriate deployment of each open source cloud solutions based on security.

References

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, pp. 50–58, Apr. 2010.

[2] Y. Tang, P. P. Lee, J. C. Lui, and R. Perlman, "Secure overlay cloud storage with access control and assured deletion," *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 6, pp. 903–916, 2012.

[3] L. M. Kaufman, "Can a trusted environment provide security?" *IEEE Security and Privacy*, vol. 8, no. 1, pp. 50–52, Jan. 2010. [Online]. Available: http://dx.doi.org/10.1109/MSP.2010.33

[4] T. Jaeger and J. Schiffman, "Outlook: Cloudy with a chance of security challenges and improvements," *IEEE Security and Privacy*, vol. 8, no. 1, pp. 77–80, Jan. 2010.

[5] R. Jhawar, V. Piuri, and P. Samarati, "Supporting security requirements for resource management in cloud computing," in *Computational Science and Engineering (CSE), 2012 IEEE 15th International Conference on*, dec. 2012, pp. 170 –177.

[6] M. Shtern, B. Simmons, M. Smit, and M. Litoiu, "An architecture for overlaying private clouds on public providers," in *8th Int. Conf. on Network and Service Management, CNSM 2012, Las Vegas, USA*, 2012.

[7] OpenStack. (2013, Jan.) Openstack cloud software. [Online]. Available: http://openstack.org

[8] B. Grobauer, T. Walloschek, and E. Stocker, "Understanding cloud computing vulnerabilities," *Security Privacy, IEEE*, vol. 9, no. 2, pp. 50 –57, march-april 2011.

[9] R. Shaikh and M. Sasikumar, "Article: Security issues in cloud computing: A survey," *International Journal of Computer Applications*, vol. 44, no. 19, pp. 4–10, April 2012, foundation of Computer Science, USA.

[10] L. M. Kaufman, "Can public-cloud security meet its unique challenges?" *IEEE Security and Privacy*, vol. 8, no. 4, pp. 55–57, Jul. 2010.

[11] W. Brown, V. Anderson, and Q. Tan, "Multitenancy - security risks and countermeasures," in *Network-Based Information Systems (NBiS), 2012 15th International Conference on*, sept. 2012, pp. 7 –13.

[12] X. Zhang, N. Wuwong, H. Li, and X. Zhang, "Information security risk management framework for the cloud computing environments," in *Proceedings of the 2010 10th IEEE International Conference on Computer and Information Technology*, ser. CIT '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 1328–1334.

[13] S. Ristov, M. Gusev, and M. Kostoska, "A new methodology for security evaluation in cloud computing," in *MIPRO, 2012 Proc. of the 35th Int. Convention, IEEE Conference Publications*, Opatija, Croatia, 2012, pp. 1808–1813.

[14] Z. Chen and J. Yoon, "It auditing to assure a secure cloud computing," in *Proceedings of the 2010 6th World Congress on Services*, ser. SERVICES '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 253–259.

[15] C. Wang, S. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for secure cloud storage," *Computers, IEEE Transactions on*, vol. 62, no. 2, pp. 362 –375, feb. 2013.

[16] S. Ristov, M. Gusev, and M. Kostoska, "Cloud computing security in business information systems," *International Journal of Network Security & Its Applications (IJNSA)*, vol. 4, no. 2, pp. 75–93, 2012.

[17] ISO/IEC, "ISO/IEC 27001:2005, Information Security Management Systems - Requirements," 2012. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=42103

[18] CloudStack. (2013, Jan.) Cloudstack opens source cloud computing. [Online]. Available: http://cloudstack.org

[19] S. Ristov, M. Gusev, and M. Kostoska, "Security assessment of openstack open source cloud solution," in *Proceedings of the 7th South East European Doctoral Student Conference (DSC2012)*, Thessaloniki, Greece, 2012, pp. 577–587.

[20] M. Gusev, S. Ristov, and A. Donevski, "Security vulnerabilities from inside and outside the cloud," University Ss Cyril and Methodius, Skopje, Macedonia, FCSE, Tech. Rep. IIT:01-13, Jan. 2013.

[21] S. Ristov, M. Gusev, and A. Donevski, "Openstack cloud security vulnerabilities from inside and outside," in *to be published in Proceedings of The 4th International Conference on Cloud Computing, GRIDs, and Virtualization (CLOUD COMPUTING 2013)*, Valencia, Spain, May 2013.

[22] KVM, "Kernel based virtual machine," Jan. 2013. [Online]. Available: http://www.linux-kvm.org/page/Main_Page

[23] C. Systems, "Xen hypervisor," Jan. 2013. [Online]. Available: http://www.xen.org/

[24] UML, "User-mode linux kernel," Jan. 2013. [Online]. Available: http://user-mode-linux.sourceforge.net/

[25] Microsoft, "Microsoft hyper-v server 2012," Jan. 2013. [Online]. Available: www.microsoft.com/hyper-v-server/

[26] Tenable, "Nessus 5," Jan. 2013. [Online]. Available: http://www.tenable.com/products/nessus

[27] N. V. Database, "Common vulnerability scoring system," Jan. 2013. [Online]. Available: http://nvd.nist.gov/cvss.cfm

Security analysis of Croatia's receipt registration and verification system

Stjepan Groš

University of Zagreb, Faculty of Electrical and Computing Engineering,
Unska 3, 10000 Zagreb, Croatia
E-Mail: stjepan.gros@fer.hr

***Abstract*—Beginning with 2013. a law in Croatia come into the force that requires owners of restaurants, café bars, and similar types of businesses that work with cash to register every receipt with a Tax Administration servers before issuing it to a customer. For the purpose of implementing the law APIS-IT, a Croatian IT company, developed a protocol based on XML, SOAP, and public key cryptography. They also implemented the server side system. It is a well known fact that developing protocols in general, and security protocols in particular, is a very tricky endeavor in which even the security professionals make mistakes. In this paper a security analysis of the protocol for receipt registration, the components of the system, and implementations is presented. Note that this is only a partial analysis, based on publicly available information, which doesn't include testings on live systems due to being illegal by the new Criminal law in Croatia. We identified two weaknesses of the current system. But the main problem of the system is the fact that many business owners are now open to different attacks and nothing has been done to remedy that situation. This is actually a broader problem since, with ever increasing number of on line services nothing is done to increase security awareness of people.**

***Index Terms*—security,xml,threats,finance,analysis**

I. Introduction

Beginning with 2013. a law [1] in Croatia come into the force that requires owners of restaurants, café bars, and other types of businesses that work directly with cash to register every receipt with the Tax Administration (TA) servers before issuing it to a customer. This is only a first phase of system deployment and during 2013. the rest of the businesses will also follow. For the purpose of implementing the law, APIS-IT, a Croatian IT company, developed a protocol based on XML, SOAP and public key cryptography, along with a server side system. It is a well known fact that developing protocols in general, and security protocols in particular, is a very tricky endeavor in which even the security professionals make mistakes. In this paper a security analysis of the protocol for receipt registration, the components of the system, and implementations is presented.

The paper is structured into 6 sections. First, in Section II we present the motivation for the introduction of the system for receipt registration. We also give an overview of its architecture and protocols used for communication. Finally, we get an overview of controls used to protect the system from deception. Then, in Section III we give a threat model, i.e. the model in which we are going to analyze the security of the system. In Section IV we enumerate different attacks against the system that allow different threat agents to circumvent controls and gain some advantage. This analysis is done under the assumption of the perfect implementation, i.e. no programming or configuration errors in the implementation. But, in Section V we review some problems that might arise in case there are errors and omissions in implementation or deployment. Finally, the paper finishes with the Section VI in which we give conclusions, recommendations for improvements and outline the further work to be done.

II. The system under the security analysis

The system under the security analysis, i.e. the *Fiscal system*, was developed to prevent fraud that was based on the manipulation with receipts done by fraudulent businesses. In essence, receipts weren't issued, or if issued they were easily erased from the cash registers since they've only been stored locally. There were a control mechanisms in place to prevent such scenario, i.e. the customers could send receipt numbers to the TA which would verify that those receipts are really reported, and also TA could do inspections of both businesses and customers. Namely, customers were required by the law to take receipts. But this control mechanisms were inefficient, i.e. customers didn't report receipts, even if they did TA had problems verifying reported receipts because of diverse ways of enumerating them and manual work involved, and inspections are inefficient and costly. Thus, the Fiscal system was introduced with the following goals:

- Having each receipt stored on TA's servers at the moment receipt is issued, in order to prevent manipulation, i.e. to assure non repudation, integrity and authenticity.
- Unifying receipt numbering schemes in order to make receipt checking easier and possible to automate.

Note that inspections by TA are still necessary in order to enforce receipts being issued. Furthermore, because of resource constraints in terms of a number of available inspectors, the customers are also encouraged to send receipt IDs to TA. This is further incentivized with different prizes that customers can win.

Based on these requirements, the architecture was developed along with appropriate protocols.

In essence, the architecture of the fiscal system is relatively simple. It consists of two entities, a client machine which submits receipts and TA's server which accepts receipts, records them and issues signed response (Figure 1).

The messages are exchanged using SOAP over HTTPS, while the data is serialized in XML format. To ensure authenticity, integrity and non-repudiation of the messages public

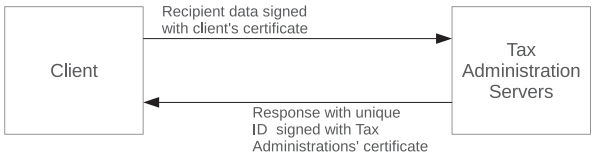


Fig. 1. The architecture of the fiscal system

key cryptography is used. Each entity in this system is issued a certificate by Financial Agency (FINA) [2] which acts as Government's Certificate Authority. The messages are signed before being sent and this signature is generated according to the XML Security specification. To ensure confidentiality (and integrity) of messages while being in transit, SSL/TLS is used. In total, six messages are defined:

- Receipt request/response pair of messages used to register receipt with TA.
- Request/response pair used to register office space.
- Simple echo request/response messages used to test TA's server availability.

Office space registration is used, presumably, for two purposes. First, for TA to know where receipts are issued, and secondly, to restrict the use of a single receipt on different locations.

But, of those, we are interested only in the first pair of messages used to register receipt with TA. So, each generated receipt has to be sent to TA prior to being issued to a customer, to obtain *uniqe receipt ID*. In this paper we will use abbreviation *JIR* for this unique receipt ID as it is so designated in appropriate documentation. An example of a JIR is:

```
6b7749c6-56c1-4cf5-b7f7-9f29cebc9f7f
```

Which is basically standard UUID format. Along with JIR, each receipt also has an *issuer's protection code*, we'll use abbreviation *ZKI* from now on as it is used in the Croatian documents. This code is generated by concatenating OIB, timestamp, receipt ID, office space ID, cash register ID, and total sum, then signing all that and finally taking MD5 of the signature data. Note that ZKI is generated by cash register without any data or involvement from TA' servers. An example of an ZKI is:

```
a1e6b1428f0cc755f0c82aa7a1327e35
```

The motivation for introducing this code is to protect issuer from a malicious third party [3] and also in cases when the issuer doesn't have connectivity to TA, then he can use this code instead of JIR.

There are two controls that force business owner to issue proper receipt:

1) A customer that checks receipt data on TA.
2) Direct controls by TA's inspectors.

Customers are required by law to take receipt with them. Additionally, they can also check JIRs by submitting them on TA's Web pages (or via SMS messages). In that way business owners are forced to issue receipt and they are prevented from falsifying JIRs. Additionally, inspectors serve as control of both business owners (which are required by the law to issue receipt) and customers (which are also required by the law to take receipt). But, the reality is that there are not a lot of inspectors, their knowledge and equipment is rather poor. Customers, on the other hand, don't usually take receipt. This makes this control mechanisms rather weak.

Customers can verify receipt using either SMS message or a Web application [4] developed specially for that purpose. In both cases, the customer sends the following information:

- Either JIR or ZKI.
- Date and time when receipt was issued.
- Total amount.

If the entered data is correct, i.e. TA has data about the receipt, the Web application will only give information that receipt is correct. If the receipt can not be found in the TA's database the system will give to a customer a message to try again later. Note that due to the length of JIR and ZKI this control is weakened since retyping those codes is very hard, especially with SMS messages. Thus, errors are very likely and this further incentivizes customers *not to* verify receipts.

Note one interesting question. ZKI looks harder to read, but is shorter. JIR, on the other hand, is easier to read, but longer to type. So, it is reasonably to believe that customers will type ZKI more frequently, even though they are the same in term of the numbers and letters and the only difference is in dashes.

III. Threat model

As a first step we have to establish a threat model, i.e. a model that defines what are the threat sources and associated threats that can impact a system. Threat sources we analyze are:

- Business owners trying to circumvent paying taxes by manipulating receipts as they used to do.
- Individuals and crime organizations trying to take financial advantage of the system.
- Political groups and movements, like Anonymous, trying to subvert a system for PR purposes.

Business owners have incentive to manipulate system in order to avoid paying taxes. The ultimate goal is not to register smaller sum of money than there actually was. This can be done in two ways, (i) by not issuing receipt, or (ii) by somehow manipulating receipts themselves. Since not issuing receipt has nothing to do with the security analysis we'll ignore that scenario so we are only interested how business owner can manipulate receipts. We also ignore non-technical manipulations, like manipulations with fake receipts (quotes). We also assume that business owners have a complete control of cash register, both hardware, network connectivity, and software.

One thing to note is that business owners can be different in size, what will have impact on their network topology and controls they implement. Here, we'll assume small business owners, i.e. owners of restaurants, café bars and similar types of businesses.

It is the fact that cash register now has to be connected to the Internet (it has to be in order to be able to communicate with

Tax Administration's servers). Some of those will be directly accessible on the Internet (by having public IP address). Many will be run by Windows operating system, and thus will be threatened by usual threats, like mallware. Last, but not least, probably business owners will offer their customers to access the Internet over the same LAN they use for communication with Tax Administration. This connectivity increases number of potential attackers, which we grouped into two categories, according to their motives for the attacks.

The first category consists of *individuals and crime organizations*. The difference is in the resources available to mount an attack. Obviously, individuals have fewer resources, while crime organizations have much larger resources on their disposal. So, individuals can try to attack at most few cash registers, while crime organizations can attack as much as they want.

The second group are political groups that aren't so much interested in financial gain but are satisfied with PR like events, like DoS attacks.

In the majority of the following text we'll assume *perfect implementation*, that is, will assume that programmers, both on server and client side, that implemented the protocol didn't make any programming mistakes. This is obviously very unrealistic assumption, especially from the point of business owners. At the end we'll outline possible mistakes that programmes could make and what this means in terms of attack possibilities. Still, this is a topic for a separate research.

IV. Attacks

A. Attacks by business owner

As we saw in Section II there are two primary controls that check invoices issued by business owners, customers and inspectors. Customers can not access cash register in order to check if invoice is handled properly or not, while TA's inspectors can to some degree.

So, here are some possibilities what business owners can do in order *not to* register valid receipt with the TA while in the same time avoiding being caught by the customers reporting invoice data to TA:

- Business owner can generate receipt only with ZKI. Then, he observers if the customer takes receipt or not. If he takes it then the Business owner registers receipt with TA, otherwise, he removes receipt from the system.
  Note also that ZKI can be a random number, there is no way for customer, or TA, to check if it is valid or not. That means that if TA receives receipt with only ZKI it *can not prove* that this receipt was issued by certain business owner. This relates to the verification of receipt via Web or SMS. In that case many businesses can coordinate to issue receipts with the same ZKI, amount, date and time. Alternatively, they can reuse one's receipt.
  Note that it is relatively easy to force customers to use ZKI instead of JIR. All they have to do is to print some random JIR. If the customer types JIR, the error will be reported and finally, if the customer is persistent enough he will finally try with ZKI.
- The way receipt checking with TA is implemented opens up possibility of reusing JIRs. Business owner can issue one receipt that will be registered with TA's servers. Then JIR and ZKI can be reused for other receipts with the same amount within a certain time frame.
- Since either JIR or ZKI is typed when the receipt is verified, this means that they don't have to be correlated. In other words, you can write JIR and ZKI belonging to a different receipts. But, for the time being it is questionable if this opens up some new possibilities to avoid registering invoice with TA.

Note that if issuing multiple receipts with the same JIR/ZKI then care must be taken about invoice numbers which are required to be sequential.

Now, second control, i.e. TA's inspectors, complicates things a bit, but not much. The first avoidance tactic in the previous list is the most dangerous. Namely, it rests on the fact that two, or more, persons don't collude, that is, they don't compare their invoices. TA's inspectors might do that, that is, they might collect somehow few invoices and if any two have the same JIR or ZKI, obviously the business owner has tried to circumvent the system.

The key difference between customers and TA's inspectors is that the letter can inspect data on cash register to some degree. What could happen is:

- Inspector comes with a bunch of invoices and asks business owner to show them records in cash register. If there are multiple invoices with the same JIR/ZKI/number then the business owner must be very careful which one he will show to inspectors. But, in the end, inspectors have no way to detect fraud *unless* they have at least two invoices with the same JIR/ZKI/number.
- For receipts that reuse JIR/ZKI numbers inspectors might notice that there is some discrepancy between dates on receipts and the actual time they were issued. But, there is no requirement on time synchronization on cash registers.
- The specification of the fiscal system defines that tax administration might ask business owner to recreate (or restore) ZKI. But it is very hard to verify that ZKI was generated using right data and, more importantly, the right certificate.

To conclude, without TA's inspections business owners have opportunities for manipulations even though they are not as easy as they were before. To which extent this manipulations can be done is a topic of a further research, especially if business owners collude. Additional problem is that TA's inspectors have to be properly equipped in order to be able to detect fraud. Finally, some fraud can be detected via advanced techniques like data mining on TA's servers.

B. Attacks by malicious individuals

Malicious individuals have enough resources to attack one, or at most several cash registers (CRs). Looking again on the Figure 1 individuals are a significant threat to business owners, and a lot less to TA's servers. Some of their motives can be the following ones:

- To harm competition by manipulating competition's receipts and in that way incur penalties by TA.
- To reuse invoices by competition and in that way no to register their invoices.
- To monitor what competition has been doing.
- Individuals that just want to do it because they can, to show off themselves.

In order to attack business owner and their CRs individuals have to be either:

- Somewhere on the communication channel, i.e. trying to perform MITM type of attacks.
- Trying to compromise cash register machine itself with a final goal of obtaining the secret key belonging to the business owner.
- Denial of service attack on Internet connection and/or cash register machine.

The communication channel is protected using HTTPS, so in case there are no bugs in the implementations it is very hard for the attacker to break into that channel. Note that two controls are of utmost importance here. The first one is that communication is performed via HTTPS (i.e. no fallback to HTTP) and second that the certificate check is properly implemented. The only way attacker can break into the channel is by exploiting bugs in the implementation. Since we assumed that there are no bugs, than this isn't an issue. Still, we'll return back to this possibility in Section V.

The second attack vector is by exploiting the client machine, i.e. cash register, itself. The grand prize in this case is stealing private key. The moment the attacker has control of the machine he is basically in a possession of private key. There could be protection in form of a password on a file but this password has to be stored somewhere. The ways in which an attacker can achieve that are usual attack vectors, of which the most dangerous one is social engineering attack. What makes things even worse here is that cash registers were old machines with old hardware and not maintained well. Now, when those machines are on the Internet, with employees that will probably use them to surf the Web which opens up a lot of possibilities to attack the machine itself. This is, in a way, *game over* situation in which attacker can do whatever he/she likes.

The third attack can prevent business owner from operating and can degrade him into the mode where it only issues ZKI, not JIR. The additional problem with this type of attack is that it manifests itself indirectly (e.g. error message failed to establish connection) and business owner can not know that he is actually under the DoS attack.

To conclude, malicious individuals are a very serious threat and to protect themselves from this attacks business owners have to have properly maintained machines as well trained stuff. Unfortunately, in a situation in which business owners, for whatever reason, try to save as much as possible it is highly likely they will not invest in security. Additionaly, they don't have a habit of maintaining machines. What TA has to do, it has to educate business owners and warn them about threats they face.

C. Crime organizations

Crime organizations have enough resources to attack potentially many business owners and their primary goal is a financial gain. Basically, what they can do is attack many business owners simultaneously and for that purpose they have same attack vectors at their disposal as individual attackers. But, additionally, they have one more attack vector and that is malicious software specifically tailored for this purpose. The problem with such approach, i.e. tailored malicious software, is that standard antivirus software is lot less efficient and this poses a significant problem.

Additionally, it is possible to attack TA's servers via some innocent clients and in that way to hide real attacker. This is viable threat under the assumption that there is some implementation error in TA's server, as we further discuss in the Section V.

D. Activist groups

Activist groups, like Anonymous, are attacking for the primary purpose of publicity, not financial gains. Thus, attacking a single business owner doesn't give them sufficient pay offs. It is more likely they will try to attack TAs servers with the primary goal of causing disruptions to the service.

This is actually very real threat because it is relatively cheap to by services from some bot owner and do prolonged DDoS attacks on the TA's servers themselves.

V. Implementation based problems

In the previous analyses we assumed that the implementations are perfect, both on the client side as well as on the server side, i.e. TA's servers. Unfortunately, it is very likely that there will be many errors, especially on the client side as there are a number of different implementations done by people that don't usually do security related protocols. In this section we are going to list some of the potential problems that might arise from different mistakes done by implementors, but in general this is open to further research.

The first and foremost question is whether clients require https connection to server and do they correctly check server's certificate. If either of those isn't fulfilled, then MITM attack is possible. At minimum, an attacker that successfully intercepts communication channel can see messages between client and server and in that way confidentiality of the business owner is violated. The next thing the attacker can achieve is to disrupt communication, either by not forwarding messages or by modifying them. Note that we assume that modifications will be detected by server or client, depending who's receiving them.

There is also question about covert channels which are very dangerous as they allow information leak. In the extreme, it would allow someone to deduce TA's secret key which would be the ultimate attack on this whole system. Note that attacks on TA's servers are supposedly protected by having each business owner sign an agreement with TA/FINA. But, this doesn't preclude some attacker from compromising business owner's machine and performing attacks using it as a proxy.

Many other potential problems here exists, like proper XML processing, schema validation, etc. To conclude, there are many potential dangers here and this is a topic for further research.

VI. Conclusions and future work

In this paper we presented initial analysis of a fiscal system introduced by Republic of Croatia's Tax Administration. It is only a partial analysis, based on publicly available information, which doesn't include testings on live systems due to being illegal by the new Criminal law in Croatia. Also, we assumed that the implementations are perfect, i.e. there are no software bugs.

The finding is that on the server side the system is relatively well protected. On the client side, the things are quite different. This is Achilles' heel of the whole system. We think that by introducing this whole fiscal system, without warning and proper education of business owners many of them are brought into a dangerous situation. It is the truth that it is the consequence of irresponsible behavior of individuals, which don't care about security, but it is also a Government's irresponsible behavior that neglects the other side. So, some actions have to be taken here.

There is also a weaknesses in the control part of the system are, the one that forces business owners to issue receipts, and as usual, the weakness is on the humans. Namely, it is questionable how much customers will be willing to type JIR or ZKI to check the receipt. Furthermore, it is mandatory that TA's inspectors be properly educated and equipped.

During the threat analysis we identified two technical weaknesses. First, there is possibility to reuse JIR's by manipulation the date and sum on receipts. That one is very hard to prevent without rigorous controls, which are infeasible. The second weakness is ZKI, which by itself is useless. There is no way for TA's inspectors to *prove* it has a correct without having private key, which by definition they are not allowed to have!

Fortunately, this systems uses a lot of advanced technology, primarily cryptography, which many people don't know much about, and this for the time being acts as a barrier for more frauds.

As for the future work, we think that the main research should be done on live systems to identify weaknesses they have. More specifically, it is our intention to build a proxy that would automatically test certain implementation for correctness. Next, a motive for attackers is not elaborate enough, but this requires financial along with technical knowledge. Finally, we think that the introduction of advanced statistical analyses on TAs servers might help in detecting some manipulations described here, i.e. in order to spot strange behavior in the registration process. But, this is a topic for a further research.

References

[1] "Odluka o proglašenju Zakona o fiskalizaciji u prometu gotovinom (in Croatian)," Nov. 2012. [Online]. Available: http://narodne-novine.nn.hr/clanci/sluzbeni/2012_12_133_2822.html

[2] "Financial agency," last checked: February 6th, 2013. [Online]. Available: http://www.fina.hr/Default.aspx?sec=1134

[3] "Fiskalizacija - Tehnička specifikacija za korisnike - Verzija 1.2 (in Croatian)," Dec. 2012. [Online]. Available: http://www.porezna-uprava.hr/fiskalizacija/dokumenti/Fiskalizacija%20-%20Tehnicka%20specifikacija%20za%20korisnike_v1.2.pdf

[4] "Web application for invoice checking," last checked: February 6th, 2013. [Online]. Available: http://www.provjeri-racun.hr/provjeraracuna

Conceptual Information Modelling within the Contemporary Information Security Policies

Aleksandar Klaić*, Marin Golub **
* Office of the National Security Council, Zagreb, Croatia
** Faculty of Electrical Engineering and Computing, Zagreb, Croatia
aleksandar.klaic@public.carnet.hr; marin.golub@fer.hr

Abstract – The contemporary information security policies are analysed in the paper. These policies are characterised by increased similarities among the information security requirements of different sectors of society. Also, they are characterised by the increased differences in comparison with the traditional approach to the security within the closed environment. Key factors of the information security policy: people, process, and technology, are closely related to the requirements and restrictions imposed on certain type of information. In that way the approach to conceptual information modelling becomes one of the central problems of contemporary information security policies. The paper elaborates the approach to the conceptual information modelling, stressing the requirements of both the protection and the sharing of information. It offers the taxonomy of the main terms, which is the base for the development of proposed conceptual model of the information definition and sharing. The conceptual model is based on the standard UML graphical notation that makes it easier to visualize and understand the proposed model and the approach applied in the paper. The proposed model introduces formalized and more structured approach to this field in order to facilitate the development of the solutions that can keep up with the growing complexity of contemporary information security policies.

I. INTRODUCTION

The contemporary information security policies that are used in both the government and private sectors are analysed in the paper. These policies are characterised by increased similarities among the information security requirements of different sectors of society. There are also considerable differences of today's approach to information security in comparison with the traditional approach within the closed environments. Key factors of the information security policy: people, process, and technology, are closely related to the requirements and the restrictions imposed on certain type of information.

Information becomes very important factor within today's cyberspace. The term cyberspace is defined as virtual global environment of mutually connected public and private information systems, in which information, including specific ones that are dominant in the view of information security requirements, are created and transmitted [1]. This development of technology and society makes the problem of the approach to different types of information to become central issue of contemporary policies. Traditionally, information is the key asset that is protected within the information security policies of the government sector [2] [3]. On the other hand, during the last two decades the approach to information security has rapidly developed, and that fact initiated the parallel process of international standardisation. The risk management methods are increasingly applied and that fact has extended the focus of the policies from information and information assets to the wider range of tangible and intangible assets. All these assets have significant values to their owners [4].

The whole range of processes that have security influence on both the government bodies and private companies were caused by the development of society and technology, and followed by the appearance of new threats and vulnerabilities. First it was the development of technology and as a consequence we have today's business processes dependency on information systems and Internet. Actually, it is the dependency on the elements of the cyberspace. Besides that, since the 1990s there were a lot of national processes of liberalization of some sectors such as telecommunications, energy, and transport. That had and still has the influence on national security [5]. All these processes have led to the changes of people's professional and private lives, deeply influencing them through the social networking sites, different internet services, and the new mobile gadget market. The result is the constant and persistent exchange of information within all parts of our professional and private lives [6]. The similar process of constant and persistent exchange of different information including confidential ones is happening on the level of legal entities. Good example is the area of national critical infrastructure, or European Union (EU) critical infrastructure protection program [7].

Closed information systems (isolated, air-gap) are decreasingly applicable even in the traditionally closed environments of government classified information systems. Such changes in the environment necessarily demand changes in the information security policy approach on the two levels. First level is the necessity of adjustment of the new approach to the policies of information system security in order to adapt them to open commercial information and communication resources [3]. Second level is even more sensitive because the approach has to be adjusted in the wider sense of the information security policy as the whole. It is the problem of categorizing different types of information within the contemporary environment based on their security criteria. This is much wider coverage of information comparing to the classified information domain, or trade secret domain for example.

Communication needs today extend the scope of previous information security policies both in the industrial and in government sectors. This includes for example the need to exchange certain information among the entities from different sectors and the demands for handling specific types of information like personal data or intellectual property.

Cyberspace represents significantly changed environment both in the sense of form, amount, and types of information, and in the sense of different and more complex regulatory requirements. Further on, there is the necessity of international cooperation and global information exchange that leads to added complexity both in the approach and in the content of contemporary information security policies. Such growing complexity looks for the new approaches to facilitate practical solutions in this field. So far, the development of information security policies has been mostly based on the best practises and standardisation processes. Such approach has offered adequate solutions within the organisation's local environment [8].

The paper presents the overview of some results of our research in the field of modelling the briefly introduced domain of contemporary information security policies. The focus of interest in the paper is primarily in the information definition and information sharing issues.

II. Conceptual Information Modelling within the Contemporary Information Security Policies

The paper elaborates the approach to conceptual modelling of information within the frameworks of contemporary information security policies, stressing not only the need for protection of information, but also the need for information sharing ("need-to-know" versus "responsibility-to-share"). Conceptual information model is elaborated within the scope of protected national and business assets and through the development of approach to information sharing. In this paper we present the conceptualization of information types and information sharing approaches based on the domain knowledge.

Conceptualization of the domain such as the contemporary information security policies makes it possible to better associate existing knowledge which is available in different forms. Due to these different forms existing knowledge has relatively weak relations among subdomains from the point of view of information security policy. Different forms of existing knowledge are for example knowledge data bases (e.g. threats and vulnerabilities), or procedural knowledge known from the best practises and comprised within some information security standards [4] [28]. According to [25], in order to make specification of conceptualization (the development of the domain ontology), three types of knowledge have to be mapped: declarative knowledge (Know-about Knowledge), procedural knowledge (Know-how Knowledge), and relational knowledge (Know-with Knowledge). Declarative knowledge is represented with the taxonomy terminology, actually the selection of concepts. Procedural knowledge is the description of meaning of such concepts, and finally, relational knowledge is represented by the relations of modelling concepts. Recognizing and mapping of the concepts with the goal to develop ontology in the field of information security policy is the research subject of several research projects [8]. So far, these research projects have been mostly focused on explicit knowledge expressed within certain information security standards [26]. In this paper authors are primarily focused on the conceptualization of implicit and tacit knowledge, contained in different policy frameworks, requirements, and standards for the development of contemporary information security policy.

A. The Issue of the Approach to Different Types of Information

The approach to different types of information within information security policies mostly has characteristics of government or business sector that certain organisation belongs to. The approach is normally adapted to the exchange of the similar type of information with the similar type of organisations. In this approach closed within the sectors there are predefined information categories essential for such approach (e.g. classified information). The sets of minimal protection measures are applied on these predefined information categories. With the approach on the lower level of individual organisational entity (legal entity), the value of certain internal sets of information is assessed. Certain security controls are applied on these sets based on the locally assessed risks. In both cases information sharing is based on the fact that the comparable type of information can be recognized on both sides that want to share information. Besides that, the risk of accepting the same protection measures for their own information, or received information has to be acceptable (reciprocity). This approach is mostly satisfactory within the narrow framework of certain government or business sector.

If we want to extend the described possibilities to cover the approach outside the framework of certain sector, the key problem is incompatibility of locally applied information security policies concerning the access to information („need-to-know"), protection of information (baseline protection or risk management), and the information sharing approach („responsibility-to-share"). Traditional approach to the information security in government sector is focused mostly on the categories of information according to their confidentiality criterion (classified information). On the other hand, regulation requirements are increasingly focused on different types of information (e.g. personal data, intellectual property) that are exposed to the different threats in cyberspace. Also, the information sharing requirements increasingly become the part of regulation demands, e.g. critical infrastructure protection. Furthermore, the increased usage of commercial communication and information resources, lead to the appearance of a range of new types of information that may become security problem to any organization [6].

Within the last few years several governments have initiated the analysis of these issues within their information security policies. The common tendency among these initiatives can be noticed. It is the separation of the domain of unclassified, sensitive, and officially used information from the domain of classified

information. On the one hand the reason for that is the requirement of transparency of work of the government sector. This means that the categories of sensitive information such as "For Official Use Only", "Controlled Unclassified Information", or the introduction of Limited Dissemination Markings, are used in many cases instead of previously used classified information (secret information). On the other hand, these are some of the mechanisms (new categorisation, limitation of distribution etc.) that should facilitate today's information sharing approaches, especially among entities from different sectors. The examples of described initiatives can be seen in a number of countries all over the world, e.g. France [9], Australia [10], or U.S.A. [11]. EU has also this type of unclassified official information marked „Limite" (sensitive but not secret) [12]. For the protection of this type of information EU uses the concept of professional secret [13]. The importance of these initiatives can be seen from the statistical information given by EU that 75% of all the information used within the Council of EU during the year 2010 was marked as „Limite" [12].

Big international organizations such as EU and NATO need to have consistent approach to the information security policy due to their complex organisation. Consistency can be hard to accomplish without certain formalization of the approach and without structured elaboration of the policy requirements and concepts. Good example of such approach is visible in the European Commission plan regarding the new concept of personal data protection in the EU [14]. This new EU regulation will be introduced by 2015 and the key novelty is the unification of the concepts regarding personal data protection across the EU member countries. Primary reason for the change of the approach is to improve consistency of the implemented measures for personal data protection of the EU institutions, EU member countries, and the legal entities registered in those countries. Cooperation of different international, governmental, non-governmental and legal entities within the complex environment of international peace keeping activities is another example of the necessity of certain level of formalization and more structured elaboration of requirements and concepts related to the different types of information. The reason is that all these different participants need to access certain type of classified and sensitive information in the field that in some cases are not recognized by their information security policies.

Conceptual information modelling today is also used in some e-Government projects primarily due to the requirements of interoperability of information systems [15]. In accordance with their purpose the focus of these projects is mostly on the certain architectural view of the system and on the certain specific project goals [16] [17] [18]. There are also researches related to the possibilities of using conceptualization of the decision making process [19], which is partly related to the problems of regulation compliance within the domain of information security policy.

B. Modelling of the Contemporary Information Security Policies

The described security challenges lead to the need of introducing formal methods and structured elaboration of requirements and concepts within more and more complex domain of information security policy. According to [20] the complexity of the security environment is the reason that the issues of the policy should be looked at within much wider context than the certain organisational entity is. Also, in [20] it is concluded that the traditional analytical role of engineers in today's circumstances of fast changes of the security environment should be extended in order to allow engineers to participate in the regulation role (regulation and standardisation, national and international level). That is the only way how to create conditions necessary to set up security relationships in the complex global and local environment according to Fig.1 [20].

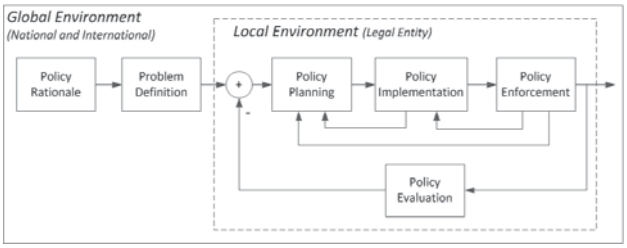


Figure 1. Lifetime phases of the information security policy

During the development process we have transformed the scheme from Fig. 1 according to the model shown in Fig. 2. This model in Fig. 2 represents the domain of contemporary security policies and it gives us the different view on the global and local levels of the policy shown in Fig.1. Fig. 2 shows very important part of the model were the global and local environment are confronted. For the purpose of this paper the term "environment" includes all the elements that are controlled through the information security policy of an organizational system, and the elements that influence on that system and its policy. Model is conceptualized as the set of mutually connected subsystems that describe certain parts of the domain of the information security policy. Conceptual modelling of information is shown through the description of the subsystems tightly related to the information issues. Primarily, it is the information definition within the interface part between the global and local environment of the model, and the information sharing part within the global environment part of the model in Fig. 2.

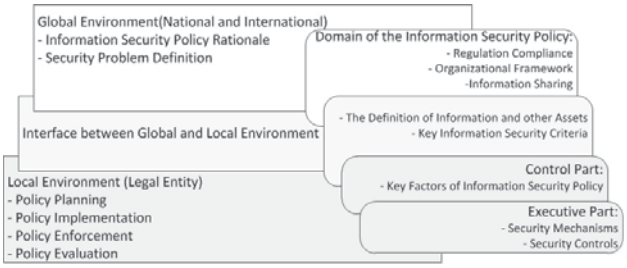


Figure 2. Modelling of the Contemporary Information Security Policies

C. Development of the Taxonomy and the Conceptualisation of the Terms

The hierarchical taxonomy of the terms is defined according to the domain of information security policy defined in [8]. Contemporary requirements and restrictions of handling different kind of information types are also taken into account, both from the point of view of government and private sectors. Taxonomy is used as a classification schema, which structures the knowledge within the domain and creates the hierarchical set of key terms. Further elaboration of taxonomy terms is based on theoretical analysis of the logical terms within the domain, as well as on the available best practises within different sectors and fields of application.

The analysis encompasses the information domains dominant from the point of view of information security (classified information, unclassified information, personal data, and intellectual property). It also encompasses societal and business sectors with specific requirements in the field of information security (e.g. government sector, telecommunication providers, and critical infrastructure protection). Different approaches to the information security policy in government and private sector are also taken into account [2]. The proposed taxonomy is elaborated with a view to future extensions, especially within the executive part of the model from the Fig 2. In that way such taxonomy is a good base for the definition of common terminology necessary in such a complex and heterogeneous field of information security policy.

The requirements imposed on the selection of categories and terms within the proposed taxonomy are used according to [21]:

1. Mutually exclusive categories that do not overlap;
2. Exhaustive categories including all possibilities;
3. Unambiguous and clear categories;
4. Repeatability;
5. Logical and intuitive acceptability;
6. Usefulness for the field of interest.

The structure of the taxonomy is elaborated following these requirements and it is divided into subsystems shown on the right-hand side of the Fig. 2. This paper is limited to the overview of the research results within the part of the taxonomy that deals with information conceptualisation. Hierarchical taxonomy is transformed into tabular view of the concepts and subconcepts in order to facilitate the elaboration of taxonomy terms into model concepts. Model concepts have to be recognized both as the categorization and as the mutual relationships of the domain terms. In this way the first part of the recognition of the basic relations among the domain terms is done. These are the groups of the relations of the type such as "is-a" (generalization), "consists-of" (composition), and "contains" (aggregation). The sample of our taxonomy is shown in tabular view in Table I.

TABLE I. THE ELABORATION OF THE HIERARCHICAL TAXONOMY, SUBSYSTEMS AND CONCEPTS/SUBCONCEPTS VIEW (SAMPLE)

DEFINITION OF INFORMATION AND OTHER ASSETS: Intangible Assets, ...	
Intangible Assets:	Information, Personnel, Software Support, Services, Intellectual Property, ...
Information:	Publicly Available information, Personal Data, Internal Sensitive Information, Confidential Information, Strictly Confidential Information, Classified Information
Personnel:	Qualification, Skills, Experience
Software Support:	Operating Systems, Software Applications, Software Support
Services:	Communication and Information Services, Other Services
Intellectual Property:	Copyright, Industrial Property Rights
Industrial Property Rights:	Patent, Trademark, Trade Secret, ...

The next step in the development process is the further elaboration of the specification of hierarchical taxonomy from Table I. into ontological model that is used for modelling information security policies. As it is already mentioned in the part of introduction of section II of the paper, ontology is treated as the explicit specification of conceptualization and represents the domain knowledge in formal and structured form [22].

One of the problems that have to be solved throughout the conceptualization of the model is the use of appropriate tools [20]. Considering the complex and very heterogeneous domain of contemporary information security policies we propose the use of standard graphical notation of Unified Modelling Language (UML) [23]. Similar approach is recommended in [24], but with the difference of using modified UML elements. UML comply with the ontology requirements in the sense of class definition and relation notation. UML graphical notation facilitates visualization and understanding of the model and the modelling approach in this paper.

D. Definitions of Information and Other Assets

Definitions of information and other assets is the interface between the upper and the lower part of our model from Fig. 2. This is one of the model subsystems that consist of the definition of general organizational assets from the business point of view. Assets are represented through the key information criteria of confidentiality, integrity, and availability.

In the model from Fig. 2 it is necessary to look at this subsystem from the global environment point of view of the requirements that are imposed on certain type of information (e.g. regulation compliance). The model also has to represent the requirements imposed on the key factors of the policy (people, process, and technology). These requirements are imposed on them because the need to handle protected assets. Key factors of the policy are treated through the control part of the model which describes local environment (lower right-hand part of the model on Fig. 2).

Modelling of the concepts in UML is done through the creation of the subsystem metaconcepts which are part of the metamodel of contemporary information security policies (Fig. 3). Metamodel allows the creation of different kind of policies that can be applied to particular organizational entities.

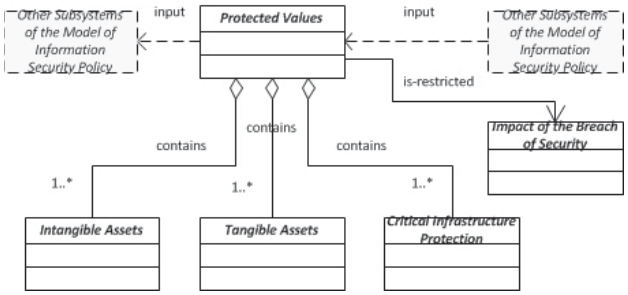


Figure 3. Basic concepts of the subsystem "Definitions of information and other assets" shown in UML class diagram

Information security criteria are generally used for the selection of security controls (assets protection in general) or security mechanisms (classified information protection). Categorisation of information is primarily done according to confidentiality levels (Fig. 4), while other two criteria (integrity, availability) are applied more on the information infrastructure and services that depends on operational needs of organizational entity (local environment). The role of confidentiality (secrecy and privacy) is important in the model as a whole because it depends not only on organizational needs, but also on global requirements (e.g. regulation compliance or information sharing).

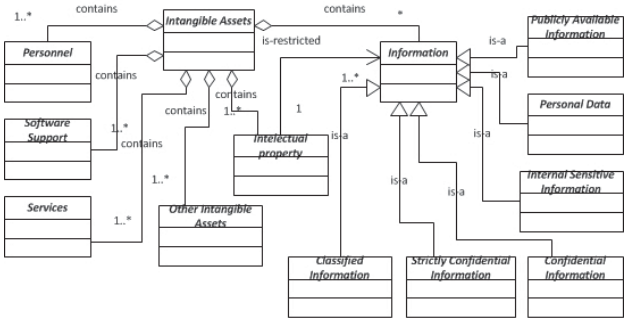


Figure 4. Elaboration of the concept of intangible assets in UML class diagram

The elaboration of the concept "Information" from Fig. 4 incorporates important category of publicly available information (key criteria are integrity and availability) which are further elaborated from the other point of view in the information sharing subsystem (Fig. 6). In Fig. 4 we have introduced the categories such as personal data, classified information, and also internal sensitive information explained in the section II.A of the paper. We encompass in this category 3-level system of classification of sensitive information proposed by [27]. This way of classification is increasingly used throughout private sector and more recently also in government sector (justice, finances, etc.).

E. Information Sharring

The subsystem of information sharing is comprised of basic concepts such as sharing subjects (entities from different sectors), sharing organization, and public information sharing. Sharing subjects are related with the concept of sharing organization according to Fig. 5. Sharing organization implies the organization of subjects in the communities of interest in such a way that there is internal community coordination (domains/users, information meaning, protocols for access) fully aligned with the external community requirements for interoperability rules (organizational, semantic, and technical rules).

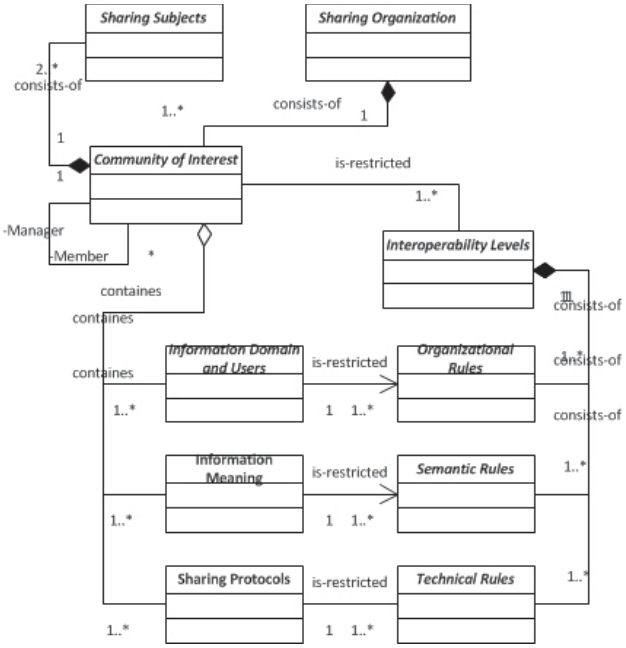


Figure 5. Elaboration of the concept of information sharing organisation in UML class diagram

Concept of public information sharing is one of the key concepts in contemporary information security policies. It is developed and shown in Fig. 6.

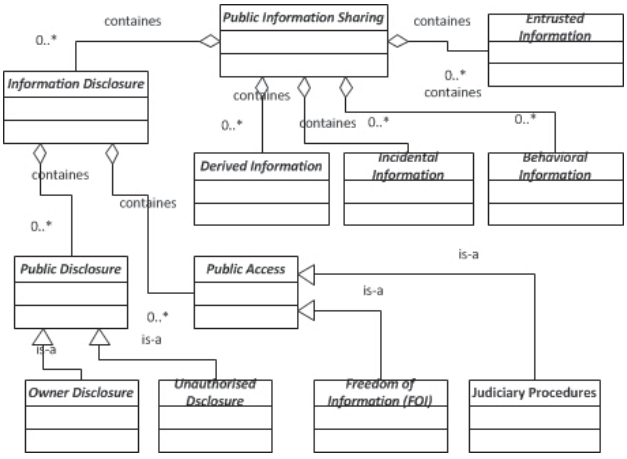


Figure 6. Elaboration of the concept of public information sharing in UML class diagram

Public information sharing concept in Fig. 6 uses several subconcepts according to [6] such as entrusted information. Entrusted information is information which is not under control of its owner because the owner entrusted it to other subjects. There are also behavioural information that is comprised of information collected by the service provider, incidental information that is comprised of what other information owners write or post on somebody or something, and derived information which can be analytically or statistically derived from other publicly available information. Information disclosure concept is further elaborated regarding the public disclosure and

public access concepts. Apart from the information definition and information sharing concepts there are some other concepts related to information in the lower, executive part of the model from Fig. 2. These concepts are not presented in this paper because they primarily present explicit knowledge defined in different regulation and standardization acts [14] [26] [28].

III. CONCLUSION

The paper presents our research related to the modeling approach of the contemporary information security policies, in the part that deals with conceptual information modeling. Presented model introduces formalized and more structured approach in order to facilitate the development of solutions that can keep up with the growing complexity of contemporary information security policies. The approach proposed in this paper consists of the elaboration of hierarchical taxonomy within the information security policy domain, the thorough analysis of the taxonomy terms, the conceptualization of the key terms, and the model development using UML class diagrams.

The proposed approach is illustrated in the paper through the conceptual information modeling that we consider as the central part of the proposed modeling approach. The paper shows the elaboration of hierarchical taxonomy in the part of information definition and sharing, the transformation of taxonomy terms in the tabular view of the concepts and subconcepts, and further elaboration of the concepts in the UML class diagrams.

Further research will be focusing on the development of the complete model of contemporary information security policies based on the approach proposed in this paper.

REFERENCES

[1] A. Klaić, A. Perešin, "The Impact of the National Information Security Regulation Framework on Cyber Security in Global Environment", Corporate Security in Dynamic Global Environment - Challenges and Risks, Institute for Corporative Security Studies, p. 85-96, Ljubljana, 2012

[2] A. Klaić, "Information Security in Business and Government Sectors", MIPRO 2005, Conference Proceedings BIS/DE/ISS, p. 193-198, Opatija, 2005

[3] Council Decision (2011/292/EU of 31 March 2011) on the security rules for protecting EU classified information, OJ L-141, 27.05.2011, p. 17-65, 2011

[4] HRN ISO/IEC 27001:2006 (ISO/IEC 27001:2005), www.iso.org

[5] A. Klaić, F. Turek, "Nacionalna sigurnost i telekomunikacije" (in Croatian), („National Security and Telecommunications"), Međunarodne studije, (1332-4756) II (2002), 4, p. 97-112, Zagreb, 2002

[6] B. Schneier, "A Taxonomy of Social Networking Data", IEEE Security&Privacy,http://www.schneier.com/essay-322.html, 2010

[7] A. Perešin, A. Klaić, „Povezanost koncepata kritične nacionalne infrastrukture i zaštite podataka" (in Croatian), „The Relation between the Concepts of the Critical National Infrastructure and the Data Protection", 3rd International Conference "Crisis Management Days", p. 13-29, Velika Gorica, May 2010

[8] A. Klaić, "Overview of the State and Trends in the Contemporary Information Security Policy and Information Security Management Methodologies", MIPRO 2010, Conference Proceedings Vol. V. DE & ISS & miproBIS &GLGPS &SP, p. 136-141, Opatija, 2010

[9] Secrétariat général de la défense et de la sécurité nationale, „Instruction générale interministérielle sur la protection du secret de la défense nationale" (in Franch), („General Interministerial Directive on the Protection of the National Defence Secret"), Paris, 2010

[10] Australian Government, "Information security management guidelines - Australian Government security classification system", V 1.0, July 2011

[11] The White House, "Executive Order 13556 - Controlled Unclassified Information", November 2010

[12] Council of the European Union, "Handling of Documents Internal to the Council", 11336/11, Brussels, June 2011

[13] Council Decision (2009/937/EU of 1 December 2009) adopting the Council's Rules of Procedure, OJ L-325, 11.12.2009, p. 35-6, Brussels, 2009

[14] European Commission, "Safeguarding Privacy in a Connected World - A European Data Protection Framework for the 21st Century", COM(2012) 9 final, Brussels, 2012

[15] S. Peristeras, K. Tarabanis, S. K. Goudos, "Model-driven eGovernment Interoperability: A Review of the State of the Art", Elsevier, Computer Standards & Interfaces, 31 (2009), p. 613-628, 2009

[16] Federal Enterprise Architecture (FEA) Program, "The Data Reference Model", Version 2.0, USA, November 2005

[17] FEA, "Consolidated Reference Model Document", Version 2.3, USA, October 2007

[18] US Department of Defence, "Guidance for Implementing Net-Centric Data Sharing ", DoD CIO, April 2006

[19] S. Al shayji, N. El Zant El Kadhi, "Building Fuzzy-Logic Ontology for Political Decision-Makers", International Journal of Mathematical Models and Methods in Applied Sciences, Issue 5, Volume 5, 2011

[20] A. Klaić, N. Hadjina, "Methods and Tools for the Development of Information Security Policy – A Comparative Literature Overview", Proceedings of the 34th International Convention MIPRO 2011, Vol. V., Conferences: DE & ISS & miproBIS & GLGPS & SP, p. 190-195, Opatija, 2011

[21] J. D. Howard, T. A. Longstaff, "A Common Language for Computer Security Incidents", Albuquerque, New Mexico, U.S.A.: Sandia National Reports, 1998

[22] T. R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", International Journal Human-Computer Studies 43, p. 907-928, 1993

[23] Object Management Group (OMG), "Unified Modeling Language Superstructure", V. 2.2, http://www.omg.org, 2009

[24] T. Dillon, E. Chang, M. Hadzic, P. Wogthongtham, "Differentiating Conceptual Modelling from Data Modelling, Knowledge Modelling and Ontology Modelling and a Notation for Ontology Modelling", Australian Computer Society, 5th Asia-Pacific Conference on Conceptual Modelling, Vol.79., Wolongong, Australia, 2008

[25] S. Fenz, A. Ekelhart, "Formalizing Information Security Knowledge", ASIACCS'09, Sidney, NSW, Australia, p. 183-194, ACM, 2009

[26] S. Fenz, G. Goluch, A. Ekelhart, B. Riedl, E. Weippl, "Information Security Fortification by Ontological Mapping of the ISO/IEC 27001 Standard", Dependable Computing, 13th Pacific Rim International Symposium, 2007

[27] Department of Trade and Industry (DTI), "Information Security: Protecting Your Business Assets", November 2004, http://ww.dti.gov.uk

[28] NIST, Special Publication 800-53 Revision 4, Initial Public Draft, Security and Privacy Controls for Federal Information Systems and Organizations, US Deaprtment of Commerce, February 2012

Influence of Established Information Security Governance and Infrastructure on Future Security Certifications

Ivan Sedinić, Zrinka Lovrić
Croatian Telecom, Data and IT-NT Security Work Unit
ivan.sedinic@t.ht.hr, zrinka.lovric@t.ht.hr

Abstract - In today business environment different security certificates are not any more "nice to have" feature but business prerequisite for service providers. PCI DSS certification is a must for card issuers and merchants and ISO27001 certification is very often prerequisite to qualify for ICT services offering. In this paper will be shown how proper security governance and security framework on which is built adequate security infrastructure could simplify and speed up certification process, while at the same time reduce cost of certification. Additionally, on examples of ISO27001 and PCI DSS, influence of one existing certificate on certification process for other certificate will be analyzed.

I. Introduction

The most frequent association to the term "telecommunication" is communication. Enabling the communication via mobile or fixed telephony or via different IP services between people is the core sense of telecommunication companies' existence. All these services generate enormous amounts of traffic data. Its adequate protection and security is no longer a question of company's care and potential misuse precaution, but an essential business need. Every respectable company will take all necessary measures to protect its valuable assets, both material and non material. In order to ensure the highest possible level of security that a company can afford, security governance will be established, strengthened with security infrastructure and confirmed by security certificates.

A description of research problem is given in the paragraph II. The Paragraph III gives an overview of terms relevant for this paper and their definitions, in order to eliminate potential ambiguities related to the meaning of the terms. The paragraph IV gives a proposition of how to minimize costs of establishment of certain security level, with the special stress on the security certification cost. The conclusion is stated in paragraph V.

II. The Aim

The times when security was all about deployed technology only are behind us. There is no such thing as 100% security, but every company aspires to establish the best possible security infrastructure. It is a very complex and very expensive job to do which, in its start, needs to have two basic prerequisites: the support of management and established corporate culture based on trust and opened to changes. The responsibility for information security should be determined within the information security governance framework, the part of the corporate governance.

Employees are the key component in building the adequate level of security. Since the success of established security infrastructure will depend upon employees approach to importance of security component of processes, activities, systems or projects, their education and awareness trainings should be the starting point.

Once the prerequisites are in place, a company starts with planning of implementation of security and its integration with other parts of the business, constructing the information security management system (ISMS). It is usually made by using ISO/IEC 27002 - the Standard for Information technology — Security techniques — Information security management systems, probably the most used security framework. Compliance with its controls creates solid basis for future security upgrades of all kind. Formal compliance with ISO/IEC 27001 can be obtained by subjecting the ISMS to ISO/IEC external audit which will confirm or reject the compliance with the standard. In case of compliance, an internationally recognized certificate is issued. The certification process is more or less the same for attestation of all certificates.

The ISO/IEC 27001 certificate is often the basic one and company needs to prove the compliance with other security standards too. For example, if a company provides e-commerce services, it must be at least compliant, if not certified, with Payment Card Industry Data Security Standard (PCI DSS).

Certifications could be very expensive and in some cases even ineffective, if they are not based on clear security governance and on defined security framework, supported by proper security infrastructure.

The reduction of security certification cost will be shown on an example of imaginary company continuously working on building the adequate information security level.

III. Definitions

For more clear understanding of the whole problem and its solution, a list of definitions is given:

Corporate Governance - Information Technology Governance Institute defines the governance as: The set of responsibilities and practices exercised by the board and

executive management with the goal of providing strategic direction, ensuring that the objectives are achieved ascertaining that risks are managed appropriately and verifying that the enterprise's resources are used responsibly [13].

Security Governance – the part of corporate governance, with focus on company's strategy for reducing the risk of unauthorized access to information technology systems, data and physical assets owned by the company. This involves the organization's enterprise risk management (ERM) and security policies [11]. It successful implementation should results in the following outcomes [3]:

- Strategic Alignment - aligning security activities with business strategy to support organizational objectives
- Risk Management - executing appropriate measures to manage risks and potential impacts on an acceptable level
- Business Process Assurance/Convergence - integrating all relevant assurance processes to maximize the effectiveness and efficiency of security activities
- Value Delivery - optimizing investments in support of business objectives
- Resources Management - using organizational resources efficiently and effectively
- Performance Measurement - monitoring and reporting on security processes to ensure that business objectives are achieved

Security framework - In general, a framework is a real or conceptual structure intended to serve as a support or guide for the building of something that expands the structure into something useful and achieves effective information security governance. Security framework is based on three basic components: people, technology, and process [4].

Security infrastructure - security infrastructure comprises the facilities, technologies, networks and processes aimed at preventing security incident, limiting the damage caused by such incident or addressing the consequences of it, including incident management process. Security infrastructure involves the combination of physical and network security, as well as its use and implementation to support a security framework [4].

Standard - International Standardization Organization [12] defines it as 'A document established by consensus and approved by a recognized body that provides for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context'.

Certification - in the sense of security governance, it is an official confirmation of compliance with certain standard.

IV. Information Security implementation

Each company, even if not aware of it, has some kind of security governance and framework. When security professional is starting preparation to gain security certificates of whatever kind, first step could be assessment of existing security governance and framework, their formalization and improvement if they are not yet formalized and extension to overall company activities. In this way, already spent efforts will be utilized for new achievements.

Security strategy, aligned with business strategy must be defined. On the example of imaginary organization, called First Company Inc., business strategic goals for the next period are:

- Increase competition capabilities of organization to win
- Transform legacy products to up to date market requirements
- Each year introduce at least one innovative product to market

To support such aggressive and demanding strategy, organizational unit for security in First Company Inc. needs to define its own strategic goals aligned with the business ones. These goals could be:

- Best in the industry security reputation
- Best of the class incident management system
- Leading in the new technology security implementation

The next step is to establish organization's own security regulation (Figure 1) which is going to support implementation of security strategic goals and should consist of:

- Security policies
- Security requirements
- Risk management methodology

General Security Policy is top of regulation pyramid and all other documents are built on its provisioning.

Level 2 are Specific Security Policies which are related to the certain topic which are covered by them. That could be for example:

- Information and IT Security Policy
- Physical Security
- HR Security
- Personal and Event Protection
- Security Policy on Business Continuity...

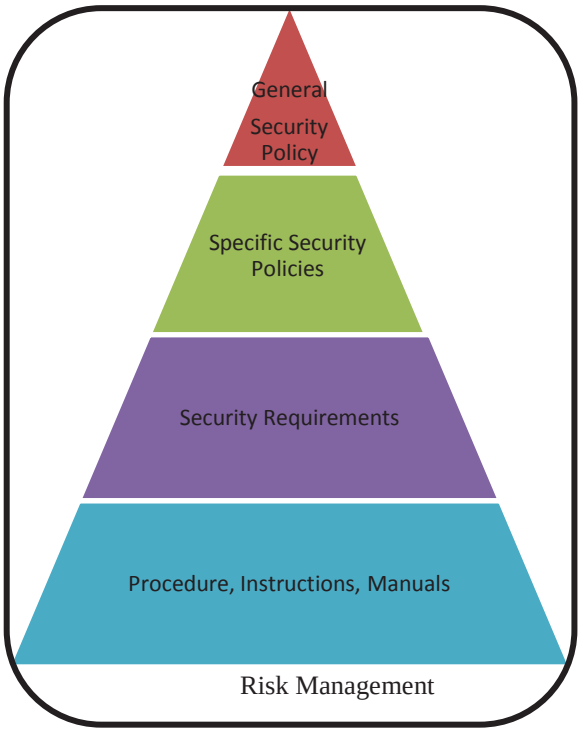


Figure 1 Security regulation structure

The number and structure of Specific Security Policies depends on the organization size, type of the business, organization structure etc. All this Specific Security Policies could be put together in one or organization could have even more than listed above, depending on mentioned specifics.

Security requirements at level 3 describe policies in greater details and they ensure that procedures, guidelines and all other operative documents on level 4 comply with policies.

Requirements are related to specific areas or topics and they could be sorted in next groups:

- Security Requirements: Operating Systems
- Security Requirements: Servers
- Security Requirements: Applications
- Security Requirements: Networks
- Security Requirements: End user devices
- Security Requirements: Databases
- Security Requirements: Incident Management

Granulation of security requirements and their grouping depends on type of the company for which they are used.

To assure that security regulation is always up to date monitoring, evaluation and update processes need to be assured. The process is shown on Figure 2.

Figure 2 Monitoring, valuation and update of company's security regulation

Once the company's security regulation is established, the last implementation step of the full security framework – the construction of security infrastructure must be done. Proper security infrastructure can be established by following up processes, security policies and requirements defined in previous steps and their implementation on proper technology (HW and SW) with trained people.

A. Compliance with ISO/IEC 27001 standard

Once the proper IT Security infrastructure, based on IT security framework and security governance is established, a comparison with ISO/IEC 27001 requirements can be done. In Table 1 mapping of ISO27001 control objectives with established security of the imaginary First Company Inc. is shown.

B. Achieving PCI DSS after ISO/IEC 27001 certification

The First Company Inc. has its virtual store with more than 1 million transactions per year. It has to achieve the PCI DSS certificate and costs of the external audit could be up to 250 000 $. If the sum for the establishment of compliance is added, the PCI DSS certification costs a small fortune. But, if the security is implemented based on the ISO/IEC 27001 controls, certain savings can be done. ISO/IEC 27001 and PCI DSS controls overlap in certain percentage, so the implemented security based on ISO/IEC 27001 can be used for achievement of compliance with PCI DSS, as shown in Table 2.

ISO/IEC 27001 control objectives and controls	First Company Inc.
A.5 Security policy	General Security Policy and Specific Security Policies
A.6 Organization of information security	Formal act about organization structure, authorities and responsibilities Information and IT Security Policy
A.7 Asset management	Formal act about organization structure, authorities and responsibilities Formal act on inventory management in organization Security requirements related to assets
A.8 Human resources security	HR Security Policy Security Requirements (applications, authorization, …)
A.9 Physical and environmental security	Security Policy on Physical Security Security Requirements on Data Centers, Servers, …
A.10 Communications and operations management	Information and IT Security Policy Security Requirements on networks, applications, outsourcing, end user devices, …
A.11 Access control	Information and IT Security Policy HR Security Policy Security Requirements on applications, authorization, network, operating systems, end user devices, …
A.12 Information systems acquisition, development and maintenance	Information and IT Security Policy Security Requirements on applications, operating system, outsourcing, …
A.13 Information security incident management	Security Policy/Requirement on incident management
A.14 Business continuity management	Security Policy/Requirement on Business Continuity
A.15 Compliance	In formal act about organization structure responsibility for compliance to law, regulation and contractual obligation need to be clearly defined

Table 1: mapping of ISO27001 control objectives to company documents and processes

V. Conclusion

When a company starts to consider obtaining certain security certificate almost always pops up the question: How much it will cost?

Some companies will immediately hesitate to approach it, since the average cost of, for example, ISO/IEC 27001 amounts more than 50.000 $ and PCI DSS even few hundred thousand dollars, calculating cost of the external audit, investment in infrastructure and employees engagement. These amounts are so big only in cases where company starts "green field", without any previous work and without existing information security. In reality, today it is very rare that company which has some data worth to protect does not have some kind of security in place. Some kind of security infrastructure can be found in almost every company, used for regular business or dedicated to perform security functions (e.g. routers, servers…).

PCI DSS [1]	ISO/IEC 27001 [10]
1. Install and maintain a firewall configuration to protect data.	A10.6. Network Security Management A11.4. Network Access Control
2. Do not use vendor–supplied defaults for system passwords and other security passwords	A10. Communications and operations management A11. Access Control A12. Information systems acquisition, development and maintenance
3. Protect stored data	A10. Communications and operations management A12. Information systems acquisition, development and maintenance A15. Compliance
4. Encrypt transmission of cardholder data and sensitive information across public networks.	A10. Communications and operations management A11. Access Control
5. Use and regularly update anti-virus software	A10.4. Protection against malicious and mobile code
6. Develop and maintain secure systems and applications.	A10. Communications and operations management A11. Access Control A12. Information systems acquisition, development and maintenance
7. Restrict access to data by business need to know.	A8.1.1. Roles and responsibilities A8.3.3. Removal of access rights A11. Access Control
8. Assign a unique ID to each person with computer access.	A8. Human Resources security A10. Communications and operations management A11. Access Control
9. Restrict physical access to cardholder data.	A8. Human Resources security A9. Physical and Environmental Security A10. Communications and operations management
10. Track and monitor all access to network resources and cardholder data.	A10. Communications and operations management A11. Access Control
11. Regularly test security systems and information security management systems with all controls specified in accordance with systems and processes.	A10. Communications and operations management A12. Information systems acquisition, development and maintenance
12. Maintain a policy that addresses information security.	All [12]

Table 2: mapping of ISO 27001 control objectives to PCI DSS requirements

ISO/IEC 27001 certification costs consist of:

1. Training and literature
2. External consultancy (not needed if the company has experienced employees)
3. Technology – HW and SW (most of it already in place)
4. Internal work (part of it already done as explained previous)

5. Certification audit (few up to few ten man days, depending on scope and company size)

From the stated it can be concluded that up to 50% of certification costs could be already covered in organization.

The PCI DSS certification costs can be grouped into two types:

1. Assessing the scope of required PCI DSS related work
2. Activities to meet the PCI DSS requirements

All these expenses will depend on how many experienced employees able to do some preparation work company has. With proper preparation, properly implemented ISO/IEC 27001 including, cost of assessing could be significantly lowered. Required activities to meet the PCI DSS requirements, when ISO/IEC 27001 is already implemented will be reduced and only 30% of the controls, where PCI DSS requires stronger control than ISO/IEC 270001, will be an extra job.

With this data, it is clear that with implementation of ISMS, using formal security governance and framework, ISO/IEC 27001 certification is significantly less expensive than when the company starts implementation of security from scratch. Efforts and expenses spent on ISO/IEC 27001 certification will also pay off with significant reduction of PCI DSS certification cost.

References

[1] Fitzgerald, Todd: Information Security Governance Simplified: From the Boardroom to the Keyboard, CRC Press, USA, 2011

[2] von Solms, S.H, von Solms, R.: Information Security Governance, Springer Science+Business Media LLC, USA, 2009

[3] Brotby, Krag: Information Security Governance: A Practical Development and Implementation Approach, John Wiley and Sons Inc., USA, 2009

[4] Whitman, M.E., Mattord, H.J.: Principles of Information Security, Cengage Learning, USA 2011

[5] Posthumus, S., von Solms, R.: A Framework for the Governance of Information Security, Computers&Security, vol.23 (638-646), USA, 2004

[6] von Solms, B., von Solms, R.: From Information Security to...Business Security?, Computers&Security, vol.24 (271-273), USA, 2005

[7] Da Veiga, A., Eloff, J.H.P.: A Framework and Assessment Instrument for Information Security Culture, Computers&Security, vol.29 (196-207), USA, 2010

[8] PCI DSS v.2, PCI DSS Council, 2011, available on: www.pcisecuritystandards.org

[9] HRN ISO/IEC 27001:2006, HZN Glasilo, 03/2006, Zagreb, Croatia, 2006

[10] HRN ISO/IEC 27002:2006, HZN Glasilo, 03/2006, Zagreb, Croatia, 2006

[11] National Cyber Security Partnership, 2004, Information Security Governance – A call to action (available on: www.cyberpartnership.org/InfoSecGov4_04.pdf)

[12] ISO/TMB Policy and Principles Statement - Global Relevance of ISO Technical Work and Publications (available on: http://www.iso.org/iso/global_relevance.pdf)

[13] Information Security Governance, Guidance for Boards of Directors and Executive Management, 2nd Edition, IT Governance Institute, 2006, USA (available on: http://www.isaca.org/Knowledge-Center/Research/Documents/InfoSecGuidanceDirectorsExecMgt.pdf)

Knowledge based authentication requirements

K. Skračić, P. Pale, B. Jeren
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Hrvatska
E-mail: kristian.skracic@fer.hr

Abstract – Published evaluation criteria for knowledge based authentication (KBA) methods do not provide a sufficiently formed framework to use as a guideline during design and testing of KBA methods and tools. The aim of this paper is to define a set of requirements for creating a secure user authentication method based on the user's knowledge. The requirements address four issues in user authentication. The first refers to eavesdropping an authentication session and using the intercepted information in the next session. By repeating the recorded response an attacker should not be able to authenticate himself as a legitimate user. The second issue is the ability to predict an authentication challenge by analyzing previous challenges. If an attacker can record a set of challenges over a long period, he should not be able to learn the next challenge beforehand. The third issue is the guessability of correct responses to authentication challenges. In general, multiple sources of information about the user are available to an attacker. The correct response to a challenge should not be obvious from such sources. The fourth issue is the authentication server's vulnerability. By this any information system component that is used to authenticate users is meant. If an attacker manages to gain partial or complete access to the authentication server and its data, the user's digital identity should not be compromised. These for requirements are proposed as a generic checklist against which KBA methods and tools should be checked.

I. INTRODUCTION

User authentication is a prerequisite for meeting basic security requirements like access control or document authenticity. Smart cards, tokens, biometric and other methods of authenticating users in information systems are gaining in popularity. However, password based authentication methods are still the dominant way of authenticating users [1], despite their numerous documented shortcomings. As stated in [1], new authentication mechanisms are yet to achieve the necessary popularity to replace passwords. Some of the reasons are explained in the next section.

In general, there are three formally accepted ways of authenticating users [21]: using something the user has, something the user knows or something the user is. These are often called authentication factors. Some research has been done to introduce a fourth factor: [22] proposes the fourth factor to be the user's location, while the research in [23] proposes using user's acquaintances. Neither of these proposals have been formally accepted yet. In this paper we focus on knowledge based authentication methods which use the factor of something a user knows.

This paper analyzes the key requirements a secure user authentication method based on user knowledge should meet. To the best of our knowledge, there are no formal requirements for user knowledge based authentication methods. However, there are several methods for evaluating knowledge based authentication methods presented in the literature. Therefore, we analyze existing evaluation methods for user knowledge based authentication methods.

The analysis of different types of knowledge base authentication (KBA) methods is outside the scope of this paper. Some insight can be found in [16] and [17]. This paper is organized as follows. Section 2 provides an overview of existing evaluation methods for knowledge based authentication. The focus is only on fact based knowledge authentication methods and textual passwords. Evaluation methods for graphical and visual passwords are beyond the scope of this paper. Section 3 describes the new proposal for user knowledge based authentication requirements. Finally, Section 4 presents authors' conclusions.

II. EVALUATION METHODS

There are several ways to evaluate authentication methods based on user knowledge. Most existing evaluation criteria are based on the experience of researchers, described as a set of security properties.

Fact based authentication methods are often called "personal verification questions". This is because most authentication questions are based on personal information from the user. However, fact based authentication methods may use any other information relating to the user. Fact based authentication methods can be divided into two types depending on the questions used to authenticate a user [25]. Static KBA constitutes authentication methods that use predetermined facts to authenticate users. For example, the user's date of birth or mother's maiden name. The use of static KBA has been widely criticized and is considered to be a weak authentication method [18]. The second type is dynamic KBA. It is similar to static KBA in that it is also based on personal information about the user. However, in dynamic KBA the user does not share the secret information beforehand. Rather the questions and answers could be derived from various data collections about user's activities like: banking, shopping, etc. This form of knowledge based authentication is new and has not been sufficiently researched. The primary problem is the acquisition of data needed to create a unique question for

the user. Some research has been done on creating dynamic KBA authentication methods [24]. At this time there are no widely used authentication methods based on dynamic KBA.

Since textual passwords are the dominant authentication method in this category, the analysis begins with them. The goal is to analyze the advantages password based authentication methods have that makes their use so widespread in today's information systems.

The research conducted in [1] introduces a wide list of properties that define user authentication methods. These properties are separated into three categories. The first category questions usability benefits. These benefits evaluate how easy it is for users to authenticate themselves to the system. For example, how easy it is to memorize the information which is needed to authenticate, does the user need to carry anything with them or how easy it is to recover from loss. The second category focuses on deployability. This is an important requirement since many authentication methods in general have been discarded because of the difficulty of deploying them on a large scale. Textual passwords have a clear advantage in this area as they are remarkably easy to implement and they scale well. Furthermore, most systems implement them by default. The final category examines security benefits. Although extensive, the list of properties in [1] is not complete, as the authors themselves conclude. It can be concluded that the popularity of authentication methods is primarily based on usability and deployability, rather than security. The authors of [1] conclude that the marginal gains of new authentication mechanisms are not enough to overcome the large cost of transitioning from one authentication method to another. In conclusion, the research provided in [1] gives a clear overview of why some authentication methods are more prevalent than others. However, the proposed security properties offer no insight into how a secure authentication method should be designed. Therefore, the security properties described in [1] cannot be used as a set of requirements for a secure user authentication method.

Another evaluation method for knowledge based authentication methods can be found in [2]. Although it focuses on introducing a new authentication method based on questions and answers, it establishes a useful set of properties. For example, users should find it easy to remember the correct answers. Also, the question set should be fairly large in order to avoid repeating the same question several times. This would make it more difficult for an attacker to use the recorded answer. A complete overview of the proposed properties is accessible in [2]. It should be noted that these properties only apply to knowledge based authentication methods based on facts. Also, these rules overlap with the usability and security benefits introduced in [1]. This overlap could confirm the usefulness of the properties in [1] and [2]. However, they are used only for comparing and evaluating authentication methods. The proposed properties should not be considered as a set of requirements for two reasons. First, their purpose is strictly aimed at evaluating authentication methods. Second, they do not express a requirement that needs to be fulfilled in order to create a secure authentication method.

One of the first formal evaluation methods for knowledge based authentication systems can be found in [3]. However, like the work in [2], these criteria are aimed at authentication systems with challenge questions and they relate to their privacy, security and usability. Also, as far as the criteria are concerned, they are mostly similar to the previous two evaluation methods. The research in [3] also establishes an overview of question based authentication methods and their qualities, but such an overview is beyond the scope of this paper. We note that the research in [3] suggests creating challenges based on user's preferences or intimate data.

Comparing these three evaluation methods, we conclude that they are similar to one another. Each of them is based on its own, limited set of properties that the authors have identified as relevant when evaluating the security of user authentication methods. Furthermore, the evaluated methods address the same issues. For example, the problem of intercepting data, guessing the correct response to an authentication question and resistance to Phishing. In essence these methods enumerate properties which can be expected to change over time. It is difficult to base a design of a new authentication method on such a list. Rather, a small number of generic requirements for user authentication could be better suited and is proposed in this paper in next chapters.

III. USER AUTHENTICATION REQUIREMENTS

The previous Section reviews a set of techniques for evaluating user authentication methods. To the best of our knowledge, there is not a single knowledge based authentication method that satisfies all evaluation methods presented in [1], [2] and [3]. Therefore, this paper proposes that current evaluation methods for user authentication are not constructive enough. It is authors' opinion that creating an authentication method by merely satisfying a set of properties may lead to the creation of vulnerable methods. This opinion is based on the fact that the properties described in [1], [2] and [3] represent the most common ways to bypass user authentication methods. They do not represent an exhaustive set of possible vulnerabilities. For example, the research in [1] suggests a number of criteria that are supposed to guard against intercepting secret data. One of them is resilience to physical observation. This constitutes a potentially endless set of attacks because physical observation constitutes a number of things. For example, an attacker can observe a user's monitor or keyboard. Guarding sensitive information on a monitor is different from guarding information on a keyboard. Another problem represents the meaning of the word observation. Does this constitute only visual observation, or can listening to and analyzing keystrokes be considered physical observation [10]?

Therefore, it can be concluded that a set of properties for secure user authentication methods does not constitute an exhaustive list of guidelines and cannot provide more information for future vulnerabilities. Therefore, it is proposed that the solution may be found by using more generic requirements to encompass all current and future vulnerabilities.

There are countless opportunities for data to be intercepted in an information system. Enumerating concrete ways of intercepting data may prove useful in reducing their occurrence. However, such a static list will be insufficient for guarding against future attacks.

Based on this, it can be concluded that there is no finite subset of properties that can ensure that an authentication method will never be compromised. Instead, a set of higher level requirements is proposed that a user authentication method should satisfy. Such higher level requirements are usable for evaluating a wide range of knowledge based authentication methods.

The following chapters define a set of (generic) requirements that an authentication method must satisfy in order to be secure. In order to be applicable for all knowledge based authentication methods, the requirements are intentionally generic. In other words, the requirements do not specify how to achieve security, but what is needed to achieve security.

A. One time challenges

When thinking about a secure authentication method, the first question one needs to ask is: what needs to be stolen in order to compromise a digital identity? Information can be intercepted in a variety of ways. Since communication is often carried out across the Internet, a lot of attention has been given to securing communication networks against eavesdropping. However, experience has proven that even the most secure mechanisms have flaws that make it possible to intercept secret data. For example, recent work [4] has shown that there are ways of circumventing SSL/TLS protection. Another obvious flaw is the inherent trust in user judgment. As explained in [5], an attacker can easily insert himself in a secure communication channel if the end user accepts the attacker's forged certificate.

Additionally, in a large information system, data can be intercepted in a non technical way with the use of social engineering techniques. For example, extracting secret data from an employee or "Dumpster diving"[6].

In the case of textual passwords, user authentication comes down to a single piece of information: the user's password. Once an attacker acquires the user's password, he can authenticate himself as that user as many times as he wants. Some policies address such issues by forcing users to change their passwords on a regular basis (usually once a month). According to [7] and [8], this leads to more problems as people tend to choose passwords that are easier to remember. At best, such a strategy only makes it slightly more inconvenient for an attacker to compromise a user's digital identity. There is still a formidable period in which the attacker can use the obtained secret information and pose as another user.

Amongst the more unusual eavesdropping techniques is the work done in [9] where the researchers managed to differentiate the sound made by using a keyboard. Their research showed that an eavesdropper can recognize what the user typed by the sound of the keys. Further research [10] has improved this to use keyboard characteristics and statistics of the text to decipher a recording of text typed for a period of ten minutes on a random keyboard. It is worth noting that the software used needs no previous training to recognize keystrokes. A similar experiment was conducted in [11]. The researcher has shown that a VDU's screen content can be recovered optically even from diffuse light reflected off room walls or the users face.

Based on the above, we propose that a secure authentication method must use one time passwords to authenticate users. If a password is used more than once, the authentication method can be compromised. Since passwords are not meant to be changed after every authentication, a more generic approach could be used in which a user must respond to an arbitrary challenge presented by the authentication server. Each challenge must be unique in order to satisfy the one time rule. An exception can be made if the challenge is constructed in a way that makes the correct response different every time. That way, even if the challenge is repeated several times the response cannot be misused in another authentication attempt. An example of such a challenge can be the user's phone bill. Since this is something that changes every month. Additionally, the information used for such a challenge can have a shorter lifespan. For example, the number of phone calls a user made today.

B. Challenge predictability

Considering that a secure authentication must use one time challenges, the next step to ensure secure authentication is that the challenge has to be unpredictable. One time challenges technique by itself does not ensure that the challenges will be sufficiently random.

Generally speaking, there are two ways of predicting a one time challenge. The first is to try to use a number of previously recorded challenges to predict the next one. The idea is that a challenge is generated using an algorithm. In case the attacker manages to reverse engineer the algorithm based on recorded previous challenges, he can learn how to predict the next one.

The second way to predict a one time challenge is the adaptive "Chosen message attack". This attack is based on adaptively querying the authentication server in order to gain information. If an attacker discovers how his inquiries influence creation of server's challenges, then he may be able to predict them. This can be the case if the challenges are functionally dependant on some form of user data. For example, user data can be the time the system is being accessed, the location of the user or the user's web browser. Research conducted in [13] calls such an attacker an interrogative adversary. The difference in this and the first approach is that the attacker managed to limit the possible set of queries to be issued. Both attacks assume some form of algorithm based approach to generation of challenges.

Based on this it can be concluded that the predictability is not only influenced by the algorithm used to create the challenges, but also by the type of attacker. Thus, it should be required that the probability that an interrogative adversary can guess the authentication challenge must be equal to or smaller than the probability

of a non interrogative attacker guessing the challenge. Furthermore, all authentication challenges must be independent of each other in order to make it impossible to predict the next challenge based on a sample of the previous challenges. It is important that not even the legitimate user cannot predict challenges he will receive in the future. It should not only protect him from blackmailing situations but could also reduce blackmailing attempts if that property of KBA would be publicly known. Based on this, it is proposed that the challenges should not be algorithm based.

C. Response guessability

The evaluation methods proposed in [1], [2] and [3] covered guessability issues in great detail. For this reason it is further explained only where guessability concerns influence onetime challenges. When using onetime challenges, guessability is less risky. Even if an attacker manages to guess a correct response, he will not be able to use the same guessed value the second time. Still, this does not condone the practice of using challenges to which answers are easy to guess. Obvious counterexamples are the static KBA questions for fallback authentication. As described in [19], responses challenges based on personal questions tend to be extremely guessable. Given the power of social media and public records, a lot of user information is accessible to attackers. One such example is described in [18]. Research [20] has proven that challenges based on dates of birth, family members and partners are often easily guessable.

It can be concluded that a single challenge should be impervious to educated guessing attacks. In other words, if an attacker has some knowledge about the user or the system, he should not be able to guess the correct response to a challenge. Ideally, the challenges should be based on private data. By this we mean the use of data which the user is unlikely to share with friends and aquantances or make public via social networking sites. We recognize that such challenges may pose a privacy concern. However, the construction of the challenges is beyond the scope of this paper.

D. Independence of the authentication server security

Knowledge based user authentication is based on shared information. When authenticating a user, the authentication server must have a piece of information that it shares or knows about the user. This means the server has to keep this information in a local storage. If an attacker manages to acquire the shared information from the server, the user's digital identity is compromised. This poses an even bigger problem if a user uses the same information on several servers.

Therefore, there should be a requirement that the authentication method must implement the use of shared knowledge in such a way which makes it independent of the authentication server security. In other words, if an attacker manages to compromise the authentication server he should not be able to compromise the user's digital identity.

Obvious counterexamples are textual passwords, even if they are stored in some encrypted form and are salted. In cryptography, password salting is the process of adding a random value to a user password before hashing it with a one way function. When an attacker gains access to a server that uses salted passwords for authentication he can extract the salt value and the hash generated from the salt and the user's password. Using dictionary or brute force password guessing attacks, the original password can be retrieved in finite time. However, the attacker does not know what the salt value is without gaining access to the authentication server. So the attacker cannot compute the dictionary before knowing the salt value. This slows the process of guessing the password because the attacker has to generate a dictionary for every salt value. Recent work [12] demonstrates that salting only complicates password guessing, but does not prevent it. Given the number of CPU/GPU cores today, salting user passwords will not be an obstacle to attackers for long. Also, some work has been done on using FPGAs [15] to parallelize dictionary attacks. Coupled with smarter algorithm [14], guessing hashes is becoming easier and increasingly faster

To the best of authors' knowledge, currently there is no authentication method that meets this requirement. The obvious reason is that most information systems use only one server for authentication and this server usually holds all the information needed for authenticating users. For example, a server has a database of user credentials.

Based on this, it can be concluded that an authentication server needs to be distributed in order to meet this requirement. In other words, the server that implements user authentication must not be the server that the user wishes to gain access to. This mechanism only creates an additional step in the process of compromising security. The attacker now needs to gain access to the server that is used for authenticating users. The level of security could be further significantly improved by using multiple different servers to hold parts of shared information. Thus an attacker would need to gain access to them all.

IV. Conclusion

This paper proposes a set of requirements for authentication methods based on user knowledge. Based on the research of current evaluation methods for user authentication, it is believed that proposed requirements are sufficient to evaluate the security of existing and future authentication methods based on the user's knowledge.

Although the proposed requirements seem generic, they were not evaluated against authentication methods that use tokens or biometrics. However, further research may involve expanding the requirements so as to include authentication methods based on other authentication factors.

References

[1] Bonneau, J.; Herley, C.; van Oorschot, P.C.; Stajano, F.; , "The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes," Security and Privacy (SP), 2012 IEEE Symposium on, vol., no., pp.553-567, 20-23 May 2012

[2] L. O'Gorman, A. Bagga, and J. Bentley, "Query-directed passwords," Computers & Security, vol. 24, no. 7, pp. 546–560, Oct. 2005.

[3] M. Just, "Designing Authentication Systems with Challenge Questions," in Security and Usability: Designing Secure Systems That People Can Use, 1st ed., O'Reilly Media, 2005, pp. 143–160.

[4] W. El-Hajj, "The most recent SSL security attacks: Origins, implementation, evaluation, and suggested countermeasures," Security and Communication Networks, vol. 5, pp. 113-124, 2012.

[5] Zhe Chen; Shize Guo; Rong Duan; Sheng Wang; , "Security Analysis on Mutual Authentication against Man-in-the-Middle Attack," Information Science and Engineering (ICISE), 2009 1st International Conference on , vol., no., pp.1855-1858, 26-28 Dec. 2009

[6] Janczewski, L.J.; Lingyan Fu; , "Social engineering-based attacks: Model and new zealand perspective," Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on , vol., no., pp.847-853, 18-20 Oct. 2010

[7] Patterson B; , letter to Communications of the ACM, vol 43, no. 4., Apr. 2000

[8] R. J. Anderson, "Chapter 2.4.3 Naïve Password Choice" in Security Engineering: A Guide to Building Dependable Distributed Systems, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 2001.

[9] Asonov, D.; Agrawal, R.; , "Keyboard acoustic emanations," Security and Privacy, 2004. Proceedings. 2004 IEEE Symposium on , vol., no., pp. 3- 11, 9-12 May 2004

[10] Zhuang L., Zhou F., Tygar JD., "Keyboard Acoustic Emanations Revisited", 12th ACM Conference on Computer and Communications Security, 2005

[11] Kuhn, M.G.; , "Optical time-domain eavesdropping risks of CRT displays," Security and Privacy, 2002. Proceedings. 2002 IEEE Symposium on , vol., no., pp. 3- 18, 2002

[12] F. Milo, M. Bernaschi, and M. Bisson, "A fast, GPU based, dictionary attack to OpenPGP secret keyrings," Journal of Systems and Software, vol. 84, no. 12, pp. 2088–2096, Dec. 2011.

[13] K. Fu, E. Sit, K. Smith, and N. Feamster, "Dos and don'ts of client authentication on the web," in Proceedings of the 10th conference on USENIX Security Symposium - Volume 10, Berkeley, CA, USA, 2001, pp. 19–19.

[14] A. Narayanan and V. Shmatikov, "Fast dictionary attacks on passwords using time-space tradeoff," in Proceedings of the 12th ACM conference on Computer and communications security, New York, NY, USA, 2005, pp. 364–372.

[15] Dandass, Y.S.; , "Using FPGAs to Parallelize Dictionary Attacks for Password Cracking," Hawaii International Conference on System Sciences, Proceedings of the 41st Annual , vol., no., pp.485, 7-10 Jan. 2008

[16] Agarwal, G.; Singh, S.; Indian, A.; , "Analysis of knowledge based graphical password authentication," Computer Science & Education (ICCSE), 2011 6th International Conference on , vol., no., pp.588-591, 3-5 Aug. 2011

[17] Gkarafli S.; Economides AA.; , "Comparing the Proof by Knowledge Authentication Techniques", International Journal of Computer Science and Security, vol. 4, no. 2, pp. 237-255, May 2010

[18] Lemos R.; , Are Your "Secret Questions" Too Easily Answered?, Technology Review, MIT, 2009.

[19] A. Rabkin, "Personal knowledge questions for fallback authentication: Security questions in the era of Facebook," in SOUPS 2008 - Proceedings of the 4th Symposium on Usable Privacy and Security, 2008, pp. 13–23.

[20] V. Griffith and M. Jakobsson, "Messin' with texas deriving mother's maiden names using public records," in Proceedings of the Third international conference on Applied Cryptography and Network Security, Berlin, Heidelberg, 2005, pp. 91–103.

[21] Guidance on Multi-factor Authentication, State Services Commission, June 2006, Version 1.0

[22] Choi, Sung; Zage, David; , "Addressing insider threat using "where you are" as fourth factor authentication," Security Technology (ICCST), 2012 IEEE International Carnahan Conference on , vol., no., pp.147-153, 15-18 Oct. 2012

[23] J. Brainard, A. Juels, R. Rivest, M. Szydlo, and M. Yung. Fourth Factor Authentication: Somebody You Know. In ACM CCS, pp. 168-78. 2006.

[24] A. Alkhalifah and G. D. Skinner, "Enhanced knowledge based authentication using iterative session parameters," World Academy of Science, Engineering and Technology, vol. 71, pp. 293–299, 2010.

[25] S. Chokhani, Knowledge Based Authentication (KBA) Metrics, Orion Security Solutions, 2004.

A New Method for the Identification of Proactive Information Security Management System Metrics

Kemal Hajdarevic*, Pat Allen**
* Faculty of Electrical Engineering, University of Sarajevo,
Sarajevo, Bosnia and Herzegovina
** Dept of Computing, The Open University, Milton Keynes, UK
*Kemal.Hajdarevic@etf.unsa.ba, **p.m.allen@open.ac.uk

Abstract – Information security is topic of everyday interest, with mainstream media reports revealing information security incidents in many different areas. These reports demonstrate the importance to any organization of having an information security management system (ISMS). Foreseeing potential security risks is usually key to successful risk management. Available information security standards such as the ISO 27000 set of standards give a formal framework for successful information security management in any size of organisation or company. In this paper we draw on experience gained during a project leading to successful ISO 27001 certification at the Central Bank of Bosnia and Herzegovina in 2009. We review recent work on proactive damage prevention, and we propose methodology based on the GQM (Goal, Question, Metrics) paradigm for determining proactive steps for detection and resolution of different information security control violations. For creating proactive measurement metrics we use the well recognised standards ISO 27004:2009, and NIST 800-55. We present several examples of proactive metrics.

I. INTRODUCTION

Every day sees new media reports on cyber crime, cyber warfare, data corruption and thefts of data by outsiders or disgruntled employees. These incidents cause financial loss and damage the public image of companies and organisations.

To successfully prevent the effects of security violations, proactive actions have to be planned in advance, defined and communicated to staff who must be trained for each specific risk resolution situation. From experience we can say that for many security incidents prior knowledge was available which could have been used to prevent specific incidents, or to stop an event escalating to the level of an information security incident. For each information security incident we often identify single or multiple unexpected events related to information security which have a significant probability of compromising the organisation's information security. This is because in the incident escalation phase as a result of an attack on a computer system every second is important for proactive action. Before any successful attack there are steps that the attacker has to perform which can be used as points for attack detection and proactive damage prevention as proposed by Beitlich [1]. Figure 1 shows the stages of a typical attack.

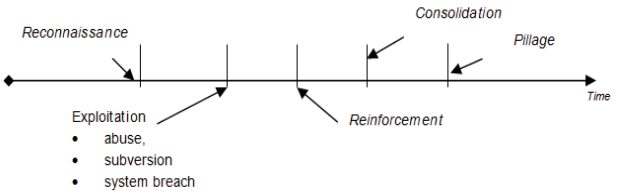


Figure 1. Attack stages (from [1])

An information security system with proactive protection mechanisms is more resilient to known threats. In this paper we present an approach for planning proactive actions supported with appropriate metrics based on the ISO 27000 series of standards. We present and compare two standards for information security metrics: ISO 27004 [2] and NIST 800-55 [3] and propose a proactive measurement method based on these standards. These and other related standards are intended to protect organisation's information assets which are essential to all of today's business operations.

Information is used for internal communication, operations, internal and external reporting, archiving and for many other purposes. Information can be presented in documents, video or audio files, stored on hard disks of PCs, servers, or removable drives. Operating system software, applications, databases, web, and file servers hold information too. Information can be used by the specific system as configuration options for its operations and can reside in configuration files. Information is vulnerable when stored in files, or transferred over communication channels.

Data stored on electronic, optical, or magnetic media are vulnerable to computer malware (malicious software), such as viruses, spyware, Trojan horses and worms which can change, steal, or delete files and information stored on all of these media.

Communication devices such as routers, switches, and modems transfer and route information in specific

directions over communication channels which have to be safe against third parties which might try to steal or damage this information.

Information can be intercepted and even changed during transfer from source to communication destination without the knowledge of the information owners and genuine information users. Physical communication channels such as fibre optic cables are vulnerable to this type of attack (although less vulnerable than copper wires), as are wireless or logical communication channels such as frame relay communication links.

Active attackers can harm information stored in data files by means of unauthorized access. Such attacks are not only threats to electronic information, since passive attacks also exist [4].

Physical channels such as copper cables can be the object of passive attacks. A wire tap is easily achieved by simply attaching wires to copper cables or by inserting a tap device, making it possible to listen traffic passing by. In some cases it is possible for an attacker to participate in the conversation and to change data packets. This is easy for the attacker if traffic is not protected with mechanisms such virtual private network (VPN) [5], or by uses of other mechanisms to protect communication channels.

Fibre optic is less vulnerable than copper cables due to its physical characteristics that make impossible to attach another fibre cable to the original one. This will not work because of the light bending of the fibre cable coats. Tapping is possible only by breaking the fibre cable, preparing appropriate connectors and installing a tap device, but for that action there has to be time spent, which means that attack can be more easily detected. Wireless communication is by the nature of its operation vulnerable to physical tapping. Because of that there are a number of cryptographic solutions such as VPN for protecting wireless and wired computer communication channels.

Threats and vulnerabilities are not limited only to electronic ones; there are others as well such as environmental threats (see for example [6], [7]) for which business and disaster recovery plans are required to support business contingency.

The risk of all the above mentioned vulnerabilities and threats can be reduced by applying specific technical or organizational solutions, and by managing information security risks using standards such as is 27005:2008 [8] and expert guidance on their application, such as [9] and [10]. The current version of ISO 27001:2005 describes 133 controls for information security management. Such controls cannot be seen as separate solutions, but need to be integrated into a holistic approach to information security management, on the principle that the whole is more than the sum of its parts [18]. One way to prevent information security incidents is to use ISO 27001:2005 [11], for which a new draft exists, and ISO 27002:2005 [12], as the basis for an Information Security Management System (ISMS). The approach we describe in this paper is based on work that was done at the Central Bank of Bosnia and Herzegovina to implement such an ISMS, and obtain ISO 27001 certification.

II. RELATED WORK

One of the starting points for this research was the approach of Calder [13], and Calder and Watkins [14] for information security governance in any business organisation. Their approach is based on implementation of an information security standard for which is possible to receive internationally recognised certification such as ISO 27001.

These authors generally describe what the specific controls of each standard covers, but not how they can be implemented and how performance of the system can be monitored. Beckers et al. [15] propose conceptual frameworks and security requirements engineering methods. These are necessary because the ISO 27001 standard does not provide precise information how the documentation for an ISMS has to be developed, which controls have to be tied to specific business roles, and how this can be achieved. Boehmer [16] argues after cost/benefit trade-off analysis that an ISMS based on ISO 27001 is equivalent to risk management, which is equivalent to cost/benefit management, and has to be considered in companies which want to avoid wasting investments in the information security of their companies.

Stambul and Razali [17], stress that without information security any organisation has no guarantee of long-term success. They present a model which suggests some elements necessary for defining levels of successful information security implementation. As a possible advance in information security data collection, Montesino and Fenz [18] identify controls for which automatic data collection can be performed and tools that can automate the process. In our work we also considered work done by Coetzee [19], where a holistic framework for information security governance for SOA is presented in a way that supports the development of a proactive information security system. One approach is described by Anwar et al. [20] who propose the establishment of a proactive measurement method which can be adjusted to any information security activity in a specific business environment.

To summarise the work reviewed above, in order to measure the effectiveness and performance of an ISMS, specific metrics have to be monitored by collecting relevant data, extracting useful information, and performing appropriate information analysis using available tools. The result of information analysis could be the creation of meaningful reports in a form of Key Performance Indicators (KPI) which can be used to provide relevant information for making appropriate decisions.

Monitoring of relevant KPIs is important for each monitored control of the ISMS such as monitoring for application security [21], which is related to the approach presented here. For some controls it is possible to implement automatic data collection [22], [23],

information extraction and analysis with predefined rules and for others human operators have to enter data in semiautomatic form. To be able to create relevant KPIs, relevant measurement specifications and meaningful metrics have to be defined and established.

III. MEASUREMENT SPECIFICATIONS

When deciding how to make meaningful metrics and how to perform measurements, information security measurement constructs have to be defined for each metric. Below we give proposals for information security measurement constructs based on the ISO 27004:2009 standard [2] and compared with NIST guidance [3].

A. Measurement Construct Identifier

To uniquely identify a measurement result, or group of measurement results, related to a specific control objective a measurement name with a unique numerical identifier has to be specified. The purpose of the measurement construct has to be defined in relation to the control process objective. As an example, we might create a unique measurement construct identification for the control specified by ISO 27002, 10.10.1 Audit logging. The number of control can be used together with a prefix specifying Measurement Method.

We use the MM abbreviation + control number to uniquely identify measurement results for specific control; for example MM10101 could be the control 10.10.1 Audit logging from ISO 27001:2005.

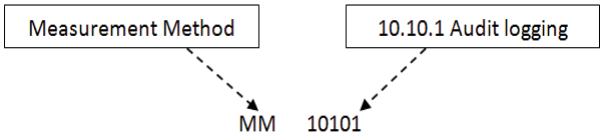


Figure 2. Measurement construct identification

In NIST document 800-55 [3] 'Measurement Construct Identification' is defined as 'Measure ID'.

B. Object Measurement and Attributes

Objects of measurement can be processes, plans, projects, resources, systems or system components. Each object is characterized by the measurement of its attributes. These attributes can be distinguished qualitatively or quantitatively. In NIST document 800-55 [3] 'Object measurement and attributes' is defined as 'Goal'.

C. Base Measure Specification

Base measure specifications are determined by the object attributes and measurement methods which are used to quantify the attributes (for example: number of incidents, number of approved system changes, number of un-patched systems).

There are two types of measurement method: subjective and objective. Objective measurement is where it is not possible for humans to alter data, and where human judgment is not used, and subjective measurement is where human judgment is considered.

Other elements of base measurement specifications are scale, type of scale, and unit of measurement, where scale are set of values or ordered marks at fixed intervals which can defined as different type of scale: nominal, ordinal, interval or ratio, and unit of measurement is consistent so that quantities can be compared. In NIST document 800-55 [3] 'Base Measure Specification' is defined as 'Measure'.

D. Derived Measure Specification

The result of a measurement which is function of two or more base measurements is called a derived measure, and calculation or algorithm which combines two or more base measurements is called a measurement function. In NIST document 800-55 [3] 'Derived Measure Specification' is defined as 'Measure, Measure Type'.

E. Indicator Specification

Indicators are the basis for analysis and for making decisions using an analytical model which is algorithm or calculation which combines one or more base and / or derived measurements. In NIST document 800-55 [3] 'Indicator Specification' is defined as 'Formula and Reporting Format'.

F. Decision Criteria Specification

Targets, thresholds, or patterns are taken as the rationale for specific decisions. In NIST document 800-55 [3] 'Decision Criteria Specification' is defined as 'Target'.

G. Measurement Results

Andicator description has to define how indicator has to be interpreted in appropriate reporting formats. In NIST document 800-55 [3] 'Measurement Results' is defined as 'Reporting Format'.

H. Stakeholders

Management are usual clients for resulting measurement reports, which has to be validated by reviewers, who collected data from information owner, by information collector and information communicator. In NIST document 800-55 [3] 'Stakeholders' is defined as 'Responsible Parties'.

I. Frequency / Period

This defines the frequency of data collection, frequency of data analysis, frequency of reporting measurement results, measurement revision, and period of measurement. In NIST document 800-55 [3] 'Frequency' is defined as 'Frequency'.

IV. PROACTIVE INFORMATION SECURITY DEFENCE SYSTEM METRICS

Every action which is done to recover from damage is a reactive action, and every action which can minimize specific negative effects in the future is proactive. The key to proactively resolve potential incidents is detection of events which may jeopardize information security in time to make it proactive, or before it affects security

(confidentiality, integrity, and availability) of a specific information asset.

Using measurement goals to determine the time when a specific event will compromise information security it is possible to determine prior steps (system behaviour, attacker behaviour, etc.) in a timeframe which can lead to proactive incident detection.

Because of this it is crucial to define events which can be used as an early warning of a possible information security incident. These events are possible to determine by conducting a relevant qualitative or quantitative risk assessment process. Below is presented a methodology for determining proactive defence points in a timeframe where proactive actions are possible.

V. METODOLOGY FOR DETERMINING PROACTIVE DEFENSE POINTS

Based on a simple Goal-Question-Metric (GQM) approach [24], [25], our method uses controls from the ISO 27001 standard as a goal or set of goals which are possible to reach by proactive detection of potential violations. The purpose of such controls is to minimize impact and likelihood of risk realisation. As shown in Figure 1. prior to successful intrusion attack there are several points in time before attacker could be stopped.

This 'divide and conquer' approach divides possible detectable events and resolves them if possible, assuming that by resolving specific issues a major issue can be resolved too.

These points could be defined as detection points of control objective violation, when the possibility exists to resolve issues proactively. We therefore name these points *Proactive Defence Points* as PDPs see Figure 3. These points in time could be predefined for any unwanted scenario or for any event which can violate information security controls. This is because every event can be put in a time frame in which a certain amount of time is needed to escalate into a specific unwanted situation.

The PDP (Figure 3.) points once reached could be used to trigger specific proactive action. For specific events relevant time measurement could be in milliseconds and for others, could be hours, days, or even months. The generalised time scale is presented in Figure 3. PDPs are points in time when specific violation is noticed and recorded.

Notifications can be obtained by specific system warning message (syslog, e-mail, SMS) or by humans (operators, external or internal auditors). In the case of events such as cyber attacks DoS or DDoS attack a PDP could be a specific number of incoming SYN packets. In proactive intrusion detection steps could be taken before system abuse or the final stage of system pillage by collecting relevant data from IDS, firewalls and other devices and systems.

Collected data are than used to create useful information that can be analyzed and used for making decisions. Together with PDP it is good practice to have already predefined proactive actions (PA) which will minimize or neutralize impact, as shown in Figure 3. To be able to reach PDPs as goals of GQM method we use questions constructed from controls defined by the relevant standard, and we are thus able to create metrics by defining acceptance criteria related to relevant periods of time for monitoring KPIs on different hierarchical levels.

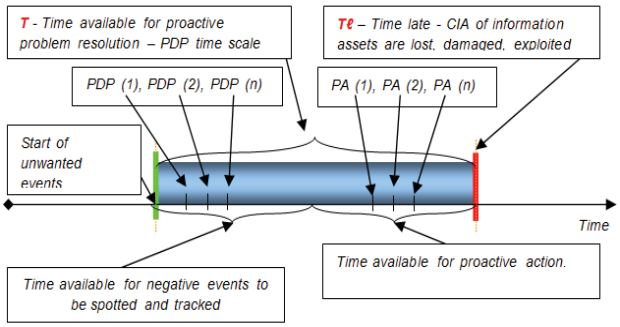


Figure 3. Time scale for proactive detections and actions

VI. PROACTIVE INFORMATION SECURITY MANAGEMENT METRICS FOR SPECIFIC ISO 27001:2005 CONTROLS

Examples of proactive information security management system metrics creation are the controls from ISO 27001:2005 with titles: Audit logging (10.10.1), Controls against malicious code (10.4.1), Security of system documentation (10.7.4), Monitoring system use (10.10.2), Administrator and operator logs (10.10.4) and for which we chose names for measurement methods MM10101, MM141, MM1074, MM10102, MM10104. The measurements are described below.

A. *Audit logging (10.10.1) MM10101*

One of the KPIs for MM10101 could be calculated as total number of log files monitored in specific time interval / number of available log files * 100

Measurement description could be: Value of MM10101 has to be very close to 100 %, and if the value is much lower then there is a risk that specific events could be undetected and unrecognized on time, and confidentiality, integrity, and availability of important system resources could be jeopardized.

Risk example: There is a risk that log information which is collected from devices capable of collecting them is not monitored, with the result that appropriate proactive action is not performed.

Proposal for acceptance criteria (n): 80% <= Acceptable, 79% > Not Acceptable

Risk: Confidentiality, Availability, Integrity

PDP time scale: monthly, quarter of the year

PDP: (1) No procedure with clear responsibilities and system of creating records for audit logging. (2) Not working by the procedure or more precisely not collecting or not monitoring log events. (3) Less than 80% log events were monitored

B. *Controls against malicious code (10.4.1) MM1041*

A number of authors have suggested that it is possible to support ISMS implementation and maintenance with

an incident management system [26], [27]. A KPI for MM1041 could be calculated as number of incidents as results of malware (malicious software) outbreaks on system / number of detected and blocked malware occurrences * 100

Measurement description: Value of MM1041 has to be very close to 100%, and if the value is much higher there is a risk that he organization's data can be deleted, changed or altered, and availability of computer and network infrastructure could be compromised because malware can attack different aspects of digital infrastructure.

Risk example: By not having software and hardware solutions and procedures against malicious code there is a risk that data can be stolen, deleted, or changed by malicious code.

Proposal for acceptance criteria (n): 100 = Acceptable, 0 < Not Acceptable

Risk: Availability, Confidentiality, Integrity, Availability

PDP time scale: monthly, quarter of the year

PDP: (1) No procedure with clear responsibilities and system of creating records of malicious code occurrences. (2) Not working by the procedure or, more precisely, detecting events or situations where anti-malware software components are not used, or no updates are performed with new virus identifications. (3) Number of virus outbreaks is higher than acceptable.

C. Security of system documentation (10.7.4) MM1074

A KPI for MM1074 could be calculated as total number of unauthorized accesses to system documentation / total number of access to system documentation * 100.

Measurement description: Value of MM1074 has to be very close to 0%, and if the value is much lower than there is a risk that confidentiality, integrity and availability of the system documentation could be jeopardized by not applying protection for these documents.

Risk example: System documentation usually contains confidential information related to structure of databases, network topology, and vendor names and versions of devices, operating system versions, applications. This information can be used by attackers to exploit known vulnerabilities against software and hardware components which have details in system documentation.

Proposal for acceptance criteria (n): 0% = Acceptable, 0 %< Not Acceptable

Risk: Confidentiality, Integrity, Availability

PDP time scale: monthly, quarter of the year

PDP: (1) No procedure with clear responsibilities and system of creating records. (2) Not working by the procedure which defines that access to system documentation has to be monitored. (3) Log record which shows that unauthorized access to system documentation was performed.

D. Monitoring system use (10.10.2) MM10102

A KPI for MM10102 could be calculated as number of unauthorized attempts to access file, folders, device attachment, attempts to change security settings.

Measurement description: Value of MM10102 has to be very close to 0 and if the value is much higher than there is a risk for confidentiality and integrity of the critical information assets, since specific data can be altered, deleted, or stolen.

Risk example: There is a risk that somebody may attempt to access specific resources for which they have not formally been given authorization rights.

Proposal for acceptance criteria (n): 0 = Acceptable, 0 < Not Acceptable

Risk: Confidentiality, Integrity, Availability

PDP time scale: monthly, quarter of the year

PDP: (1) No procedure with clear responsibilities and system of creating records. (2) Not working by the procedure or, more precisely, not monitoring for unsuccessful access attempts. (3) Recorded unsuccessful access attempts to specific resource.

E. Administrator and operator logs (10.10.4) MM10104

A KPI for MM10104 could be calculated as total number corrective actions taken after specific event is recorded in log / total number of events recorded in log * 100

Measurement description: Value of MM10104 has to be very close to 100%, and if the value is much lower than that result there is a risk to the confidentiality, integrity and availability of the critical information assets because corrective actions were not taken.

Risk example: There is risk that specific tasks are performed with administrative privileges which can have impact on data or system but and without recording or logging those changes.

Proposal for acceptance criteria (n): 100 = Acceptable, 100 > Not Acceptable

Risk: Confidentiality, Integrity, Availability

PDP time scale: monthly, quarter of the year

PDP: (1) No procedure with clear responsibilities and system of creating records for managing administrator logs. (2) Noticed that task is performed and there are no log records. (3) Using correlated events it is noticed that system has failed after administrator intervention.

VII. CONCLUSIONS

Information security is important for any size of organization or company because of the high importance of the confidentiality, integrity and availability of the organization's information assets. Without information no operations are possible, and with right and timely information many tasks can be performed, and many decisions made. While it is not easy to predict future development in the information security arena, it is clear that there must be continuous improvement in information security standardization and management. Rationale for this is clear: standards and good practices have to be implemented if not for internal efficiency or partner confidence, than because of regulation. However media

reports reveal many cases of neglect of information security management.

In previous work we have shown how KPIs can be visualised in real time [28]. In this paper we have presented a methodology for proactive information security incident resolution based on the GQM approach, and given examples which might help in efforts to manage information security better in any business ecosystem [29]. This is achieved by defining points in time when specific events could be detected, and provide information which can be used to prevent information security incidents.

REFERENCES

[1] Bejtlich R., (2004) The Tao of Network Security Monitoring, Pearson Education Inc. Second printing 2004

[2] Information Technology – Security techniques – Information security management – Measurement, ISO / IEC 27004:2009, 15.12.2009 First edition.

[3] Elizabeth Chew, Marianne Swanson, Kein Stine, Nadya Bartol, Anthony Brown, and Will Robinson, Perforamnce Measurement Guide in Information Security, 2008 Available from: http://csrc.nist.gov/publications/nistpubs/800-55-Rev1/sp800-55-rev1.pdf [accessed on 25.01.2013]

[4] Passive and active attacks, SECURITY CONCEPTS FOR CORPORATE NETWORKS, Available from: http://www.tc11.uni-frankfurt.de/CONF/SEC94/roppl.html [Accessed on 08.05.2008]

[5] Cisco, VPN Services, Available from:

http://www.cisco.com/en/US/netsol/ns341/ns121/ns193/networking_solutions_solution.html, [Accessed on: 07.05.2008]

[6] Toigo Jon William, Diastser Recovery Planning Preparing for the Unthinkable, Prentice Hall PTR 2003.

[7] Syed Aakhtar and Syed Aafsar, Business Continuity Planning Methodology, Sentryx, 2004

[8] ISO 27005:2008 Information technology – Security techniques – Information security risk management, Geneva: ISO, 2008

[9] Stoneburner Gary, Goguen Alice, and Feringa , Risk Management Guide for Information Technology Systems, NIST, 2002 Avaialble at: http://csrc.nist.gov/publications/nistpubs/800-30/sp800-30.pdf [accessed on 24.11.2012]

[10] Calder Alan and Watkins G Steve, Information Security Risk Management for ISO27001 / ISO 17799, IT Governance Publishing, 2007

[11] ISO 27001:2005 Information technology – Security techniques – Information security management system – requirements, Geneva: ISO, 2005

[12] ISO 27002:2005 Information technology – Security techniques – Code of practice for information security management, Geneva: ISO, 2005

[13] Calder Alan and Watkins Steve, (2006) International IT Governance, An executive Guide to ISO 17799 / 27001, Kogan Page 2006

[14] Calder Alan (2006), implementing Information Security based on ISO 27001 / ISO 17799 – A Managament Guide, Van Haren Publishing 2006

[15] Kristian Beckers, Stephan Faßbender, Maritta Heisel, Holger Schmidt, Using Security Requirements Engineering Approaches to Support ISO 27001 Information Security Management Systems Development and Documentation, 2012 Seventh International Conference on Availability, Reliability and Security. pp. 242-248

[16] Wolfgang Boehmer, Cost-benefit trade-off analysis of an ISMS based on ISO 27001, 2009 International Conference on Availability, Reliability and Security. pp. 392-399.

[17] Mohd Asri Mohamad Stambul and Rozilawati Razali, An Assessment Model of Information Security Implementation Levels, 2011 International Conference on Electrical Engineering and Informatics, 17-19 July 2011, Bandung, Indonesia

[18] Raydel Montesino and Stefan Fenz, Information security automation: how far can we go?, 2011 Sixth International Conference on Availability, Reliability and Security. pp 280 – 285

[19] M. Coetzee, Towards a holistic Information Security Governance Framework for SoA, Seventh International Conference on Availability, Reliability and Securtiy, 2012, pp. 155 – 160

[20] M.M. Anwar, M.F. Zafar, Z. Ahmed, A Proposed Preventive Information Security System, ICEE '07. International Conference on Electrical Engineering, Digital Object Identifier: 10.1109/ICEE.2007.4287288

[21] Vasudaven Vinod, Anoop Mangla, Application Security in the ISO 27001 Environement, IT Governance LTD 2008

[22] Architecture and Infrastructure for Governing Information Security in Central Banks, Journal of Central Banking Theory and Practice,Volume 1, Number 2, pp. 5-17,K. Hajdarevic, K. Kozaric, J. Hadzigrahic, Published in October 2012. Available at: http://www.cb-mn.org/slike_i_fajlovi/fajlovi/journal/vol1/no2/architecture_and_infrastructure.pdf [Accessed on 17.01.2013]

[23] Savola R. M. Heinonen P.A, Visualisation and Modeling Tool for Security Metrics and Measurements Management. Information Security South Africa (ISSA) 2011 IEEE, ISBN 978-1-4577-1481-8.

[24] Aversano, Lerina ; Bodhuin, Thierry ; Canfora, Gerardo ; Tortorella, Maria, A framework for measuring business processes based on GQM, International Conference on System Sciences, 2004. Proceedings of the 37th Annual Hawaii

[25] Yixi, Guo ; Hui, Guan, Research on Method of Metrics for Information Ferry in Secret Intranet, International Conference on Intelligent Computation Technology and Automation (ICICTA), 2011

[26] Stefan Metzger, Wolfgang Hommel, Helmut Reiser, Integrated Security Incident Management — Concepts and Real-World Experiences, Sixth International Conference on IT Security Incident Management and IT Forensics, 2011 pp. 107-121

[27] Juhani Anttila, Kari Jussila, Jorma Kajava, Ilkka Kamaja, Integrating ISO/IEC 27001 and other managerial discipline standards with processes of management in organizations, 2012 Seventh International Conference on Availability, Reliability and Security, pp 425-436

[28] Information Security Measurement Infrastructure for KPI Visualization, K. Hajdarevic, C. Pattinson, K. Kozaric, A. Hadzic, MIPRO 35 Proceedings of the 35th International Convention, Publication Year: 2012, Page(s): 1543 – 1548, IEEE Conference Publications.

[29] J. F. Moore, J. F. The Death of Competition: Leadership and Strategy in the Age of Business Ecosystems, New York: HarperBusiness, 1996

Cyber crime and violation of copyright

B. Vukelić, K. Škaron
Polytechnics of Rijeka/Business department, Rijeka, Croatia
Polytechnics of Rijeka/Business department, Rijeka, Croatia
bvukelic@veleri.hr
kskaron@veleri.hr

Abstract - This paper describes cyber crime and violation of copyright. All the copyright laws that have to be obeyed in the real world apply to the virtual world as well. The problem arises when people violate these laws, whether they are aware of it or not. One of the primary causes is the lack of awareness of the existence of copyright law. A group of 110 examinees, of different sex, age, education level and social status took part in a survey for the purposes of this research. The results are shown in three parts – cyber crime in general, intellectual property and copyright.

I. INTRODUCTION

Development and improvement of information and communication technology, together with information availability, have an influence in the developed society. However, there are some negative aspects of these processes, such as misuse of content availability on the Internet. This means that people often feel "free" to download a movie, a photo or any other content just because it is available.

This paper describes the concept of copyright. Copyright arises from the creation of something new, unseen and original in literature, science and art in general. In order to create a copyright, it is necessary to have an author. As a creator, the author is a copyright owner. In this way, author's interests, moral, property and other rights are protected. [1]

This research was conducted in order to show that many people violate copyright laws and therefore break the laws of their country. It doesn't matter whether they are aware of the existence of these laws or not, because if they got caught, they wouldn't avoid sanctions. The violation of copyright can be sanctioned by civil, criminal, misdemeanor and administrative laws. [2][3] The research method in this paper was survey. The method of survey is a procedure based on a questionnaire investigating and collecting data, information, views and opinions on the subject of research. [5] The data were collected by an online survey for the purposes of this questionnaire. The questions were close-ended. The assigned data were analyzed by relevant statistical methods and results were used for interpretation.

Figure 1. Request for removal copyrighted content from Google [4]

The research, which was conducted between the end of 2012 and the beginning of 2013, analyzes people's awareness of computer crime, that is, copyright. The pattern was chosen to represent overall population, that is, people who use Internet on daily bases, for different purposes. This choice was intentional because the objective of this research was to find out what average Internet users know about the topic of this research.

II. RESEARCH

The survey included 110 examinees, of different sex, education level and status; 60.9% of the examinees were female and 30.09% male. (Table I.) Their age was also different: 3.64% of them were younger than 18, 38.18% were aged between 18 and 29, 40% were aged between 30 and 40, 9.09% were aged between 40 and 49, 4.55% were aged between 50 and 59 and 4.55% were older than 60 years of age. (Table II.) When it comes to their education level, the results were also very different. 2.73% of the examinees have lower educational level; 43.64% had secondary education, 20.91% of examinees have a two-year degree and 32.73% have an university degree. There were 3.64% pupils, 09.09% students, 74.55% empolyed, 9.09% unemployed and 3.64% retired examinees.

TABLE I. THE NUMBER AND PERCENTAGE OF EXAMINEE BY GENDER

Gender	Examinee	%
Female	67	60.91%
Male	43	39.09%

TABLE II. THE NUMBER AND PERCENTAGE OF EXAMINEES BY AGE

Age	Examinee	%
Under 18	4	3.64%
18 - 29	42	38.18%
30 - 39	44	40.00%
40 - 49	10	9.09%
50 - 59	5	4.55%
60 and older	5	4.55%

TABLE III. THE NUMBER AND PERCENTAGE OF EXAMINEES ACCORDING TO COMPUTER USAGE FREQUENCY

Answer	Examinee	%
Daily	105	95.45%
Rarely	4	3.64%
Never	1	0.91%

TABLE IV. THE NUMBER AND PERCENTAGE OF EXAMINEES BY REASON FOR INTERNET USAGE

Answer	Examinee	%
Research	104	94.55%
Shopping	48	43.64%
E-mail	97	88.18%
Social networks	76	69.09%
Online banking	63	57.27%
Work	16	14.55%

Tables III. and IV. show the examinees' answers to the questions about using computers and the Internet, as well as the reasons for the usage. It is evident that the majority of examinees use their computer and Internet daily (95.45%). They were offered multiple answers to the question about the reason for using the Internet, and the results show that the majority of the examinees spend the time on the Internet searching for information (94.55%), communicating via e-mail (88.18%) and on social networks (69.09%).

TABLE V. NUMBER AND PERCENTAGE OF EXAMINEES' ANSWERS TO THE QUESTION "ARE YOU ACQUAINTED WITH THE CONCEPT OF CYBER CRIME"?

Answer	Examinee	%
Yes	98	89.09%
No	12	10.91%

TABLE VI. NUMBER AND PERCENTAGE OF EXAMINEES' ANSWER TO THE QUESTION "ARE YOU ACQUAINTED WITH THE CONCEPT OF COPYRIGHT?"

Answer	Examinee	%
Yes	107	97.27%
No	3	2.73%

In the second part of the survey, the examinees were asked about their knowledge of certain concepts of cyber crime.

The second part of the survey also included the questions concerning copyright. The question "Are you acquainted with the concept of copyright?" showed that 107 out of 110 examinees, that is 97.27%, answered "Yes".

But, when it comes to downloading content from the Internet, the majority of the examinees answered that they download images (84.55%), music (80.00%), movies (61.82%) and software (65.45%), as indicated in Figure 2.

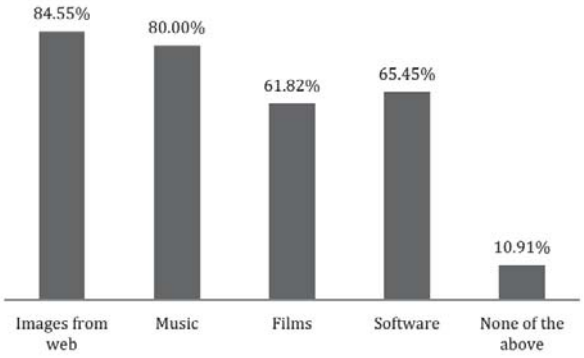


Figure 2. Downloading content from the Internet

The fact that 58% of the examinees answered that they aren't acquainted with copyright laws of the Republic of Croatia indicates that they aren't informed well about legislation.

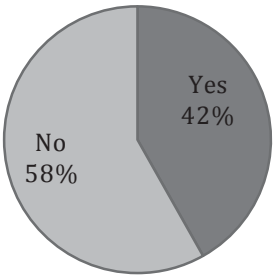


Figure 3. Percentage of examinees who are acquainted with copyright laws

Here, the question was "Have you ever downloaded illegal software?" The answers were the following: 59% of the examinees have, 26% haven't done that and 15% of the examinees aren't sure.

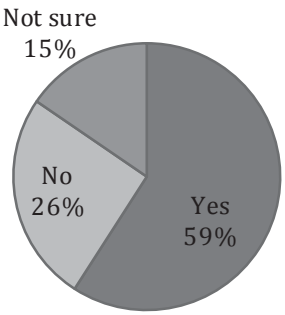


Figure 4. Downloading illegal software

The next question was: "Have you ever downloaded any other copyrighted content, such as images, music or videos?" The answers were the following: 53% of the examinees have downloaded some other kinds of copyrighted content, 29% haven't done that and 18% of the examinees aren't sure.

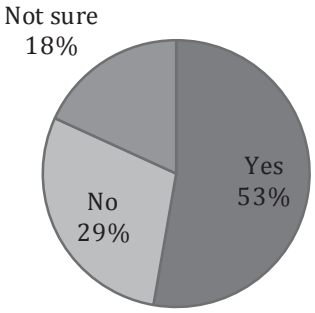


Figure 5. Dowloading other illegal content

Figure 6. shows the answers to the following question: "Have you ever been notified by your Internet provider (T-Com, Iskon) or Internet service (WWW, Youtube, blogs, forums) that you are violating their copyright policies?" The majority of the examinees (83%) answered "No", while only 17% of them answered "Yes".

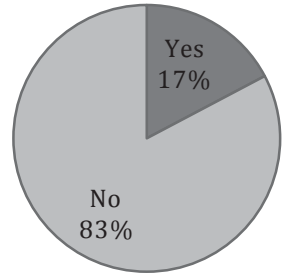


Figure 6. Have you ever been notified by your Internet provider (T-Com, Iskon) or Internet service (WWW, Youtube,blogs, forums) that you are violating their policies or copyright?

The answers to the question "Have you ever used copyrighted content, although you were aware that was illegal?" were the following: 62% of the examinees haven't used content when they were aware it was copyrighted and 38% used it anyway.

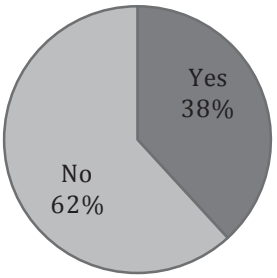


Figure 7. Have you ever used copyrighted content, although you were aware that was illegal?

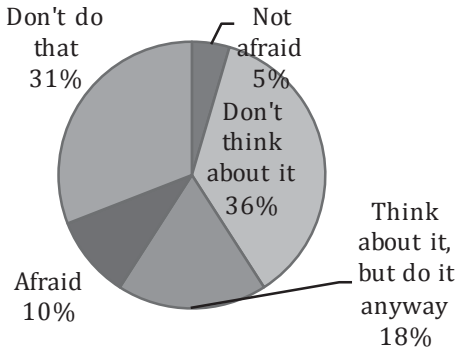


Figure 8. If you are aware of the fact that you're violating laws, are you afraid of legal sanctions?

The question was: "If you are aware of the fact that you're violating laws, are you afraid of legal sanctions?" Only 10% of the examinees answered that they are afraid of possible legal sanctions, while 36% of the examinees answered that they don't think about it and 18% answered that they "Think about it, but do it anyway".

This percentage may be explained by the question that followed: "Have you ever been sanctioned/do you know a person who has been sanctioned for violating copyright?"

The majority of examinees (93%) answered that they have never been sanctioned and that they don't know anyone who has been sanctioned for violating copyright.

The third part of the survey included the questions about violating examinees' copyright laws, such as downloading photos from their own profiles on social networks or personal web page content. Answers to the following question indicate that the majority of examinees don't know what should be done in such cases or that they don't react at all.

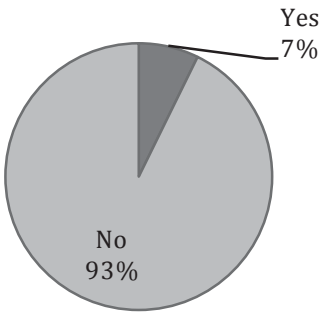


Figure 9. Have you ever been sanctioned/do you know a person who has been sanctioned for violating copyright?

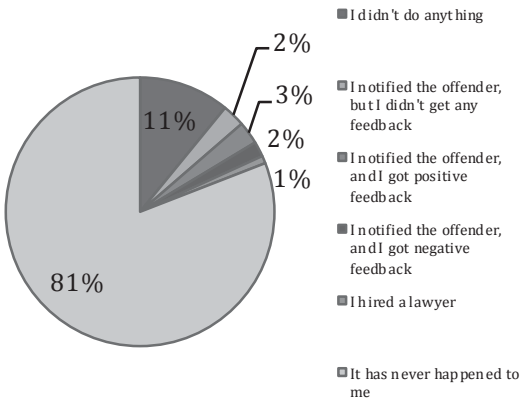


Figure 10. If your copyright had been violated, what did you do about that?

The last two questions were intended to discover whether people will react differently in the future, that is, take some actions if their copyright gets violated.

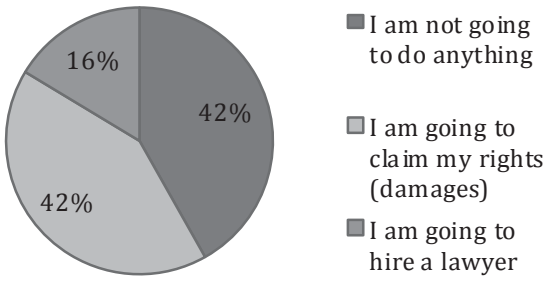


Figure 11. If anyone violates your copyright in the future, what will you do?

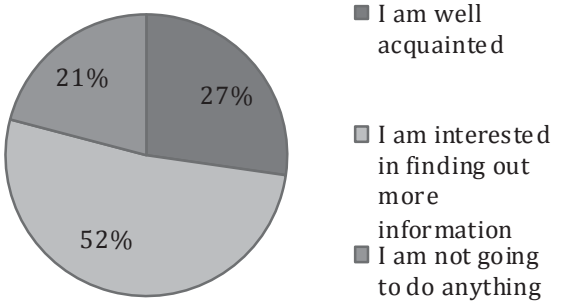


Figure 12. If, prior to filling this questionnaire, you haven't thought of the mentioned regulations, are you going to change that in the future?

III. CONLUSION

Results of the research, which was conducted among examinees of different sex, age, education level and status, showed that the majority of the examinees use the computer and the Internet daily and that they are acquainted with the concept of cyber crime and copyright. They have also been victims of misuse or they have been offenders themselves. Regardless of the fact, more than half of the examinees will search for more information about the abovementioned problems.

There are still some aspects of this topic to research in the future. For example, there is the question of the source of information for the users, that is, where will they find proper information about copyright laws and how should they react when dealing with copyrighted material on the Internet as a global network. In that case, they should be able to find all the necessary information in institutions whose responsibility is raising awareness of the existence of copyright problems.

Also, if there will be sanctions, will they be imposed by the user's or the author's national legislation?

References

[1] Gliha, I., Autorsko pravo: zbirka propisa, uvodni tekst, stvarno kazalo, Informator, Zagreb, 2000.

[2] Kazneni zakon RH, NN 125/11

[3] Zakon o autorskom pravu i srodnim pravima, NN167/03

[4] Google Transparency Report, Google Copyright centar, http://www.google.com/transparencyreport/removals/copyright/

[5] Žugaj M., Dumičić K., Dušak V.:Temelji znanstvenoistraživačkog rada : metodologija i metodika, FOI Varaždin, Varaždin, 2006.

Pravna regulativa kompjuterskog kriminala s osvrtom na Bosnu i Hercegovinu

doc. dr. sc. Feriz Gabela
Dioničko društvo BH Telecom Sarajevo
Obala Kulina bana broj 8Sarajevo
Bosna i Hercegovina
Telefoni: 387 – 33 586 508; 387 – 61 507 530 e-mail: feriz.gabela@bhtelecom.ba

Sažetak – Suvremena informacijska tehnologija, odnosno globalne računarske mreže, stvorile su mogućnosti za nove oblike društveno neprihvatljivog ponašanja. Sve učestaliji načini zloupotrebe kompjutera, potaknuli su znanstvenu i stručnu javnost da se pozabavi ovim oblikom kriminalnog ponašanja. Pod kompjuterskim kriminalom obično se podrazumijeva kriminal koji angažira kompjuter kao sredstvo ili kao cilj izvršenja krivičnih djela. Porast broja krivičnih djela iz ove oblasti posebno se odnosi na one oblike izvršenja krivičnih djela s elementima prekograničnog i internacionalnog kriminala. Unifikacija i efikasna međunarodna suradnja, osnovne su pretpostavke za bolju koordinaciju nadnacionalnih napora za suzbijanje ove vrste krivičnih djela. Kompjuterski kriminal je nedovoljno tretiran u pravnoj regulativi. Relevantna analiza pravne regulative nije moguća bez osvrta na Konvenciju o kibernetičkom kriminalu donijete od strane Savjeta Europe i ukazivanja na njen značaj u globalnim okvirima. U ovom radu izložen je i pregled propisa koji su na pravnoj snazi u Bosni i Hercegovini, a odnose se na suzbijanje kompjuterskog kriminala. Ono što je neophodno u suzbijanju kompjuterskog kriminala jest stabilan pravni sistem s jakom zakonskom regulativom.

Summary - Modern information technology, and global computer networks, have created opportunities for new forms of socially unacceptable behavior. All common types of abuse of computers fostered scientific and professional community to deal with this form of criminal behavior. Computer crime usually involves a crime which uses computer as a mean or a consequence in the execution of a crime. The increase in the number of crimes in this area specifically refers to those forms of criminal offenses with elements of cross-border and international crime. Efficient international cooperation is the first assumption for better coordination of efforts to fight transnational types of crimes. Computer crime is under-treated in the legal regulations. Relevant analysis of legislation is not possible without reference to the Convention on Cybercrime adopted by the European Council and an indication of its importance in the global framework. This paper presents a review of the regulations that are in legal force in Bosnia and Herzegovina, in the fight against cyber crime. What is essential in combating cyber crime is stable legal system with strong legislation.

Ključne riječi - Kompjuterski kriminal, informacijska tehnologija, zloupotreba podataka.

Key words - Computer crime, information technology, misuse of data.

I. UVOD

Kompjuter (računalo) predstavlja jednu od najznačajnijih tekovina razvoja tehničko–tehnološke civilizacije. Pored ogromne koristi čovječanstvu, kompjuter je postao i savršeno sredstvo za izvršenje različitih oblika protupravnih i društveno opasnih djelatnosti, zloupotrebe pojedinaca, grupa i organizacija. [1]

Kompjuterski kriminal obuhvaća nezakonite aktivnosti koje se vrše na kompjuteru ili kod kojih je kompjuter sredstvo izvršenja. U krivičnopravnoj literaturi za kompjuterski kriminal upotrebljavaju se termini kao što su:

1. informatički kriminal,
2. zloupotreba kompjutera,
3. kompjuterska prevara,
4. računarski kriminal,
5. delikti uz pomoć kompjutera.[3]

Informatizacijom društva i ulaskom Interneta u sve oblasti društvenog i privatnog života ljudi, kompjuterski kriminal postaje dominantan oblik kršenja zakona.

Tipovi kompjuterskog kriminala su brojni, a najčešći su:

- neovlašten pristup računalu, računalnoj mreži i elektronskoj obradi podataka,
- sprječavanje i ograničavanje pristupa javnoj računalnoj mreži,
- neovlaštena upotreba računala i računalne mreže,
- oštećenje računarskih programa i podataka,
- pravljenje i unošenje računarskih virusa,
- računarska prevara i sabotaža,
- krađa računarskih servisa,
- djela protiv intelektualnog vlasništva, imovine i pravnog prometa kod kojih se kao objekat ili sredstvo izvršenja krivičnih djela javljaju računari, računarske mreže, računarski podaci, kao i njihovi proizvodi i
- djela gdje se računala javljaju kao sredstvo izvršenja krivičnih djela prevare, kod zloupotreba platnih kartica na Internetu, zloupotreba u oblasti elektronske trgovine i bankarstva, zloupotrebe djece u pornografske svrhe na Internetu (dječja pornografija), govora mržnje na Internetu (širenje nacionalne, rasne i vjerske mržnje).[5]

Osnovne karakteristike kompjuterskog kriminala su:

- društveno opasna, protupravna ponašanja za koja zakon propisuje krivične sankcije i
- poseban način i sredstvo vršenja krivičnih djela od strane ljudi uz pomoć ili posredstvom kompjutera.

Učinioce kompjuterskog kriminala možemo podijeliti na:

- zlonamjerne, koji djeluju radi ostvarenja imovinske koristi ili u cilju nanošenja štete i
- na učinioce koji ne djeluju radi ostvarenja imovinske koristi, – tzv. hakeri, koji koristeći svoje računarsko znanje upadaju neovlašteno u tuđe kompjuterske sisteme.[6]

II. MEĐUNARODNO REGULIRANJE KOMPJUTERSKOG KRIMINALA

A. Reguliranje kompjuterskog kriminala od strane međunarondnih organizacija

Budući da preventivne mjere kojima se društvo suprotstavlja naraslim oblicima i vidovima zloupotrebe kompjutera u različite svrhe nisu dovoljne, logično je da sva suvremena krivična zakonodavstva u sistemu inkriminacija poznaju više kompjuterskih krivičnih djela za koja su propisane različite mjere i vrste krivičnih sankcija. [2] Za međunarodno reguliranje i borbu protiv kompjuterskog kriminala, najznačajniji su: Ujedinjeni narodi, Države članice G-8 i Savjet Europe.

Generalna skupština Ujedinjenih naroda donijela je Rezoluciju o kompjuterskom kriminalu na VIII Kongresu, kojom se od svih država članica Ujedinjenih naroda traži da pojačaju suzbijanje manipulacija s kompjuterima. U Rezoluciji 45/121, državama članicama preporučuje se poboljšanje kompjutorske sigurnosti, pritom vodeći računa o uvažavanju ljudskih prava i temeljnih sloboda i zaštiti privatnosti. [5] Određenim aspektima ili tipovima kompjuterskog kriminala bave se i druge međunarodne organizacije kao što su Svjetska organizacija za intelektualno vlasništvo, Svjetska trgovinska organizacija, Međunarodna privredna komora, Azijsko pacifička organizacija za ekonomsku suradnju i druge. [8]

Najveći doprinos borbi protiv kompjuterskog kriminala dala je grupa osam najrazvijenijih država (G-8) formiranjem tzv. Lionske grupe, nakon samita 1995. godine u Halifaksu. Već 1996. godine u Lionu, ova grupa je predstavila četrdeset preporuka za borbu protiv organiziranog kriminala. Ove preporuke su podržane od strane zvaničnika G-8. Nakon samita u Lionu, formirane su podgrupe koje se bave specifičnim problemima vezanim za organizirani kriminal (kompjuterski kriminal, trgovina ljudima, dokazima u krivičnim procesima, imigracijske prevare i terorizam). Godine 1997. usvojeno je deset principa za borbu protiv kompjuterskog kriminala. Ključni momenti u radu ove grupe su: kreiranje globalnog direktorijuma za kritične informacije (direktorijum je dostupan samo vladinim institucijama država članica), razni dokumenti o najboljoj praksi (uputstva za sigurnost kompjuterskih mreža, internacionalni zahtjevi za podršku, globalno praćenje komunikacija i dr.) i održavanje konferencija za obuku nacionalnih službi sigurnosti, a u vezi s kompjuterskim kriminalom. Ova grupa je odgovorna za preispitivanje mnogih pravnih sistema u cilju lakšeg suočavanja s kompjuterskim kriminalom, kao i za kreiranje principa transnacionalnih pretraga (usvojenu Moskvi 1999. godine). Postojanje ovih principa, kao i brojnih dokumenata koji opisuju najbolju praksu, značajno olakšava borbu protiv globalnog komjuterskog kriminala.[5]

B. Pravno reguliranje kompjuterskog kriminala u Europi

Vijeće Europe je još 1991. godine donijelo Direktivu o zaštiti kompjuterskih programa u kojoj su dane osnovne postavke zaštite kompjuterskih programa. Od država potpisnica se traži da usvoje legislativne mjere kojima bi se sankcioniralo namjerno bespravno pristupanje kompjuterskim sistemima kao cjelini ili nekom njegovom dijelu. Najznačajniji međunarodni akt donijet u okviru Europske unije 1998. godine pod vodstvom Ulricha Siebera sa Sveučilišta u Würzburgu Studija o pravnim aspektima kompjuterskog kriminala u informacijskom društvu *Legal Aspects of Computer-related Crime in the Information Society* – COMCRIME *study*, koja je obuhvaćala i kompjuterski kriminal. Ovaj akt u kombinaciji sa dokumentima s Lisabonskog sastanka 2000. godine, na kojem je istaknut značaj tranzicije u konkurentnu i na znanju zasnovanu ekonomiju, predstavlja smjernice za aktivnosti vezane za razumijevanje kompjuterskog kriminala. Akcijski plan iz iste godine vezan je za sigurnost mreže i uspostavljanja suradnje država članica i njihovog zajedničkog pristupa kompjuterskom kriminalu do 2002. godine. Iste godine donosi se Pravni okvir odlučivanja vezanog za napade na informacijske sisteme (*Proposal for a Council Framework Decision on attacks against information systems*). Godinu dana kasnije dokument je dopunjen s nedozvoljenim pristupom informacijskim sistemima i nedozvoljenim ometanjima sistema.[8]

C. Konvencija o kibernetičkom kriminalu - Convention of Cybercrime Council

Osnovni dokument u europskim okvirima u ovoj oblasti je Europska konvencija o kibernetičkom kriminalu donesena 23. studenog 2001. godine u Budimpešti, pod okriljem Vijeća Europe (u daljem tekstu Konvencija). Ona je prvi cjelovit formalni potpisani i široko prihvaćeni multilaterarni

sporazum usmjeren na probleme kompjuterskog kriminala i jedan od najznačajnijih dokumenata koji su pored europskih država prihvatili i SAD, Japan, Kanada, Južna Afrika. Ova Konvencija spada u krug takozvanih okvirnih konvencija s obzirom na činjenicu da njene odredbe nisu direktno primjenjive, tako da bi ih svaka država trebala implementirati u nacionalno zakonodavstvo. Konvencija je prvi međunarodni ugovor kojim se sankcioniraju kriviična djela i druge vrste kriminala izvršene preko Interneta i drugih kompjuterskih mreža i veoma je značajna za zaštitu intelektualnog vlasništva i dalji razvoj međunarodnog kompjuterskog prava. Jedan od razloga za njeno donošenje, jest uvjerenje da efikasna borba protiv ove vrste kriminala zahtjeva brzu, funkcionalnu i uvećanu međunarodnu suradnju u krivičnim stvarima.

Konvencija je predvidjela četiri grupe krivičnih djela:

1. Krivična djela protiv povjerljivosti kompjuterskih podataka i sistema, i to:

- nezakonito presretanje,
- ometanje podataka i
- zloupotreba uređaja.

Od država potpisnica Konvencije se traži da nacionalnim zakonima, kao krivično djelo okvalificiraju bespravno presretanje prijenosa kompjuterskih podataka koji nisu javne prirode, unutar kompjuterskog sistema, bespravno ometanje kompjuterskih sistema putem brisanja, oštećenja, prikrivanja unošenja kompjuterskih podataka, ukoliko je učinjeno s namjerom radi upotrebe i distribucije kompjuterskih programa u svrhu izvršenja nekog od krivičnih djela.

2. Kompjutersko krivično djelo falsificiranja i prevare koje su u vezi sa kompjuterom.

Od država potpisnica Konvencije se traži da nacionalnim zakonima, kao krivično djelo okvalificiraju unošenje, izmjenu ili brisanje kompjuterskih podataka, u namjeri da oni budu korišteni u legalne svrhe kao da su autentični, kao i nanošenje imovinske štete drugim osobama, kada se to učini bespravnim unošenjem, mijenjanjem, brisanjem ili prikrivanjem kompjuterskih podataka, s namjerom prijevare, da se bespravno pribavi ekonomska dobit za sebe ili za druge osobe.

3. Krivična djela u vezi sadržaja koja se odnose na dječju pornografiju.

Od država potpisnica Konvencije se traži da nacionalnim zakonima, kao krivično djelo okvalificiraju sljedeće radnje, ako se učine namjerno i bespravno:

- proizvodnja dječje pornografije u cilju njene distribucije preko kompjuterskih sistema,
- stavljanje na raspolaganje dječje pornografije preko kompjuterskih sistema,
- prenošenje dječje pornografije preko kompjuterskih sistema,
- pribavljanje dječje pornografije preko kompjuterskog sistema za sebe,
- posjedovanje dječje pornografije u kompjuterskom sistemu za smještanje kompjuterskih podataka.

„Dječja pornografija“ obuhvaća pornografski materijal koji vizualno prikazuje maloljetnike koji sudjeluju u eksplicitnom seksualnom činu ili slike koje predstavljaju maloljetnika koji sudjeluje u seksualnom činu.

4. Krivična djela u vezi napada na intelektualno vlasništvo - kršenje autorskih i njima sličnih prava.

Od država potpisnica Konvencije se traži da nacionalnim zakonima, kao krivično djelo okvalificiraju kršenje autorskih prava definiranih u zakonima te članice koji se odnose na obaveze koje je ona preuzela po Praškom aktu od 24. lipnja 1971. godine, Konvenciji iz Berna o zaštiti književnih i umjetničkih djela, Međunarodnoj konvenciji iz Rima o zaštiti izvođača, proizvođača fonograma i emisijskih organizacija, Sporazumu o komercijalnim pravima na intelektualno vlasništvo, izuzimajući moralna prava sadržana u tim aktima, kada su ta djela učinjena namjerno, da imaju komercijalni karakter, a učinjena su pomoću kompjuterskih sistema. [14]

D. Dodatni protokol uz Konvenciju o kriminalizaciji akta rasizma i ksenofobije počinjenih putem kompjuterskih sistema

Dodatni protokol uz Konvenciju o kriminalizaciji akta rasizma i ksenofobije izvršenih preko kompjuterskih sistema sačinjen je u Strasbourgu 28. siječnja 2003. godine. Formalno-pravno, međutim, obaveze su postale aktualne tek nakon ratifikacije Konvencije i Dodatnog protokola. Dodatni protokol zahtijeva od država potpisnica kriminalizaciju širenja rasističkih i ksenofobnih sadržaja putem kompjuterskih sistema, kao i rasističko i ksenofobno obojene prijetnje i uvrede, te negiranje holokausta i ostalih genocida, i to:

- širenje rasnog i ksenofobnog materijala pomoću kompjuterskog sistema,
- prijetnja motivirana rasizmom i ksenofobijom,
- uvreda motivirana rasizmom ili ksenofobijom i
- poricanje, umanjivanje, opravdanje genocida ili zločina protiv čovječnosti.

Od država potpisnica Konvencije se traži da nacionalnim zakonima, kao krivično djelo okvalificiraju omogućavanje dostupnim javnosti, rasnog i ksenofobnog materijala, pomoću kompjutera, kao i poricanje ili umanjivanje, odobravanje ili opravdanje genocida ili zločina protiv čovječnosti, koje mora biti počinjeno u namjeri da potakne mržnju ili nasilje prema nekom pojedincu ili grupi na temelju rase, boje kože, porijekla, nacionalnog ili etničkog porijekla te vjere, ako se ona koristi kao povod za bilo koje od spomenutog. Naravno većina razvijenih država ima

zakone koji u većoj ili manjoj mjeri sankcioniraju i kompjuterski kriminal. Međutim, u manje razvijenim državama situacija je mnogo nepovoljnija, mada nema pravila. [14]

III. KOMPJUTERSKA KRIVIČNA DJELA U KRIVIČNIM ZAKONIMA BOSNE I HERCEGOVINE

Bosna i Hercegovina je 2006. godine ratificirala Konvenciju. Ta Konvencija za države potpisnice podrazumijeva uspostavljanje nacionalnih zakona o kompjuterskom kriminalu, osiguranje da domaći policijski organi mogu istraživati i krivično goniti kriminalce te osiguravanje međunarodne suradnje s drugim strankama pri istrazi ovog kriminala. Bosna i Hercegovina je potpisala i Dodatni protokol uz Konvenciju 9. veljače 2005., ratificirala ga 19. svibnja 2006. a stupio je na snagu 1. rujna 2006.[1] Nakon stupanja na snagu Konvencije u odnosu na Bosnu i Hercegovinu slijedila je implementacija odredbi krivičnog materijalnog i procesnog prava Konvencije u državni i nedržavna krivične i krivično procesne zakone.

Izmjenama i dopunama Krivičnog zakona Bosne i Hercegovine, u Članu 145a, za izazivanje nacionalne, rasne i vjerske mržnje, razdora i netrpeljivosti, propisana je kazna zatvora od tri mjeseca do tri godine. Nakon usvajanja izmjena i dopuna na nivou Bosne i Hercegovine, Krivični zakoni Federacije Bosne i Hercegovine, Republike Srpske i Brčko distrikta su usklađeni i imaju odredbu o zabrani izazivanja mržnje. [9]

U Članu 4. Zakona o zabrani diskriminacije, gdje se definiraju ostali oblici diskriminacije, zabranjuje se "poticanje na diskriminaciju" koje se precizira kao "svako zagovaranje nacionalne, rasne ili vjerske mržnje." I pored pokušaja usvajanja Zakona o zabrani fašističkih i neofašističkih organizacija i upotrebe njihovih simbola u Bosni i Hercegovini i preporuka CERD-a, koji savjetuje usvajanje ovakvog zakona, to se nije dogodilo.

Područje interneta kao usluge definira Član 7 Zakon o komunikacijama Bosne i Hercegovine, koji propisuje neophodnost posjedovanja dozvole za pružanje internet usluga i utvrđuje obavezu sklapanja ugovora s korisnikom/com. Predstavnici/ce Regulatorne agencije za komunikacije su se u više navrata izjasnili protiv reguliranja sadržaja na internetu i javno negirali da imaju nadležnosti da traže uklanjanje govora mržnje sa BH interneta.

U okviru Državne agencije za istrage i zaštitu (SIPA) ne postoji organizacijska jedinica koja je zadužena samo za borbu protiv kompjuterskog kriminala, ali Kriminalističko-istražno odjeljenje Državne agencije za istrage i zaštitu vrši istrage u vezi s širenjem mržnje preko interneta.

Na inicijativu Tužilaštva Bosne i Hercegovine u okviru sektora Kriminalističke policije u Federalnoj upravi policije postoji Odjeljenje za borbu protiv ekonomskog kriminala, korupcije i pranja novca i kompjuterskog kriminala, dok je Uprava kriminalističke policije Republike Srpske oformila Odjeljenje za sprječavanje visoko tehnološkog kriminala 1. siječnja 2010. godine, a zbog sve većeg broja krivičnih djela u kojima je sredstvo izvršenja računar, a formirana je i Radna grupa za borbu protiv kompjuterskog kriminala kao neformalno radno tijelo. Ova Radna grupa okuplja predstavnike policijskih agencija i pravosudnih agencija i nadležnih ministarstava, a njen je cilj da se utvrde prioritetni zadaci koji se trebaju izvršiti kako bi se kapaciteti u pravosuđu i policiji poboljšali i ojačali za rad na otkrivanju i procesuiranju krivičnih djela iz ove oblasti.

Krivična djela iz oblasti kompjuterskog kriminala regulirana su entitetskim zakonodavstvom, i to:

- Krivičnim zakonom Federacije Bosne i Hercegovine i
- Krivičnim zakonom Republike Srpske.

Kompjuterska krivična djela uvedena su Krivičnim zakonom Federacije BiH, u Glavi 32, gdje su predviđena „krivična djela protiv sistema elektronske obrade podataka" (čl. 393-397).

Krivičnim zakonom Federacije BiH, propisane su sankcije za krivična djela kod oštećenja računarskih podataka i programa, i to:

(1) Tko ošteti, izmijeni, izbriše, uništi ili na drugi način učini neupotrebljivim ili nepristupačnim tuđe računarske podatke ili računarske programe, kaznit će se kaznom zatvora do jedne godine.

(2) Tko neovlašteno pristupi računarskim podacima ili programima ili neovlašteno presreće njihov prijenos, onemogući ili oteža rad ili korištenje računarskog sistema, računarskih podataka ili programa ili računarsku komunikaciju, kaznit će se novčanom kaznom ili kaznom zatvora do tri godine.

(3) Ako je navedno krivično djelo učinjeno u odnosu na računarski sistem ili program organa vlasti, javne službe, javne ustanove ili privrednog društva, kaznit će se kaznom zatvora do pet godina.

(4) Tko neovlašteno izrađuje, prodaje, posjeduje ili čini drugom dostupne sredstva, računarske programe stvorene radi učinjenja krivičnog djela, kaznit će se kaznom zatvora do tri godine.

(5) Posebne naprave, sredstva, računarski programi ili podaci stvoreni, korišteni ili prilagođeni radi učinjenja krivičnih djela, kojima je navedeno krivično djelo učinjeno, oduzet će se.

Prema odredbama Člana 394. računarsko krivotvorenje je:

(1) Tko neovlašteno izradi, unese, izmijeni, izbriše ili učini neupotrebljivim računarske podatke ili programe, s ciljem da se upotrijebe ili sam

[1]Dodatni protokol uz Konvenciju koji se odnosi na kažnjavanje akta rasizma i ksenofobije učinjenih putem kompjuterskih sistema donijet je u Strasbourgu dana 28.1.2003. godine, a ratificiran je Zakonom o potvrđivanju dodatnog protikola uz konvencije o kompjuterskom kriminalu koji se odnosi na inkriminaciju djela rasističke i ksenofobične prirode izvršenih preko računarskih sistema, koji je objavljen u Službenom glasniku Bosne i Hercegovine dana 18.3.2009. godine.

upotrijebi takve podatke ili programe, kaznit će kaznom zatvora do tri godine.
(2) Ako je navedeno krivično djelo učinjeno u odnosu na računarske podatke ili programe organa javne službe, javne ustanove ili privrednog društva, kaznit će se kaznom zatvora od do pet godina.
(3) Tko neovlašteno izrađuje, prodaje, posjeduje ili čini drugom pristupačnim posebne naprave, računarske programe radi učinjenja krivičnog djela, kaznit će se kaznom zatvora do tri godine, a posebne naprave, računarski programi ili podaci, korišteni radi učinjenja krivičnih djela, oduzet će se.

Prema odredbama Člana 395. računarske prijevare su:
(1) Tko neovlašteno unese, ošteti, izmijeni ili prikrije računarski podatak ili na drugi način utječe na ishod elektronske obrade podataka s ciljem da sebi ili drugom pribavi protupravnu imovinsku korist i time drugom prouzrokuje imovinsku štetu, kaznit će se kaznom zatvora do pet godina.
(2) Ako je navedenim krivičnim djelom pribavljena imovinska korist, učinitelj će se kazniti kaznom zatvora od dvije do dvanaest godina.
(3) Tko krivično djelo učini s ciljem da drugog ošteti, kaznit će se kaznom zatvora do tri godine.
Prema odredbama Člana 397. neovlašteni pristup zaštićenom sistemu i mreži elektronske obrade podataka je:
(1) Tko se neovlašteno uključi u sistem ili mrežu elektronske obrade podataka kršenjem mjera zaštite, kaznit će se kaznom zatvora do jedne godine, a ako upotrijebi takve podatake, kaznit će se kaznom zatvora do tri godine.
(2) Ako su navedenim krivičnim djelom prouzrokovane drugom teške posljedice, učinitelj će se kazniti kaznom zatvora do pet godina.
(3) Tko unese, izmijeni, izbriše ili prikrije računarski podatak ili se na drugi način umiješa u računarski sistem, ili uništi ili ošteti elektronsku obradu podataka s ciljem da onemogući elektronsku obradu podataka organima vlasti, javnim ustanovama, privrednim društvima, kaznit će se kaznom zatvora do osam godina. [10]

Krivični zakon RS (2003. g.) predviđa krivično djelo upad u kompjuterski sistem u Glavi 16 A kao „krivična djela protiv sigurnosti računarskih podataka" (čl. 186a-čl. 186e).

U neovlašteno korištenje računara i računarske mreže spada:

- oštećenje računarskih podataka,
- računarska sabotaža,
- izrada i unošenje računarskih virusa,
- računarska prevara i
- neovlašteni pristup zaštićenom računalu, računarskoj mreži, telekomunikacijskoj mreži i elektronskoj obradi podataka. [11]

Analizom sadržaja odredbi državnog i nedržavnih zakona se može zaključiti da se obaveza preuzeta ratifikacijom Konvencije u pogledu krivično metrijalnog prava uglavnom ispoštovala i da su krivična djela predviđena odredbama Konvencije u modificiranoj formi i sadržaju ugrađena u nove odredbe nedržavnih krivičnih zakona u Bosni i Hercegovini. S obzirom na to da zakoni Bosne i Hercegovine nepredviđaju postojanje posebnog krivičnog djela koje bi se moglo podvesti pod govor mržnje na internet, policije dva entiteta i distrikta Brčko ne vode posebne evidencije o broju ovakvih krivičnih djela.

Kompjuterski kriminal u Bosni i Hercegovini je u porastu, a naročito je prisutan kod pristupa zaštićenim računalima i bazama podataka. Kriminalne grupe sve češće koriste mogućnosti novih tehnologija za ostvarivanje međusobne komunikacije. Predviđa se povećanje organiziranog kriminala putem kompjutera i kada je u pitanju dječja pornografija, zlouporaba platnih kartica i slično. To je još jedan od razloga da se Bosna i Hercegovina brzo uključi u europske i svjetske trendove suzbijanja ove vrste neželjenog ponašanja.

IV. ZAKLJUČAK

Ubrzani razvoj kompjuterskih tehnologija i njihova masovna primjena u poslovne i privatne svrhe, uslovili su pojavu nove vrste nedozvoljenog ponašanja – kompjuterskog kriminala. Razmatrajući karakteristike kompjuterskog kriminala, a naročito njegove posljedice, očigledno je da se isti razlikuje od svih drugih oblika kriminalnog djelovanja postepenu opasnosti u odnosu na napadnuta dobra. S obzirom da su načini zloupotrebe kompjuterske tehnologije svakim danom sve složeniji i savršeniji za otkrivanje, i da je vrlo teško ići u korak s tim kriminalnim aktivnostima, potrebno je ulagati napore u to da javnost bude svjesna s kakvim se fenomenom današnje društvo suočava, te je potrebno konstantno raditi na što adekvatnijem odgovoru na različita kriminalna djelovanja u ovoj oblasti. Odlučno suprotstavljanje različitim oblicima kriminalnih aktivnosti su bitne odrednice u težnji da se kompjuterski kriminal, svede u određene, za društvenu zajednicu, podnošljive okvire.

Ohrabruje činjenica da su mnoge države postale svjesne ove pojave i da su u svojim krivičnim zakonima predvidjele pojavne oblike kompjuterskog kriminala kao posebna krivična djela. S druge strane, ohrabrujuće je i to što se u sve većem broju znanstvenih i stručnih radova pažnja posvećuje upravo ovom obliku kriminalnog ponašanja. Na taj se način dolazi do razotkrivanja mnogih specifičnosti kompjuterskog kriminala, a istovremeno se otvara mogućnost suprotstavljanju njegovim oblicima od strane društvene zajednice.
Protiv međunarodnog organiziranog kriminala, s kompjutorskim krivičnim djelima kao njegovom bitnom dijelu, koju posebno karakteriziraju pornografija, krađe tuđih intelektualnih ostvarenja i tuđeg identiteta, krivotvorenja i prijevare, međunarodna zajednica i svaka država trebaju djelovati *nunc aut numquam!* Analiza najvažnijih međunarodnih dokumenata i nacionalnih

zakonodavstava pojedinih država otkriva kakve su promjene nastale u ovoj oblasti. Ratificiranje Konvencije i Dodatnog protokola uz tu Konvenciju koji se odnosi na kriminalizaciju akta rasizma i ksenofobije izvršenih preko kompjuterskih sistema, otvara značajne mogućnosti u tom pravcu. Na osnovu stanja pozitivnog zakonodavstva u Bosni i Hercegovini, a imajući u vidu činjenicu da je kompjuterski kriminal nov oblik kriminala, možemo konstatirati da su država i entiteti usvajanjem odgovarajućih pravnih propisa, omogućili krivični progon počinitelja krivičnih djela kompjuterskog kriminala. S druge strane, neophodno je vrlo ozbiljno analizirati preporuke koje dolaze od stručnih osoba koji su angažirani na poslovima otkrivanja i gonjenja kompjuterskog kriminala. Dakle, može se zaključiti da je potrebno stalno pravno regulirati kompjuterski kriminal i štititi informacijske sisteme, jer će nedostatak adekvatnih propisa značajno umanjiti razvoj informacijskih tehnologija i elektronskog poslovanja. U Bosni i Hercegovini nedostaju propisane procedure za uklanjanje nedozvoljenog ili nelegalnog sadržaja, kao i suradnja civilnog sektora i internet industrije kako bi se uspostavili i uskladili standardi koji bi zabranjivali upotrebu kompjuterskog kriminala za širenje govora mržnje.

LITERATURA

[1] Pavlović, Š.: Kompjutorska kaznena djela u Kaznenom zakoniku, Hrvatski ljetopis za kazneno pravo i praksu (Zagreb), broj 2/2003, str. 625-664.
[2] Britz, M.T.: Computer forensis and cyber crime An introduction, New Jersey, 2004.
[3] Casey, E.: (edit), Handbook of computer crime investigation, forensic tools and technology, Elsevier Academic Press, San Diego, 2004.
[4]Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain aspects of Information Society services, in the Internal Market.
[5] Jerković, R.: „Visokotehnološki kriminal“, Revija za sigurnost, broj 3/2009, str. 27-34.
[6] Nikolić, L.: Suzbijanje visokotehnološkog kriminala, Udruženje tužilaca Srbije, Beograd 2010.
[7] Martin, W: The computer misuse act, The Criminal Review, 1990. godine, str. 767.
[8]. Grubor G.: Osnovi žaštite informacionih sistema,Univerzitet Singidunum, 2006.
[9] Krivični zakon BiH (“Sl. glasnik BiH”, broj: 3/03, 32/03, 37/03, 54/04, 61/04, 30/05, 53/06, 55/06, 32/07, 8/10),
[10] Krivični zakon Brčko Distrikta BiH (“Sl. glasnik Bd BiH”, broj: 10/03, 6/05, 47/11, 52/11),
[11] Krivični zakon Federacije BiH (“Sl. novine Federacije BiH”, broj: 36/03, 37/03, 21/04, 69/04, 18/05, 42/10, 42/11),
[12] Krivični zakon RS (“Sl. glasnik RS”, broj: 49/03, 108/04, 37/06, 70/06, 73/10, 1/12),
[13] Zakon o krivičnom postupku BiH (“Sl. glasnik BiH”, broj: 3/03, 32/03, 36/03, 26/04, 63/04, 13/05, 48/05, 46/06, 76/06, 29/07, 32/07, 53/07, 76/07, 15/08, 58/08, 16/09,93/09),
[14] Zakon o krivičnom postupku Brčko Distrikta BiH (“Sl. glasnik Bd BiH”, broj: 10/03, 48/04, 06/05, 06/05 (2), 12/07,14/07, 21/07),
[15] Zakon o krivičnom postupku Federacije BiH (“Sl. novine FBiH”, broj: 35/03, 37/03, 56/03, 78/04, 38/05, 55/06, 27/07, 53,07, 09/09, 12/10),
[16] Zakon o potvrđivanju Konvencije o cyberkriminalu Vijeća Evrope i Dodatni protokol uz Konvenciju o kibernetičkom kriminalu („Sl. glasnik BiH“, br. 06/06, Međunarodni ugovori).

A. Internet izvori:

[1] http://www.aci.net/kalliste,15.01.2013.
[2] http://www.apisgroup.org/sec.html/,15.01.2013.
[3] http://www.coe.int/cybercrime, 15.01.2013
[4] http://www.coe.int/t/dghl/cooperation,15.01. 2013.
[5] http://conventions.coe.int/Treaty, 16.01.2013.
[6] http://www.crime-research.org, 16.01.2013.
[7] http://www.eur-lex.europa.eu, 16.01.2013.
[8] http://www.nij.gov/nij/pubs-sum, 16.01.2013.
[9] http://www.ncjrs.gov/pdffiles1/nij,17.01.2013.
[10]http://www.justice.gov/cybercrime,17.01.2013. 17.01.2013.
[11] http://megatrender.blog.rs/blog,17.01.2013.

miproBIS

International Conference on

BUSINESS INTELLIGENCE SYSTEMS

Steering Committee

Chairs: Mirta Baranović, University of Zagreb, Croatia
Matteo Golfarelli, University of Bologna, Cesena, Italy
Boris Vrdoljak, University of Zagreb, Croatia
Roberto Sandri, Invenio Consulting Ltd., Zagreb, Croatia

Members: Alberto Abello Gamazo, Universitat Politècnica di Catalunya, Barcelona, Spain
Marko Banek, University of Zagreb, Croatia
Mirta Baranović, University of Zagreb, Croatia
Ladjel Bellatreche, Poitiers University, Futuro Scope Chasseneuil, France
Alfredo Cuzzocrea, University of Calabria, Rende, Italy
Jamel Feki, Multimedia, Information, Systems and Advanced Computing Laboratory, Sfax, Tunisia
Dragan Gamberger, Ruđer Bošković Institute, Zagreb, Croatia
Matteo Golfarelli, University of Bologna, Italy
Igor Mekterović, University of Zagreb, Croatia
Damir Pintar, University of Zagreb, Croatia
Zoran Skočir, University of Zagreb, Croatia
A Min Tjoa, Vienna University of Technology, Austria
Mladen Varga, Faculty of Economics and Business, Zagreb, Croatia
Boris Vrdoljak, University of Zagreb, Croatia

Robert Wrembel, Poznan University of Technology, Poland

INVITED PAPER

The Generalized Physical Design Problem in Data Warehousing Environment: Towards a Generic Cost Model

Ladjel Bellatreche
LIAS/ISAE-ENSMA
Futuroscope, Poitiers, France
bellatreche@ensma.fr

Sebastian Breß
University of Magdeburg
D-39016, Germany
bress@iti.cs.uni-magdeburg.de

Amira Kerkad, Ahcène Boukorca, Cheikh Salmi
LIAS/ISAE-ENSMA
Futuroscope, Poitiers, France
firstname.lastname@lisi.ensma.fr

***Abstract*—Over the years, plenty of cost models for physical database design were developed. However, solving the physical design problem in a generic way is still a hard task. In this paper, we (1) summarize the development stages of cost models for database systems, (2) propose a generic cost model for the generalized physical design problem, and (3) apply our model on the joint query scheduling and buffer management problem.**

I. Introduction

The physical design phase got more attention from the database community when query optimizers became sophisticated enough to cope with *complex decision support queries* running on extremely large database such as *data warehouses* and large scientific databases. During this phase, the database administrator (DBA) has to select optimization techniques such as materialized views, indexes, or partitioning modes [12]. Two types of optimization structure selection are distinguished: (1) *isolated selection*, where the DBA selects only one optimization technique and (2) *multiple selection*, where several techniques may be selected. Several studies concentrated mainly on the isolated selection. In this case, the physical design problem can be formalized using three components: (1) A workload $W = \{Q_1, Q_2, \ldots, Q_m\}$ containing a set of queries, (2) a set of *storage structures* $ST = \{ST_1, \ldots, ST_n\}$ that can be used by the DBA during the physical design, and constraints related to these techniques and hardware (e.g., storage and maintenance cost, or maximal number of partitions). The problem of the physical design consists of two sub-problems: First, selecting a schema of optimization techniques that reduces the cost of executing a workload W; while second, fulfilling the constraints. Several studies showed the hardness of instances of this problem (e.g., materialized view selection problem).

By exploring the literature, we observe that cost models are well present to quantify the cost of solutions proposed by the aforementioned algorithms. This importance of cost models is not new, because they were used in query optimization, where rule based approaches were substituted by cost based approaches. Cost models needed to be developed to estimate the cost of each relational operation as well as whole query plans.

Cost models can also be used to measure the quality of an optimization technique. Several metrics have been considered by cost models such as CPU time, number of input/output (I/O), and network transfer cost. Usually, I/O is the most frequently used metric.

One of the main characteristics of cost models is that they follow the evolution of database technologies. In the *first generation*, cost models were simple, because they estimated only the cost of relation operations. If we view a cost model as a function, in this generation, a cost model had only two components: a query and a database schema. In the *second generation*, these estimations were enriched by considering optimization techniques such as indexes, and materialized views. The inputs of a cost model were: queries, a database schema and optimization structures. *Third generation* cost models included, the deployment architecture of the database such as distributed or parallel database systems, database clusters, or cloud environments. Inputs were: queries, a database, and the nature of the architecture. With the development of storage models dedicated to databases (e.g., column stores), the *fourth generation* proposed cost models including these storage models. Figure 2 summarizes the evolution of cost models. To the best of our knowledge, existing cost models do not consider all layers (database schema, optimization structures, queries, deployment architecture, storage model) as depicted in figure 1. We argue that in order to develop a generic cost model for physical design, all these layers should be included.

Therefore, the contribution of the paper is as follows: (1) We show the development of cost models. (2) We propose a generic cost model for database systems for physical design problems. (3) We formulate all parameters of our generic cost model. (4) We apply our approach to the joint query scheduling and buffer management (QSBM) problem in the context of relational data warehouses to show applicability. Note that our approach can be used for *any* physical design problem, e.g., partitioning, or index selection.

The paper is structures as follows. We introduce our generic cost model in Section II and we instantiate our

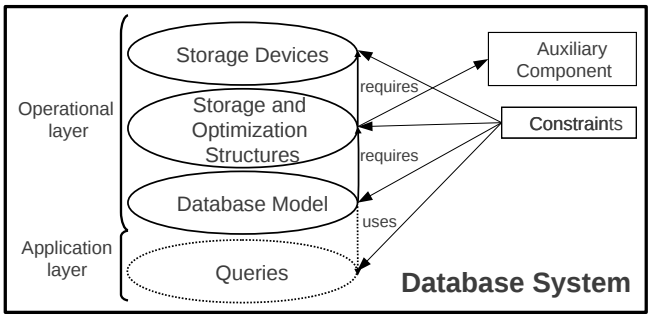


Figure 1. Layers addressed by Cost Models

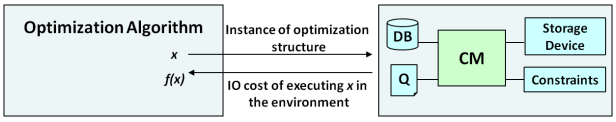


Figure 3. The use of Generic Cost Model in optimization algorithms

generic model for the joint QSBM problem in Section III to prove applicability. Section IV discussed the related works. The paper concludes in Section V.

II. Cost Model Architecture

In this section, we present our generic cost model for the physical design problem, where we highlight the need of use of meta model to represent different components of our cost models.

A. Motivation

To create a generic model for a data warehouse system, we have to take into account several aspects, namely different storage devices, storage structures and database models.

Different Storage Devices: The primary used storage devices nowadays are Hard Disc Drives (HDDs) and Solid State Drives (SSDs). SSDs provide better performance for random access patterns than HDDs, but SSDs are much more expensive [21]. Hence, it is a trade off which storage device is used in the system. Therefore, our meta model has to be able to express characteristics for different storage devices.

Different Storage Structures: The used storage structures have a high impact on database query performance, because they have a large influence on the number of pages that has to be read from disc to process a query. Even plain tables can be stored differently, e.g., row oriented [10] or column oriented [1]. Furthermore, query performance can be accelerated using precomputed results like materialized views [11] or alternative access paths using index structures, e.g., B-tree (one-dimensional) [5] or R/R^*-tree (multi-dimensional) [16], [6]. Each of this storage structures have different performance characteristics. Hence, we have to take them into account in our meta cost model.

Different Database Models/Operation Sets: Nowadays, the most used database model is the relational model introduced by Codd [13]. However, over the years, new database models, and therefore different operations on data sets were developed, e.g., the object-oriented database model [4] or semistructured data models like XML. To cope with this variety of database models, we have to develop a coding for database models in our meta model.

Existing studies are specific to storage devices, storage structures and assume certain database models. Therefore, we need to have a generic model for database systems, in order to create a generic algorithm, which solves the general physical design problem. Our approach is to build a generic cost model, which we will refer to as the *meta model* in the remainder of this paper. The presence of the meta model forces us to develop several types of cost models taking into account the implementation of the storage structures.

B. Layer design meta model

Query: The query is expressed by a formalism which depends closely on the type of database used. However we can represent all queries with different syntax in one meta model. The query has an identifier and a textual description formalism which characterizes the language of the query. The query takes as input a set of concepts which are used to perform a set of relational algebra operations (join, union, etc.) or data manipulation operations (update, insert,etc). An algebra operation can be an unary or a binary function. The result of the query can be restricted by a set of predicates using logical and arithmetic operators or textual operators [27], [22]. The query meta model can easily be extracted from the Common Warehouse Metamodel (www.omg.org/cwm/). Due to the lack of space, we are unable to present all meta models.

Database: We describe an overview of the database meta model. The database is composed of instances which are constituted of tables, indexes, triggers, views and procedures. Tables consist of one or more columns. For each column we must determinate the size, the type of data, the name and the constraints if they exist. The indexes aim at organizing the data of one column on the storage device for minimizing the time of data access. The role of the trigger is the control of data integrity. The stored procedures are a set of SQL statements configured, pre-compiled, stored and executed on demand by the DBMS [25].

Database architecture: There are many deployment architectures of the database as centralized, distributed, parallel or cloud architecture. Centralized architecture has the characteristic that all the data of the database is stored on one storage device. In the distributed architecture, the data is stored in many storage devices in different sites which are connected by the network. When the data is identical in the different sites of the distributed architecture, it is called replicated architecture. The parallel architecture

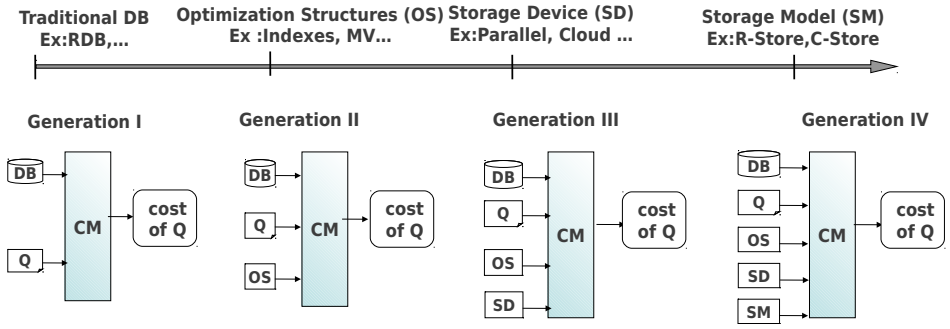


Figure 2. The evolution of Cost Models in Physical Design Problem

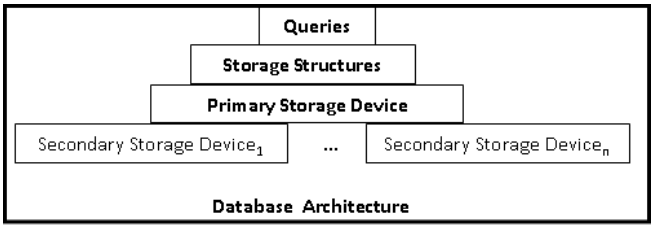


Figure 4. Components of Meta Model

has the same characteristics as the distributed architecture, except that the storage devices are in the same site with many CPU.

Storage Model: There are two storage methods of tables in the storage device, the row-stores and the column-stores. In the first method, the database system stores the table in a sequence of rows, the benefit of this method is easier to insert and to update. In the second method, the database system stores the table in a sequence of columns, the benefit of this method is that any of the column can be as an index and the database system affects only the column to be read during the selection [2].

C. The Generic Cost Model (GCM)

In this section, we propose a generic cost model to describe costs in arbitrary database systems. This means that all existing cost models can be unified in our meta model. To model a database system, one need three components. First, we need the properties of the storage device such as HDD, SSD, main memory, or Cloud Storage. Second, we need to consider the database components, which can be modeled as a set of storage structures $\{ST_1, \ldots, ST_n\}$. Third, we need an abstraction of queries on the database system. A *storage device* is a 5-tuple: SD (L_{read},L_{write},B_{read},B_{write},$size$)

where L_{read}, L_{write} quantify the latency fetching/writing the first page of a data stream ; B_{read}, B_{write} specify the read/write bandwidth of the storage device ; $size$ quantifies the total storage capacity of the device. We differentiate between read and write operations to be able to model storage devices with asymmetric properties, e.g., SSDs.

A *storage structure* ST abstracts how data is managed in a system and is stored on pages $P_1 \ldots P_n$. We denote by $|ST|$ the number of pages of ST.

We define the following functions on a storage structure.

$$rs(ST) = \begin{cases} 1, & \text{if } ST \text{ is row-store} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

$$cs(ST) = \begin{cases} 1, & \text{if } ST \text{ is column-store} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

$$comp(ST) = \begin{cases} 2, & \text{if } ST \text{ is heavy compressed} \\ 1, & \text{if } ST \text{ is lightly compressed} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

The function K(SD, ST) returns the percentage of pages that are dormant in the database buffer located on the primary storage device, e.g., main memory.

The *operation set* OS: $\{O_1, \cdots O_m\}$ is the set of all operations. Each $O_i \in OS$ can be an unary or a binary function. An unary function gets ST_x as input and returns ST_y as output. Accordingly, a binary function gets ST_x, ST_y as input and returns ST_z as output. Hence, a *query* is a tree of operations O_i with $O_i \in OS$.

Since we abstract from storage devices, we need cost functions to estimate the time to fetch a part of a storage structure in the buffer from storage device SD. One possible approach would be to construct analytical cost models for each instance of a (storage devices, storage structure) pair. However, this would introduce high effort in model maintenance and is in contrast to a generic model that can express everything. Hence, we choose a learning based approach to estimate execution times like [9], [3], [20].

We call $T_{read}(ST,SD)$ and $T_{write}(ST,SD)$ the loading and writing time of all pages of the storage structure ST in the storage device SD.

$$T_{read}(ST,SD) = L_{read} + |ST| \cdot page_{size} \cdot B_{read} \quad (4)$$

$$T_{write}(ST,SD) = L_{write} + |ST| \cdot page_{size} \cdot B_{write} \quad (5)$$

here $|ST|$ is the number of page of the storage structure ST. Note that $|ST|$ has to be estimated using storage structure

dependent cost functions. We are now able to formally introduce the Database System (DBS) as a tuple

$$DBS = (\{ST_1, \ldots, ST_n\}, OS, SD, \{SD_1, ..., SD_n\}) \quad (6)$$

where $\{ST_1, \ldots, ST_n\}$ is a set of storage structures, OS is a set of operations, SD is the primary storage device where the database buffer is located and the set $\{SD_1, \ldots, SD_n\}$ represents the secondary storage devices. A primary storage device represents the working memory of a database system and a secondary storage device has a persistent memory, larger storage and slower access speed than a primary storage device. Figure 4 summarizes the different "bricks" involved in our generic cost model. With this representation, we can model different types of databases such as in-memory database, database with raid system, database with shared discs.

For each pair (ST, Q_i) we get a sequence $\{O_1, \ldots, O_m\}$ of operations which may vary depending on the chosen optimizations. In particular, for two different optimization choices, the order in which the operations are executed may differ.

We now discuss the cost formulas to estimate I/O cost. The I/O cost of a query Q is the sum of the I/O cost of each operation $O_i \in Q$.

$$IOCost(Q, ST, SD) = \Sigma_{i=1}^{m} IOCost(O_i, ST_i, SD) \quad (7)$$

where ST_i is a set of storage structures resulting from the $i-1$ previous operations. Possible values of ST_i are

$$ST_i = \begin{cases} \{\emptyset, \{ST_x\} & \text{if } O_i \text{ is unary} \\ \{ST_x, ST_y\} & \text{if } O_i \text{ is binary}\} \end{cases} \quad (8)$$

Finally, we need to estimate the cost of executing operation O_i on the storage structure ST_i. We distinguish between row-stores and column-stores:

$$IOCost(O_i, ST_i, SD) = rs(ST) \cdot IOCost_{row}(O_i, ST_i, SD) + cs(ST) \cdot IOCost_{col}(O_i, ST_i, SD) \quad (9)$$

where $IOCost_{row}(O_i, ST_i, SD)$ and $IOCost_{col}(O_i, ST_i, SD)$ are straightforward from the general $IOCost(O_i, ST_i, SD)$ computation formula. The following formulas compute the number of pages $|ST_i|$ of a storage structure:

$$|ST_i| = \frac{\#rows(ST_i) \cdot avgSize(c_i)}{pagsize} \quad (10)$$

for column stores, where c_i is the column involved in ST_i used by operation O_i, and $avgSize(c_i)$ is the average size of the values in column c_i and

$$|ST_i| = \frac{\#rows(ST_i) \cdot avgSize(r_i)}{pagsize} \quad (11)$$

for row stores, where $avgSize(r_i)$ is the average size of a row r_i.

Symbol	Description
ST	storage structure
$P_i(ST)$	ith page of ST
OS	set of operations $\{O_1, \cdots, O_n\}$
W	set of queries $\{Q_1, \ldots, Q_n\}$
SD	storage device
L_{read}	read latency of storage device
L_{write}	write latency of storage device
B_{read}	read bandwidth of storage device
B_{write}	write bandwidth of storage device
$size$	size of storage structure
CM	cost model

Table I
SUMMARY OF NOTATION

D. Examples

We instantiate our model for two scenarios with two common open source database management systems, namely Postgres [24] and MonetDB [7].

Postgres: In the first scenario, we assume that the system has a HDD as secondary storage device. Postgres supports two main index structures: the B^+ Index and the Hash Index[1]. Since Postgres is a row-store, the function *rs(Table)* is true for all tables and *cs(Table)* is always false. Postgres does support compression[2], so the function *comp(Table)* can either return true or false, but the (default) B^+ Index and Hash Index do not support compression. Therefore:

$$\begin{aligned} Postgres = &(\{Table, HashIndex, B^+Index\}, \\ &\{\sigma, \pi, \bowtie, \cup, \cap, groupby, sort, aggregate\}, \\ &MainMemory, \{HardDiscDrive\}) \end{aligned}$$

with

$$comp(HashIndex) = comp(B^+Index) = 0$$
$$rs(Table) = 1 \text{ and } cs(Table) = 0$$

MonetDB: In the second scenario, we assume that the system has a SSD as secondary storage device. MonetDB is a column store, where each column is stored as a Binary Association Table (BAT) [7]. Since MonetDB support compression techniques, the function $comp(BAT)$ can return either one or zero. Therefore

$$\begin{aligned} MonetDB = &(\{BAT\}, \\ &\{\sigma, \pi, \bowtie, \cup, \cap, groupby, sort, aggregate\}, \\ &MainMemory, \{SolidStateDrive\}\}) \end{aligned}$$

with $rs(BAT) = 0$ and $cs(BAT) = 1$

[1]We are aware of extensions like Gist or GIN. However, adding more index structures, although easily applicable, would increase the example complexity.

[2]http://www.postgresql.org/docs/current/static/storage-toast.html

III. Application of the GCM : BMQS problem

As an application example, we propose to instantiate our GCM in a hard optimization problem in physical design. This problem combines the buffer management problem (*BMP*) with the query scheduling (*QSP*) to optimize workloads. In this section, we will give the formalization of the joint problem, namely *Buffer Management and Query Scheduling (BMQS)* problem, followed by an instantiation of our GCM to deal with BMQS.

A. Formalization of the BMQS problem

In this work, we consider some assumptions: **(1)** prior knowledge of the workload (offline scheduling), **(2)** a centralized $\mathcal{RDW}$ environment and **(3)** the initial cache content is considered as empty.

To represent our workload and facilitate handling queries and buffer objects, we use the representation proposed by Sellis et al. by merging all query plans in one graph called Multi View Processing Plan (*MVPP*) [23].

To facilitate the understanding of the formalization of the combined problem *BMQS*, we start by presenting a separate formalization of both *BMP*and *QSP*.

BMP is formalized as follows: **(1)** Inputs : (i) $\mathcal{RDW}$, (ii) a workload with a set of queries $\mathcal{Q} = \{Q_1, Q_2, ..., Q_n\}$ represented by a MVPP, (iii) a set $\mathcal{N} = \{no_1, no_2, ..., no_l\}$ of intermediate nodes of the MVPP candidates for caching, **(2)** Constraint: a buffer size $\mathcal{B}$ and **(3)** Output : a buffer management strategy $\mathcal{BM}$ that allocates nodes in the buffer to optimize the cost of processing $\mathcal{Q}$.

QSP is formalized as follows: **(1)** Inputs : (i) $\mathcal{RDW}$, (ii) a workload with a set of queries $\mathcal{Q} = \{Q_1, Q_2, ..., Q_n\}$, (iii) a buffer management strategy $\mathcal{B}M$. **(2)** Output : scheduled queries $\mathcal{QS} = \{SQ_1, SQ_2, \ldots, SQ_n\}$.

BMQS is described based on the above formalizations as follows:

- Inputs : (i) A $\mathcal{RDW}$, (ii) a set of queries $\mathcal{Q} = \{Q_1, Q_2, ..., Q_n\}$ represented by a MVPP, (iii) a set $\mathcal{N} = \{no_1, no_2, ..., no_l\}$ of intermediate nodes of the MVPP candidates for caching ;
- Constraint: a buffer size $\mathcal{B}$;
- Output : (i) scheduled queries $\mathcal{SQ} = \{SQ_1, SQ_2, ..., SQ_n\}$ and (ii) a buffer management strategy $\mathcal{B}M$, minimizing the overall processing cost of Q.

The *BMQS* problem is NP-hard [14], [17] and impact seriously the performance of overall data warehouse systems. As a consequence, the development of efficient solutions becomes a crucial issue. This efficiency is measured by a cost model, including different layers.

B. Used Cost Model

In order to instantiate our database system (DBS) in the proposed meta model, we describe each component of our system. (1) The storage structure ST used (database component) is relational tables with Row-Oriented Storage. We ignore indexes and materialized views, and suppose there is no data compression. (2) The operation set OS is the set of algebraic operations, e.g projection, selection, join, aggregation, sort, union, which represents the query component in our GCM. (3) The Primary storage device is the main memory where the buffer lies, and the secondary storage device is the Hard Disc Drive. Therefore, our DBS is modeled as :

$$DBS = (\{Tables\}, \{\sigma, \pi, \bowtie, \cup, \cap, groupby, sort, aggregate\}, MainMemory, \{HardDiscDrive\})$$

$$\begin{aligned} comp(Tables) &= 0 \\ rs(Tables) &= 1 \\ cs(Tables) &= 0 \end{aligned}$$

To instantiate our cost model, we have to estimate the size of intermediate results. However, the number of pages to fetch from disc depends on the buffer content. Therefore, we need to check whether some results (pages) already exist in the buffer. In case some pages are in the buffer, they may be read with much lower latency and higher bandwidth, making T_{read} and T_{write} negligible. Thus, if R_i pages are dormant in the buffer then $T_{read} = T_{write} = 0$. Otherwise, the number of fetched pages is estimated as in [17].

C. Queen-Bee algorithm

As we said before, the joint problem of BMQS is NP-hard. Existing techniques are based on greedy algorithms, costly heuristics etc. Other faster algorithms, which are affinity based, exist but are less efficient. To get a trade-off between speed and efficiency, we propose a divide and conquer algorithm based on queries interaction.

The idea is to group queries by affinity, and to run locally (inside each group of queries) the scheduling and the buffer management strategy. We call these groups "hives", and inside each hive, a query is chosen to be executed first.

This elected query is the "Queen-Bee" of its hive, and it allows filling the buffer with objects that satisfies queries in the same hive. Remaining queries are executed in an order depending on the buffer content. The choice of the queen-bee is done by minimal cost. The next queries of the same hive are ordered by minimal cost considering buffer content.

The algorithm takes a workload to be optimized represented by the merged query plans ($MVPP$) [26]- and starts by generating the query graph with connected components (QGCC), where connected vertices are interacting queries, and each vertex is tagged by its execution cost. The edges are tagged by the number of shared operations (Figure 5)

An optional step is sorting the components, depending on whether queries have priority or not, or if some quality-of-service constraints are given to avoid making some queries

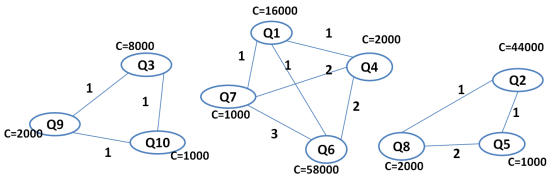


Figure 5. An example of QGCC

wait a long time. The final step schedules queries inside each component by traversing its queries. Contrary to the scheduling strategy proposed in [8], called Dynamic Query Scheduler (DQS), that takes into account only the cache content, our scheduler considers other parameters regarding the query execution cost, nodes frequency etc.

The cost model is the kernel of the algorithm because it allows to decide on the order of queries.

D. Experimental Study

To show the efficiency of the Queen-Bee algorithm, an experiment is done on Star Schema Benchmark of 100 GB and 30 queries. An instance of the GCM adapted to the dataset and the environment is used.

Our algorithm is compared with 5 baselines : (1) No cache representing initial cost; (2) LRU policy; (3) combining LRU with DQS; (4) a dynamic buffer management policy (DBM); and finally (5) a genetic algorithm using DBM and DQS. Varying Buffer size gives different I/O costs. The results show the high efficiency of the Queen-Bee compared to simplistic and greedy algorithms, as depicted in Figure 6.

To validate the simulation results, the same dataset is considered on a Server (32GB of RAM, 2x2.4GHz of CPU). The used DBMS is Oracle11g and the buffer size is set at 10GB. Figure 7 shows the real execution cost in seconds. In addition to the efficiency of Queen-Bee, the validation shows the similarity between simulation cost estimated by our cost model, and the real cost. This proves the quality of our GCM which is able to consider different environment parameters.

IV. Related Work

The cost models are considered an important part of query optimizers and physical design selection algorithms. With the spectacular interests that the database community gave to the physical design in the 90's, developing algorithms for selecting optimization techniques became a necessity. Note that physical design represents one of the hardest problems for database management systems [18]. To satisfy this requirement, several studies proposed algorithms dealing with one instance of the physical design with simple cost models. They take into account database parameters (tables' size, attributes' length, etc.) and disk parameters (e.g. disk page size) [26], [15]. These cost models have been extended

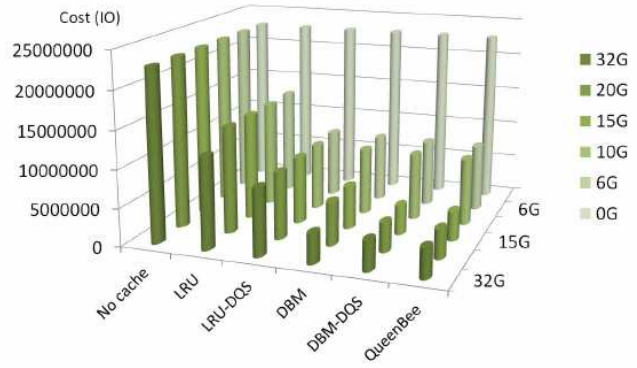


Figure 6. Simulation results using the cost model

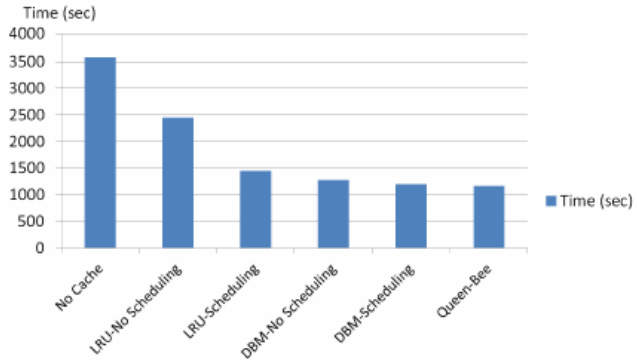


Figure 7. Real costs obtained by deploying simulation results on a DBMS

by taking into account the different deployment architectures of database applications (distributed, parallel, database cluster, and recently cloud). Afterwards, with development of advanced devices such as flash, researchers proposed cost models including the characteristics of these devices in their components. Some research studies considered other auxiliary aspects related to the DBMS components, such as buffer management. Manegold et al. propose a framework [19], which allows to automatically create cost functions of database operations for each layer of the memory hierarchy by combining the memory access patterns of database operations. Our approach is designed to be able to address all physical design problems, whereas the approach of Manegold et al. is tailor made for estimating database operations cost during query processing in a generic way. Hence, they do not focus on the general physical design problem, as we do. One of the major differences despite that, is that Manegold et al. differentiate between sequential and random access latency and bandwidth, whereas we consider different values for read/write latency and bandwidth to model storage devices. We see a possible link of our approach and of Manegold et al. for future work, because our approaches can complement one another, e.g., the framework of Manegold et al. could deliver cost metrics for database operations on

a system, whereas our model can utilize this cost metrics to build our generic cost model for a given database system.

V. Conclusions and Outlook

In this paper, we showed the great impact of cost models along the different generations of databases. Usually, they are largely used in the context of the physical design phase of extremely large databases. Recall that physical design represents one of the hardest problems for database management systems. This importance was amplified by the birth of decision support applications. In this paper, we establish a clear discussion about the cost model evolution and the different layers that have to be considered in their development. We identify four layers : (1) Queries, (2) Database, (3) Database Architecture, and (4) Storage Model. Based on this discussion, we propose a generic cost model and provide an example of its instantiation. To make the connection between our cost model and the physical design problem, we consider one exiting problem by combining two well known NP-hard optimization problems, namely buffer management and query scheduling. This joint problem is solved by using an algorithm inspired from bee life and using our generic cost model. A theoretical validation using our mathematical cost model is given and the obtained results are validated on Oracle11g. The results are encouraging.

References

[1] D. J. Abadi, P. A. Boncz, and S. Harizopoulos. Column-oriented Database Systems. In *VLDB*, pages 1664–1665. VLDB Endowment, 2009.

[2] D. J. Abadi, S. R. Madden, and N. Hachem. Column-stores vs. row-stores: how different are they really? In *SIGMOD*, pages 967–980. ACM, 2008.

[3] M. Akdere and U. Çetintemel. Learning-based Query Performance Modeling and Prediction. In *ICDE*, pages 390–401. IEEE, 2012.

[4] M. P. Atkinson, F. Bancilhon, D. J. DeWitt, K. R. Dittrich, D. Maier, and S. B. Zdonik. The Object-Oriented Database System Manifesto. In *SIGMOD*, pages 395–408. ACM, 1990.

[5] R. Bayer and E. McCreight. Organization and Maintenance of Large Ordered Indices. In *SIGFIDET*, pages 107–141, 1970.

[6] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The R*-tree: An Efficient and Robust Access Method for Points and Rectangles. In *SIGMOD*, pages 322–331. ACM, 1990.

[7] P. Boncz and M. L. Kersten. Monet: An Impressionist Sketch of an Advanced Database System. In *BIWIT Workshop*, pages 240–251. Computer Society Press, 1995.

[8] L. Bouganim, F. Fabret, P. Valduriez, and C. Mohan. Dynamic query scheduling in data integration systems. In *ICDE*, pages 425–435. IEEE, 2000.

[9] S. Breß, F. Beier, H. Rauhe, E. Schallehn, K.-U. Sattler, and G. Saake. Automatic Selection of Processing Units for Coprocessing in Databases. In *ADBIS*, pages 57–70.

[10] D. D. Chamberlin, M. M. Astrahan, and M. W. e. a. Blasgen. A History and Evaluation of System R. In *Commun. ACM*, volume 24, pages 632–646. ACM, 1981.

[11] S. Chaudhuri, R. Krishnamurthy, S. Potamianos, and K. Shim. Optimizing Queries With Materialized Views. In *ICDE*, pages 190–200. IEEE, 1995.

[12] S. Chaudhuri and V. R. Narasayya. Self-Tuning Database Systems: A Decade of Progress. In *VLDB*, pages 3–14. VLDB Endowment, 2007.

[13] E. F. Codd. A Relational Model of Data for Large Shared Data Banks. *Commun. ACM*, 13(6):377–387, 1970.

[14] A. Diwan, S. Sudarshan, and D. Thomas. Scheduling and Caching in Multi-Query Optimization. In *COMAD*, pages 150–153. Computer Society of India, 2006.

[15] H. Gupta. *Selection and maintenance of views in a data warehouse*. Thesis, Stanford University, 1999.

[16] A. Guttman. R-Trees - A Dynamic Index Structure for Spatial Searching. In *SIGMOD*, pages 47–57. ACM, 1984.

[17] A. Kerkad, L. Bellatreche, and D. Geniet. Queen-bee: Query interaction-aware for buffer allocation and scheduling problem. In *DaWaK*, pages 156–167, 2012.

[18] W. Labio, D. Quass, and B. Adelberg. Physical database design for data warehouses. In *ICDE*, pages 277–288, 1997.

[19] S. Manegold, P. Boncz, and M. L. Kersten. Generic Database Cost Models for Hierarchical Memory Systems. In *VLDB*, pages 191–202. VLDB Endowment, 2002.

[20] A. Matsunaga and J. A. B. Fortes. On the Use of Machine Learning to Predict the Time and Resources Consumed by Applications. In *CCGRID*, pages 495–504. IEEE, 2010.

[21] M. Polte, J. Simsa, and G. Gibson. Comparing Performance of Solid State Devices and Mechanical Disks. In *PDSW*, pages 1–7, 2008.

[22] M. A. Roth, H. F. Korth, and D. S. Batory. Sql/nf: A query language for 1nf relational databases. *Information Systems*, 12(1):99 – 114, 1987.

[23] T. K. Sellis. Multiple-query optimization. *ACM Trans. Database Syst.*, 13(1):23–52, 1988.

[24] M. Stonebraker and L. A. Rowe. The Design of POSTGRES. In *SIGMOD*, pages 340–355. ACM, 1986.

[25] T. J. Teorey, S. S. Lightstone, T. Nadeau, and H. Jagadish. *Database Modeling and Design:Logical Design*. University of Michigan, fifth edition, 2011.

[26] J. Yang, K. Karlapalem, and Q. Li. Algorithms for materialized view design in data warehousing environment. In *VLDB*, pages 136–145, 1997.

[27] M. M. Zloof. Query-by-example: A data base language. *IBM Systems Journal*, 16(4):324 –343, 1977.

PAPERS

Analysis of World Bank Indicators for Countries with Banking Crises by Subgroup Discovery Induction

Dragan Gamberger, Dražen Lučanin and Tomislav Šmuc
Rudjer Boskovic Institute, Zagreb, Croatia
dragan.gamberger@irb.hr, drazen.lucanin@gmail.com, tomislav.smuc@irb.hr

Abstract – The paper presents results of the analysis of countries that experienced banking crises in the period 1976-2011. The analysed data are World Bank indicators in the periods before the onset of the crises. The work is done as part of the EU FP7 project "Forecasting Financial Crises". For the analysis we have used the subgroup discovery methodology, which is especially useful when different subtypes of the main concept are anticipated and when different, possibly even contradictory properties may characterize these subtypes. We describe the data preparation process and the Data Mining Server application for subgroup discovery induction. The applied methodology may also be useful for other applications in business and marketing domains. Our main result is the classification of five subsets of countries with banking crises. Three of them may be recognized as financially driven types of banking crises, while significant socio-economic problems are characteristic for the remaining two subgroups.

I. INTRODUCTION

In systemic banking crises, multiple banks in a country fail simultaneously and the effects on the country's economy may be significant. In the period following the year 2007, a significant number of crises occurred in developed countries – especially in the EU. This wave of crises has stimulated research aiming to better understand the causes of these crises and to give recommendations to policy makers in order to prevent future crises. The work presented in this paper forms one part of the research done within the EU FP7 project "Forecasting Financial Crises"[1] (FOC). The main goal of the project is to achieve a better understanding of connections among banks and the detection of connection patterns that can lead to instabilities in financial networks [1].

Rudjer Boskovic Institute is a participating institution in this project. A part of our task is to evaluate and better understand general social and macroeconomic conditions that are characteristic for countries in a period before the outburst of crises. The data used for the analysis is publicly available on the World Bank indicators portal[2].

Detecting novel and potentially useful relations in data is a typical data analysis task. In the domain of banking crises, it involves the identification of potentially complex relations among indicators that characterize countries in the period before crises in a form that can be easily interpreted by domain experts and used either for understanding of causes of crises or used for the identification of early warning signals. Induction of subgroups is a rule-learning methodology that is especially appropriate for such descriptive data analysis tasks [2]. The methodology has already been successfully applied in medical and scientific applications, while financial and marketing use cases are still rare. In contrast to standard statistical data analysis, the main advantage of the approach is effective segmentation of the analysed domain. This is especially relevant in tasks in which it may be expected that the analysed concept, like banking crises, may have very different origins and characteristics. For example, it is known that excessive monetary activity of a central bank is a risk factor for banking crises, but it can also be expected that insufficient availability of credits may reduce economic activity and finally cause a banking crisis in a country.

The main goal of the paper is the presentation of some of the results obtained on the FOC project by using the subgroup discovery methodology. Additionally, the intention is to illustrate how the methodology may be used also in other domains and applications. Section II presents the data preparation process while Section III introduces the Data Mining Server as a simple and useful tool for subgroup induction tasks. The main part of the paper is Section IV, which presents five subgroups constructed for the banking crises domain. Their preliminary expert evaluation is in given in Section V.

II. DATA

The results presented in this paper are based on the list of banking crises reported in [3]. In total, 147 crises in the period 1976-2011 have been described in this document and they have been used as the positive cases for our analysis. 29 out of 147 crises are in the period 2008-2011. Examples of crisis cases are China in year 1998 and USA in years 1988 and 2007. As a control group we have used 287 cases. They are selected for countries that have experienced banking crises, but in a way that the observed years are at least 10 years far from those crises. Examples of non-crises are Finland in years 1971, 1981, and 2001 and UK in years 1977, 1987, and 1997, because Finland had a crisis in the year 1991 and

[1] http://www.focproject.eu/
[2] http://data.worldbank.org/indicator

UK in the year 2007. A period of 10 years without crisis is assumed as long enough to demonstrate that a country is a good representative of a non-crisis case.

The crises and non-crises cases have been described using 105 World Bank indicators. At first we included 5 indicators suggested as potentially relevant in an International Monetary Fund study [3]. They are: current account balance as percentage of GDP, central government debt as percentage of GDP, domestic credit to private sector as percentage of GDP, foreign direct investments as percentage of GDP, and bank capital to assets ratio. Aside from these, we have also included 100 other indicators from the World Bank database. The included indicators cover various domains, including – economic policy, health, agriculture, and gender data. From each field we tried to select a few of the most representative indicators. An additional criterion has been the decision to use only relative indicators that do not need normalization in order to be comparable between countries of different size. Examples of selected indicators are: percentage of rural population, life expectancy at birth, percentage of unemployed people with tertiary education, and research and development expenditures as percentage of GDP.

In order to study predictive connections among indicator values and the occurrences of crises, the indicator values in the period 1 to 3 years before crisis and non-crisis have been used. The values are denoted as t_1, t_2, and t_3 data. From these short data sequences we have computed 6 additional values representing slope, mean value, minimal value, position of the minimal value, maximal value, and position of the maximal value. The result is a dataset consisting of 147 positive examples and 287 negative examples so that each of these examples is described with 945 (105 times 9) numerical attributes. Some of these attributes have unknown values. The dataset is public and is available for download from http://dms1.irb.hr/ from the page with prepared datasets.

In the dataset used in our previous work [4] we have used as negative cases the countries that have never experienced banking crises and the countries that have experienced crises, but so that negative cases are at least 10 years far from the period of crises. Due to the fact that in the period 2008-2011 crises have been experienced in many developed countries and that this period is the most interesting for the analysis, it has been decided that negative cases can be only from countries that have experienced banking crises. The intention behind such a definition of negative cases is to avoid the detection of differences between countries experiencing crises and those that never had a banking crisis, especially to avoid detection of differences between developed economies in which crises are relative often and undeveloped economies in which banking crises are rare.

III. METHODOLOGY

Subgroup Discovery (SD) was first introduced as a methodology by Kloesgen [5] and Wrobel [6]. SD techniques aim to discover distinct subsets of the population that are statistically unique or interesting and at the same time as large as possible. The goal is the induction of human-interpretable descriptions of subgroups. The input is a set of cases consisting of a group of positive cases P (countries experiencing banking crises in a specific year) and a control group of negative cases N (countries in a period of no banking crises). The subgroup discovery algorithm constructs rules that are true for positive cases and false for negative cases. It is not necessary that rules are true for all positive cases and false for all negative cases, but the intention is to find short rules that are true for large subsets of positive cases and at the same time that are false for large subsets of negative cases. Subgroup sizes are not defined in advance but the algorithm tends to make them as large as possible.

The subgroup discovery methodology based on ILLM (Inductive Learning by Logic Minimization) system has been implemented at the Rudjer Boskovic Institute, Zagreb, Croatia. At http://dms1.irb.hr there is a publically available Data Mining Server that may be used for subgroup discovery tasks on user-submitted data. The server presents a very simple and user-friendly interface to the data analysis process, but it is limited to 1000 cases and 1000 descriptors for the sake of public availability. Fig. 1 presents the main page of the Server. From this page the user has access to the instructions about the format in which the data has to be prepared and to the instructions on how to start the induction process.

Fig. 2 presents the data upload page, which is used to select two parameters and to start the induction. It is assumed that the user has prepared the data file on his computer and the first step is to select the name of this file.

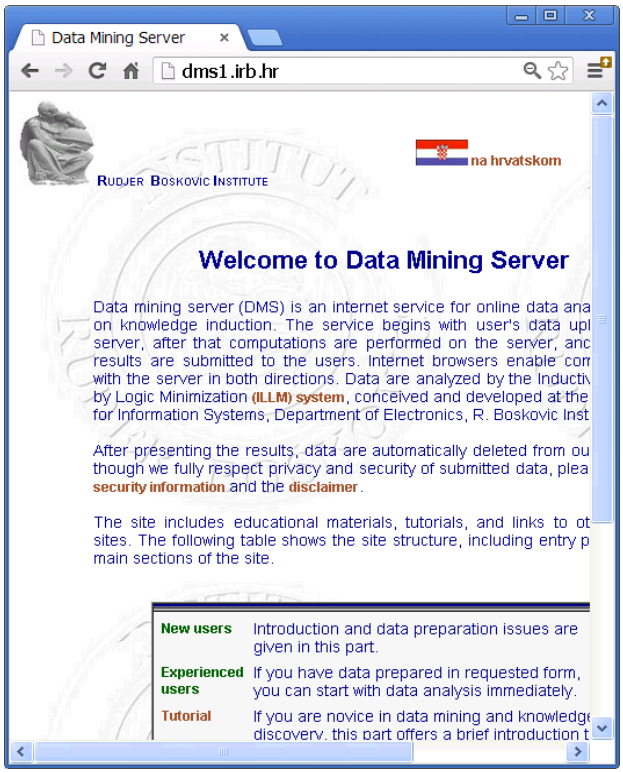


Figure 1. The Data Mining Server for subgroup discovery tasks.

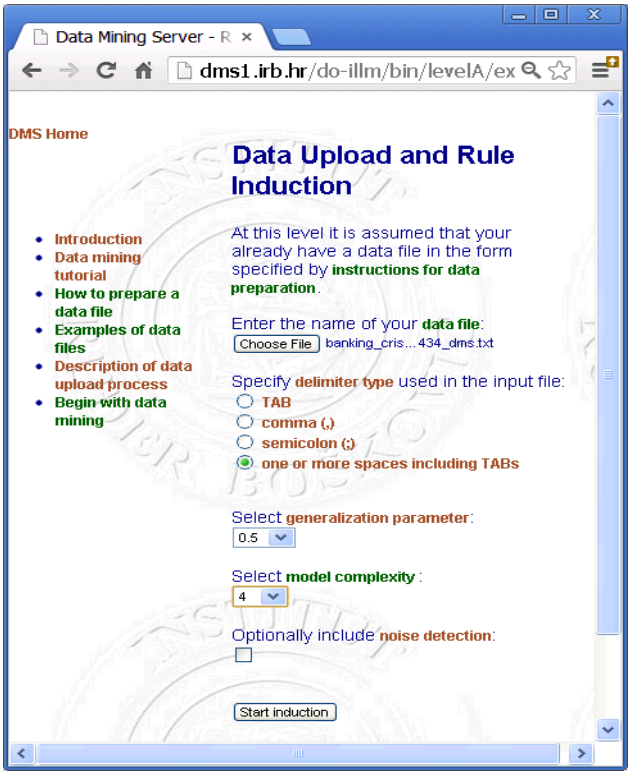


Figure 2. Data Mining Server: The main page for data upload and selection of parameters for the expected generality and complexity of induced subgroups.

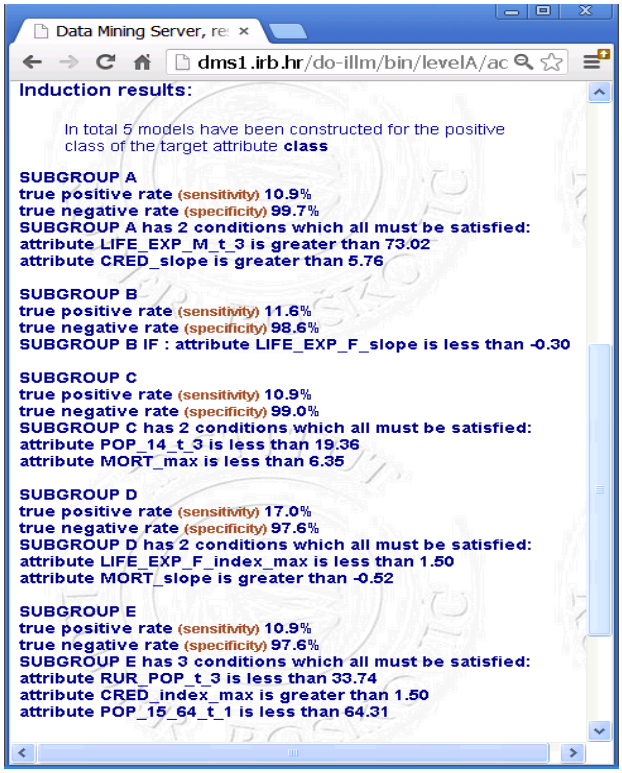


Figure 3. Data Mining Server: Results for banking crises domain.

The second step is the selection of the generalization parameter, which can be in the range 0.1-100, and the complexity parameter, which can be in the range 1-10. For each domain it is strongly suggested to experiment with different parameter values.

The induction starts by pressing the "Start induction" button. Fig 3 presents the solution obtained for the banking crises dataset with parameters 0.5 for the generality and 4 for the complexity. The server needs around 15 minutes to compute the solution. The parameter values have been selected so that the induced subgroups have very high specificity (true negative rate) and that rules are short in order to enable easy expert analysis.

In the next section the details of the induced subgroups are presented. The order of their presentation has been changed with the intention of grouping financially and socio-economic driven subgroups. Also, with the intention of facilitating expert evaluation, some conditions have been modified. For example, the condition life expectancy for males > 73.0 in Subgroup A in Fig. 3 has been substituted with life expectancy for females > 80.2 in Subgroup 1 which is based on Subgroup A. The condition population below 14 years < 19.7% in Subgroup C has been substituted with condition population ages 65 and more >14.2% in Subgroup 2, which is constructed from Subgroup C. The new form of conditions is the result of a statistical and logical analysis of constructed Subgroups A-E.

IV. SUBGROUPS OF CRISES

The subgroups are defined with a list of included positive cases. Properties of the subgroups are defined with necessary conditions that must be satisfied for a positive case to be included into the subgroup, and a list of supporting conditions that are satisfied by the majority of the included cases. Supporting conditions are selected so that they are statistically most significantly different between the cases included into the subgroup and the complete set of 287 non-crises cases.

According to the preliminary expert evaluation of detected conditions each, subgroup received a name.

Subgroup 1: *Fast growing credit activity in economies with an aging population.*

List of 16 included crises: Sweden in the year 1991, USA and UK in year 2007, Belgium, Denmark, France, Greece, Ireland, Island, Italy, Luxemburg, Netherlands, Portugal, Slovenia, Spain and Sweden in the year 2008.

The necessary conditions:

- The slope of credits in the period of three years before the crisis > 5.8 % per year
- Life expectancy for females three years before the crisis > 80.2 years.

The supporting conditions are: low mortality of children, low percentage of young population, high

percentage of old population, high companies' capitalization.

Subgroup 2: *High credit activity in economies with high social security.*

List of 16 included crises: Japan in the year 1997, UK in the year 2007, Austria, Belgium, Germany, Denmark, France, Greece, Luxemburg, Netherlands, Portugal, Spain, Slovenia, Sweden, Switzerland in the year 2008.

The necessary conditions:

- Maximal under-five mortality rate in the period of three years before the crisis < 6.3 (per 1000 live births)
- Mean value of population ages 65 and above three years before the crisis > 14.2 % of total population.

The supporting conditions are: high life expectancy, low percentage of young population, high credits to private sector, high money availability, high companies' capitalization, high public health expenditures, low number of agriculture employees.

Subgroup 3: *Increasing credit activity in developing economies.*

List of 16 included crises: Chile in the year 1976, Israel and Spain in the year 1977, Argentina in the year 1980, Uruguay and Chile in the year 1981, Jordan in the year 1989, Lebanon and Brazil in the year 1990, Sweden in the year 1991, Kuwait in the year 1992, Brazil and Mexico in the year 1994, Argentina in the year 1995, Colombia in the year 1998, Uruguay in the year 2002.

The necessary conditions:

- Increasing credit activity in the period of three years before the crisis
- Population aged 15-64 in the year before the crisis < 64.3 % of total population
- Rural population three years before the crisis < 33.7 % of total population.

The supporting conditions are: low mortality of children, high life expectancy, high annual population growth, high road sector energy consumption.

Subgroup 4: *Socio-economic problems shown through a decreasing life expectancy.*

List of 17 included crises: Congo in the years 1991, 1992, and 1994, Latvia in the years 1995 and 2008, Sierra Leone in the year 1990, Kenya in the year 1992, Uganda and Burundi in the year 1994, Belarus, Central African Republic, Cameroon, Kyrgyzstan, Lithuania, Swaziland, Zambia, Zimbabwe in the year 1995.

The necessary condition:

- Slope of life expectancy for females in the period of three years before the crisis < -0.3 years per year

The supporting conditions are: stagnating or increasing mortality of children, stagnating or decreasing annual population growth, high mean percentage of added value by agriculture.

Subgroup 5: *Socio-economic problems shown through a non-increasing quality of public health.*

List of 25 included crises: Sierra Leone in the year 1990, Finland, Liberia, Nigeria, Norway, and Sweden in the year 1991, Kenya and Poland in the year 1992, Burundi in the year 1994, Belarus, Central African Republic, Latvia, Lithuania, Swaziland, and Zimbabwe in the year 1995, Bulgaria in the year 1996, Ukraine in the year 1998, Uruguay in the year 2002, Belgium, Greece, Hungary, Island, Italy, Portugal, and Spain in the year 2008.

The necessary conditions:

- Non-increasing life expectancy for females in the period of three years before crisis
- Slope of under-five mortality rate in the period of three years before crisis > -0.5 (per 1000 live births per year)

Supporting conditions are: high money growth two years before the crisis.

V. ANALYSIS OF SUBGROUPS

Subgroups 1 and 2 are relevant because they include many banking crises in developed countries in the years 2007 and 2008. There has been a strong "avalanche" effect causing that the crises occurred in many countries in the same time. But the detected conditions demonstrate that there have been also common patterns in many countries characterized by strongly increasing credit activity in economies with ageing population (Subgroup 1) or high credit activity in economies with high social security (Subgroup 2). The subgroups are actually very similar and there are 10 countries that are in both of them (like Italy, Netherlands, and Spain). The crisis in USA in year 2007 is perhaps the most representative example for the first group while Germany and Switzerland are representative examples for the second one. Although Subgroups 1 and 2 are mostly characteristic for the crises in years 2007 and 2008, it is very relevant to notice that Subgroup 1 includes also the crisis in Sweden in year 1991 while Subgroup 2 includes the crisis in Japan in year 1997. This demonstrates the generality of the result.

Subgroup 3 consists of a very different set of countries. Most of them are developing countries characterized by relative low percentage of active population (below 64.31%) in which increase of credit activity has happened. Representative examples are Israel in year 1977, Argentina in year 1980, and Uruguay in year 2002. It is interesting to notice that this subgroup includes also the crisis in Spain in year 1977 and the crisis in Sweden in year 1991. Similarity among

Subgroups 1, 2, and 3 is that they all include the same driving force (increased credit activity) in the societies that are not able to absorb these credits in a proper way. In Subgroups 1 and 2 this is mainly due to an aging population while in Subgroup 3 this is due to a relatively small percentage of active population that in many developing countries is a consequence of very high percentage of young population.

For socioeconomic Subgroups 4 and 5 a common characteristic is decreasing quality of life. The interpretation is that decreasing quality of life is not a cause of banking crises but actually a sign of problems in the country that may be a macroeconomic environment for the development of banking crises. There can be more causes of systemic problems including ethnic or civil wars, significant changes of economic system, and deep political crises. In this respect, individual countries may differ significantly.

Subgroup 4 includes undeveloped and developing countries in which in some cases banking crises are related with turbulent conditions. For examples, Congo in the period 1991-1994 is faced with the suspension of military and financial assistance for the Mobutu regime, Kenya in year 1992 experiences significant violence in certain parts of the country before presidential elections, and finally Belarus, Kyrgyzstan, Lithuania, and Latvia in the year 1995 are post-communist countries trying to implement novel economical models. In all these cases, a common result has been an instable socioeconomic system prone to banking crises.

Besides stagnating life expectancy Subgroup 5 includes the necessary condition of stagnating mortality of children till the age of 5 years. Both conditions to some extent reflect the quality of the health care in the country and its decrease or stagnation seems to be able to detect various socioeconomic problems in the society. It is relevant to notice that besides crises in countries like Nigeria and Ukraine it has been able to detect problems in Finland, Norway, and Sweden for the crises in the year 1991, as well as in a few EU countries for the crises in the year 2008.

Based on the presented analysis it may be concluded that we have two types of subgroups. In the first are Subgroups 1-3 that have in common the increased values for the indicator of domestic credit to the country's private sector. One of these subgroups has also a supporting condition representing money and quasi money as percentage of GDP. Because of these indicators, the first group of subgroups may be recognized as a *financially driven type* of banking crises. Subgroups 4 and 5 have stagnating or decreasing life expectancy of females as a common necessary condition. Additionally, both subgroups have increasing or stagnating mortality of children as either a necessary or a supporting condition. They may be recognized as a type of banking crises *related to socio-economic problems*.

VI. CONCLUSION

The induced subgroups present the basis for further data analysis, which mainly depends on the domain experts' preferences. In the domain of banking crises the first significant and partially surprising result is that socio-economic indicators turned to be relevant for a relatively large number of crises cases. The second interesting issue is that most of EU countries with crises in year 2008 are included into Subgroups 1 and 2, but that some of them are also included in Subgroup 5, which is a socio-economically-driven type. This fact may present an indication that the crises in some EU countries are significantly different and it is the topic of our further research. The good property of the methodology is that we explicitly know the countries that are included into specific subgroups and that we explicitly know indicators that characterize the subgroups. That enables the concentration of further analysis on these countries and these indicators.

The subgroups presented in this work are extensions of the two models reported in our previous work [4] that have been obtained using the same methodology, but on the smaller set of crises until the year 2007. The first model obtained in [4] is high money growth in countries with a low percentage of active population. It corresponds to financially driven Subgroups 1 and 2. The second model is a negative trend of the percentage of active population and a negative trend of life expectancy for males. The model is to some extent a combination of Subgroups 4 and 5. Both banking crises in the USA in the years 1988 and 2007 are positive cases of the second model. The subgroups presented in this work are more appropriate for the analysis of recent banking crises in EU countries.

ACKNOWLEDGMENT

The work presented in this paper has been supported through the EU FP7 project "FOC- Forecasting Financial Crises" and the project "Machine Learning Algorithms and Applications" financed by Croatian Ministry of Science, Education and Sport.

REFERENCES

[1] S. Battiston, D. Delli Gatti, B. C. Greenwald, and J. E. Stiglitz, "Default Cascades: When does risk diversification increase stability?", Journal of Financial Stability, vol. 8, pp. 138-149, 2012.

[2] D. Gamberger and N. Lavrač, "Expert-guided subgroup discovery: Methodology and Application", Journal of Artificial Intelligence Research, vol. 17, pp. 501–527, 2002.

[3] L. Laeven and V. F. Valencia, "Systemic banking crises database: An update", IMF Working Paper No. 12/163, International Monetary Fund, 2012.

[4] D. Gamberger, D. Lučanin, and T. Šmuc, "Descriptive modeling of systemic banking crises", Proceedings of 15th International Discovery Science Conference, Springer, 2012, pp. 67-80.

[5] W. Kloesgen, "Explora: A multipattern and multistrategy discovery assistant," in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.) Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, 1996, pp. 249–271.

[6] S. Wrobel, "An Algorithm for Multi-relational Discovery of Subgroups", in J. Komorowski and J. Zytkow (Eds.) Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97), Springer, 1997, pp. 78–87.

Transformation of OWL Ontology Sources into Data Warehouse

M. Gulić
Faculty of Maritime Studies, Rijeka, Croatia
marko.gulic@pfri.hr

Abstract - The Semantic Web, as the extension of the traditional Web, provides the semantic annotations of the information generated by different organizations. Semantic annotations are stored within ontologies. Ontologies are expressed using ontology language. Web Ontology Language (OWL) is one of the most popular ontology languages. As a result of the increasing use of ontologies, large quantity of complex, heterogeneous and semi-structured semantic data sources exists. There is plenty of useful information in these data sources that can be analyzed and used in the decision making process of an organization. In order to facilitate the analysis of semantic data, new data warehouse tools that support semantic data analysis must be made. Data warehouse is used in traditional business analysis and decision making processes. A star schema is the most common design of data warehouse. In this paper a method that transforms OWL structure into the star schema of data warehouse is proposed. After the designer chooses the fact in the ontology, a method transforms OWL into the star schema. At the end, the designer selects which elements in the schema remain while creating physical data warehouse.

I. INTRODUCTION

The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries [1]. Therefore, the goal of Semantic Web is the creation of standards and technologies that help machines to understand more about the Web data. These standards and technologies will improve user search results, data integration, navigation etc. Semantic annotations for Web data are stored within ontologies.

Ontology is a shared understanding of some domain of interest [2]. Ontology defines a set of entities and relations between them in a way that both humans and machines understand it. There are various data and conceptual models that can be thought of as ontologies (e.g. folksonomies, UML models, XML schemes, formal ontologies, etc.). Ontologies are expressed in an ontology language. OWL [3] is one of the most popular languages that is recommended by W3C organization.

As the Semantic Web is rapidly increasing, a large quantity of heterogeneous, composite and semi-structured semantic data sources exists. In these data, there is a lot of useful information that can be used in the decision making process of some company. For example, the data of the sales that were completed through the Web using a common ontology between two organizations can be used in decision making process. Data warehouse has proved as a good solution in decision making process. According to [4], a data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process. The process of creating the data warehouse includes business demands, data design, architecture design, implementation, and deployment [4]. A dimensional fact model is one of the most popular models. If the presentation area is based on a relational database, then this dimensional fact model is implemented with a star schema [5]. The star schema is different from typical relational databases that are in the third normal form.

In order to facilitate the analysis of semantic data sources in the data warehouse, new warehouse tools need to be made. The tools [7, 8, 9 and 10] transform OWL ontology into a relational database, but do not deal with the star schema. Therefore, the additional transformation from relational database to the star schema is necessary. Furthermore, there is a risk of losing relevant information during multi-step transformation. Some relations (hierarchy between classes, symmetric relations etc.) within the OWL ontology could be lost after transformation of ontology into a relational database. Therefore it is better to implement a method that transforms OWL ontology into the star schema directly.

Furthermore, these solutions (except the [7]) deal with all subclasses of certain class in a way that the particular table is created for every subclass. This solution is not good for transforming relational database into the star schema because every subclass is managed like specific entity which is not the true. However, we take advantage of their transformation and improve several procedures to transform ontology into the star schema. Only the authors in [11] propose the direct transformation from the OWL ontology to the star schema, but with several drawbacks. The main drawback of their work is that data warehouse does not include the data of instances of certain class and its subclasses that does not have the values for certain attributes in the ontology that are specific for several subclasses. In this paper a transformation of the OWL ontology into a conceptual multidimensional model is proposed.

The limitations described in the previous paragraph are also resolved. Our method deals with a class and its subclasses in a way that these classes represent one entity. Therefore the certain class and its subclasses become one dimension in the data warehouse. Our method also stores

the data of all individuals of parent class and all its subclasses despite some individuals do not have values for several attributes that are specific only for certain subclasses. In this way, the data warehouse contains total data from the ontology, which is essential if the analyst wants to get the accurate results.

The paper is organized as follows. In Section II the terminology of OWL Lite ontology, dimensional fact model and star schema is described. In Section III related work is discussed. In Section IV the method for transforming OWL structure into the star schema is presented. Finally, the conclusion is given in section V.

II. TERMINOLOGY

A. OWL Lite ontology

As we stated before, OWL (figure 1) is one of the most popular ontology languages. OWL represents the meaning of terms in an ontology and the relationships between them. OWL has three sublanguages: OWL Lite, OWL DL, and OWL Full. In this paper, the OWL Lite ontology is used. According to [12] the definitions of the OWL Lite elements that are essential for transformation are given below:

An *owl:Class* represents a group of individuals that have the same properties. For example, *HP1005* and *Canon505* are both members of the class Printer (figure 1). A *rdfs:subClassOf* establishes a hierarchy between one or more class. For example, the class Printer could be stated to be a subclass of the class Product (figure 1). *An Individual* is instance of the certain class. Individuals carry data that can be used for analysis.

A *rdf:Property* is used to define relationships between individuals (owl:ObjectProperty) or between individuals and data values (owl:DatatypeProperty). Example of properties can be named as hasProduct, hasPrice etc. A *rdfs:domain* of a property defines the individuals to which the property can be applied. For example, the property hasProduct can have the domain of Invoice (figure 1). A *rdfs:range* of a property defines the individuals that the property may have as its value. For example, the property hasProduct can have the range of Product (figure 1). An *inverseOf* element defines a property that can be the inverse of another property. If the property *P1* is the inverse of the property *P2*, and *X* is related to *Y* by the *P2*, then *Y* is related to *X* by the *P1*. Properties may be stated to be *SymmetricProperty*. If a property is symmetric, and the pair (x,y) is an instance of the symmetric property P, then the pair (y,x) is also an instance of P. If a property is *FunctionalProperty*, than a unique value is added to each individual that has this property. *Transitive* and *InverseFunctionalProperty* also exist in the OWL Lite ontology, but this type of property does not affect the transformation of OWL ontology into a star schema.

```
<owl:Class rdf:ID="Product">
</owl:Class>
<owl:Class rdf:ID="Invoice">
</owl:Class>
<owl:Class rdf:ID="Printer">
        <rdfs:subClassOf>
                <owl:Class rdf:about="#Product"/>
        </rdfs:subClassOf>
</owl:Class>
<owl:ObjectProperty rdf:ID="hasProduct">
        <rdfs:range rdf:resource="#Product"/>
        <rdfs:domain rdf:resource="#Invoice"/>
</owl:ObjectProperty>
<owl:DatatypeProperty rdf:ID="hasPrice">
     <rdfs:domain rdf:resource="#Product" />
     <rdfs:range rdf:resource="&xsd;float"/>
</owl:DatatypeProperty>
<Invoice rdf:ID="Invoice145">
          <hasProduct rdf:resource="#HP1005"/>
          <hasProduct rdf:resource="#Canon505"/>
</Invoice>
<Printer rdf:ID="HP1005">
          <hasPrice>95</hasPrice>
</Printer>
<Printer rdf:ID="Canon505">
          <hasPrice>90</hasPrice>
</Printer>
```

Figure 1 A part of OWL ontology

OWL Lite restrictions define the rules for using properties by particular instance. There are 5 restrictions: *allValuesFrom*, *someValuesFrom, minCardinality, maxCardinality and cardinality*. *AllValuesFrom* defines a local range restriction of some range class. *SomeValuesFrom* defines that at least one value of restricted property related to some instance is of the certain type. MinCardinality defines the minimal number of values of certain property that the individual must have. MaxCardinality defines the maximal number of values of certain property that the individual must have. Cardinality defines the exact number of values of certain property that the individual must have.

OWL uses the *XML Schema dataTypes* for defining the range of *owl:DataTypeProperties*.

B. Dimensional fact model and star schema

The dimensional fact model is a conceptual model that presents a data warehouse by a set of fact schemes with facts, measures, dimensions and hierarchies [16]. A fact is the center of interest in the decision making process.

A fact represents an event that occurs dynamically in the business process (e.g. invoice). A fact consists of measures. Measures are attributes that specify a fact. The most useful measures are additive and numeric. For example, every invoice has the total price, total tax etc.

Dimensions consist of discrete attributes that describe the fact event. Dimensions define the grain of the fact. For example, dimensions for every invoice are product, date, time, store etc. In this example, an analyst may request to see the money amount by product, by week, every evening from 7pm to 9pm. An analyst usually uses the OLAP (On-Line Analytical Processing) tools for analyzing data warehouse. These tools are based on a multidimensional conceptual view of data. In this paper, a transformation from OWL ontology to ROLAP (Relational On-Line Analytical Processing) system is proposed. The ROLAP uses the relational model for representing dimensional fact model. The implementation of dimensional fact model in the relational database is called star schema (figure 2). The

star schema consists of a set of dimension tables and the fact table. Each dimension has a set of attributes that describe the dimension. The fact table has a primary key that is a set of foreign keys of dimension tables. As it was stated before, the fact table also has the numeric and additive attributes that are measures of the fact.

III. RELATED WORK

As it was stated before, new data warehouse tools need to be made for more accurate analysis of semantic data sources. The focus of this paper is on the direct transformation of OWL ontology into the star schema because when a multi-step transformation is performed, some relations defined in the OWL ontology could be lost. The authors, who explore the transformation of OWL ontology into the relational database [7, 8, 9 and 10] deal with the transformation in the third normal form, not with the star schema. There are many solutions that transform the relational database in third normal form into the star schema. However, the main problem here is that the OWL ontology has a hierarchy of entities that appear in the ontology. The authors in [8, 9 and 10] propose the transformations that create an extra table in the relational database for every class in class hierarchy. For example, the class *Product* has subclass *Food* that has the additional attribute named lifetime. Transformations in [8, 9 and 10] create two tables, one for *Product* and one for *Food*. These two tables have all common attributes except for the attribute lifetime. The relation *rdfs:subClassOf* defined in OWL ontology (*Food* is a *subClassOf Product*), between *Product* and *Food* is lost after the transformation in the third normal form hence *Product* and *Food* will represent two different entities in the third normal form. When the transformation from third normal form into the star schema is made, the star schema has two dimensions for *Food* and *Product* instead of one dimension because these two classes represent the same entity, except that the *Food* class has the additional attribute. Hence, the wrong transformation could happen after losing a class hierarchy in the third normal form. However, some solutions of transformation proposed in [7, 8, 9 and 10] are used and integrated in the method proposed in this paper. A table for the parent class is created in all transformation. If a class has the *ObjectProperty* and the *maxCardinality* is 1, a column is created and it is a foreign key for the range class of the *ObjectProperty*. If a class has the *DataTypeProperty* and the *maxCardinality* is 1, a column is created and the type of the column is the most similar type in the relational database comparing XML data types added as a range of the *DataTypeProperty*. The authors in [10] propose the rules for transforming every *XML* data type into the certain data type in the relational database. If the *ObjectProperty* is *Functional*, the cardinality of the property is 1 and the column in the table is created. If the cardinality of any property is greater than 1, a new table will be created that has the primary key which is a combination of two foreign keys. The first key references to the domain class and the second references to the range class.

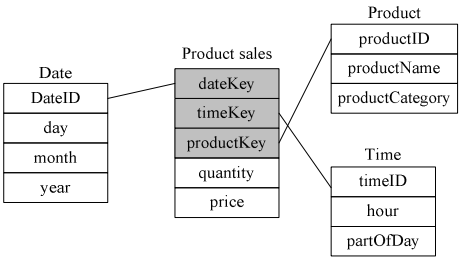


Figure 2 An example of the star schema

Although the authors in [7] present the transformation from OWL ontology in third normal form, they propose an interesting solution. The subclasses of parent class are recognized as the same entity but the ontology is transformed into the object relational database. For example, the class *Food* inherits all attributes of class Product and has the additional attribute lifetime. This solution of recognizing hierarchical classes as the same entity is used in this paper.

The authors in [11] provide the transformation from OWL ontology to the star schema but with notable limitations. For example, the analyst wants to analyse certain class according to the first attribute which is common for certain class and its subclasses in the ontology, and to the second attribute which is common only for several subclasses. Only the individuals of these several subclasses that have both attributes will be stored in the data warehouse. Therefore, when the analyst wants to see analysis just according to this first common attribute of a class and all of its subclasses, she will get only the individuals of these several subclasses that have the first and the second attribute in the OWL ontology.

IV. THE METHOD FOR AUTOMATICALLY TRANSFORMING OWL STRUCTURE INTO THE STAR SCHEMA

Transformation of OWL ontology into the star schema cannot be fully automated. The designer always has to choose which data she wants to analyze. Hence, the most important topic is that the designer selects what entity will represent the fact. As it was already mentioned, the direct transformation of OWL ontology into the star schema is proposed. The *Invoice* ontology (figure 4) will be taken as example for transformation of OWL Lite ontology into the star schema. In figure 4, The *Invoice* ontology is displayed as a directed graph to facilitate the steps of the transformation. That ontology is a combination of two real ontologies. First ontology [13] describes the main parts of the e-invoice while the second ontology [14] is the professional web vocabulary for e-commerce that describes a large number of products and services. These two ontologies are merged to show the real situation that can happen in a company while analyzing sales. For example, an analyst of the company wants to analyze sale by certain month, year, product etc. Invoice ontology that is used for example has only the specific parts of the ontologies [13] and [14] through which it can be shown all capabilities of transforming OWL Lite ontologies into the star schema. The names of ontology classes in the example are shorter than real names in the ontologies [13] and [14] to facilitate the

transformation steps. When the designer selects the fact, the transformation starts.

First, a dependency graph will be created from the OWL ontology. The dependency graph (DG) is an intermediate structure used to provide a multidimensional representation of the XML data describing the fact [15]. The DG is adapted from attribute tree [16]. Therefore, in the case of OWL ontology, the DG is a directed rooted graph whose vertices are classes or their data attributes in the OWL ontology. We decided to use DG instead of attribute tree because the DG has the possibility of including many-to-many relationships. When a property in the OWL has cardinality greater than 1, the designer determines if this property represents a many-to-many relationship. If it is a many-to-many, then the designer decides whether to include it into the DG or not. The inclusion of a many-to-many relationship is one of the special cases that we try to resolve in our transformation from an OWL structure to the DG. We also examine the appearance of convergence, shared hierarchies and partly the appearance of recursion within OWL ontologies. Whenever a class in the OWL has more than one property that has the same range class or a class has different properties whose successors have the same terminating range class within the OWL structure, it has to be determined if a convergence or a shared hierarchy may be implied in the DG. For resolving these cases, we adopted the approach from [18], but it is not described in the algorithm (figure 3) due to lack of space. In this paper, the appearance of recursion is resolved only for the case when the inverse property occurs. When an inverse property of some property exists within the ontology, it can happen that the range class of the inverse property is already inserted within the DG. In that case, a range class of inverse property should not be inserted into the DG because the enlargement of DG would repeatedly include the existing DG node. Pseudo code of transforming OWL ontology into the DG is shown in the figure 3.

```
{Create DG starGraph
Select a class or an ObjectProperty (with cardinality greater than 1) in the OWL ontology that will represent the fact;
Create the lists factDataTypeProperty and factObjectProperty;
IF fact is ObjectProperty THEN
  Get all objectProperties and dataTypeProperties of the domain class and range class of the fact and put them in the lists factObjectProperty and factDataTypeProperty;
ELSE
  Get all objectProperties and dataTypeProperties of the fact class and its first parent classes and put them in the lists
  factObjectProperty and factDataTypeProperty;
END IF
ProcessDataTypeProperties(factDataTypeProperty, true);
ProcessObjectProperties(factObjectProperty, true);

ProcessDataTypeProperties(dataTypeProperties,all){
        FOR every dataTypeProperty in dataTypeProperties
                ProcessDataTypeProperty(dataTypeProperty,all);
        END FOR
}
ProcessObjectProperties(objectProperties,all{
        FOR every objectProperty in objectProperties
                ProcessObjectProperty(objectProperty,all);
        END FOR
}
ProcessDataTypeProperty(dataTypeProperty,all){
        Cardinality(dataTypeProperty, all);
}
Cardinality(property, all){
        IF property cardinality > 1 AND the property represents a many-to-many relationship and the designer wants to include in the starGraph
        THEN
                Create node (with name of the property range Class) with cardinality > 1 (sign -> arc labeled as multiple) and insert into starGraph; (arc marked with a dash if all is false)
        ELSE
                Create node (with name of the property range Class) with cardinality = 1 and insert into starGraph; (arc marked with a dash if all is false)
        END IF
}
```

Figure 3 Pseudo code for transforming OWL into the DG

```
ProcessObjectProperty(objectProperty,all){
        IF objectProperty is Functional THEN
                Create node (with name of the property range Class) with cardinality = 1 and insert into starGraph; (arc marked with a dash if all is false)
        ELSE IF objectProperty is Symmetric THEN
                IF starGraph does not contain range Class of this objectProperty THEN
                        Cardinality objectProperty,all);
                        SubClasses(range Class);
        ELSE
                IF objectProperty is inverseOf THEN
                        IF starGraph does not contain range Class of this objectProperty THEN
                        Cardinality objectProperty,all);
                        SubClasses(range Class);
                        END IF
                ELSE
                        Cardinality objectProperty,all);
                        SubClasses(range Class);
                END IF
        END IF
}
SubClasses(class){
        IF class has subclasses THEN
                Create node with name SubClassName with cardinality = 1 and insert it into starGraph; (squared node in the DG)
        END IF
Get all common dataType and object properties of parent class and all subclasses of certain class and put them in the lists classObjectProperty and classDataProperty;

ProcessDataTypeProperties(classDataProperty, true);
ProcessObjectProperties(classObjectProperty, true);

Get all dataType and object properties of subclasses of certain class that are not common for every subclass and parent class and put them in the lists classObjectPropertyNotAll and classDataPropertyNotAll;

ProcessDataTypeProperties(classDataPropertyNotAll, false);
ProcessObjectProperties(classObjectPropertyNotAll, false);
}

}
```

For example, an analyst wants to examine the product sale of every invoice line. The designer selects the *InvoiceLine* class in the figure 4 to be a fact. The *InvoiceLine* class becomes the first node (it will be later renamed as *Product sales*) in the DG. Then, the algorithm gets all *dataType* and *object* properties of the *InvoiceLine* class and its first parent class (*Invoice*) and puts them into a list of properties. The *Invoice* class has one *dataType* property (*hasInvoiceID*) and four *object* properties (*hasDelivery*, *hasTime*, *hasDate*, *hasInvoicePrice*). All properties have the cardinality equal to 1 and therefore the *range* classes (*Date*, *Time*, *Delivery*, *InvoicePrice*) or *dataTypes* (xsd:string named *InvoiceID*) of these five properties are added to the DG (figure 5 a)). The *InvoiceLine* class has two *dataType* properties (*hasInvoiceLineID*, *hasQuantity*) and two *object* properties (*hasProduct*, *hasInvoiceLinePrice*). All properties have the cardinality equal to 1 therefore the *range* classes are added to the DG (figure 5 a)). Thereafter, for every *range* class inserted into the DG, the algorithm obtains all *object* and *dataType* properties.

The properties of the *Delivery* are *hasDate*, *isDeliveryFor, hasInvoiceID* and *hasTime*. These properties are inserted into the list of properties. An interesting *object* property is *isDeliveryFor* because it is an *inverseOf* property of the *hasDelivery* property in the *Invoice* class. The range class (*Delivery*) of the *hasDelivery* property is already inserted in the DG through the *Invoice* node. Hence, the node for the *range* class (*Invoice*) of the *object* property *isDeliveryFor* will not be created because of possible recursion. Other properties have the cardinality equal to 1 therefore the *range* classes (*Date*, *Time*) or *dataTypes* (*xsd:string* named *DeliveryID*) of these three properties are added to the DG (figure 5 b)). The same procedure is made for the classes *Date* and *Time*. It can be seen that the DG stops expanding in some direction when the remaining properties are only *dataType* properties.

The properties of the *Date* and *Time* in the *Invoice* node are the same as the properties *Date* and *Time* in the *Delivery* node. Hence, the shared hierarchy will be introduced at the end (figure 5 c)) according to [18]. The *dataType* property *hasPriceValue* is the only property of the class *InvoicePrice* therefore the node *PriceValue* is created (figure 5 b)). The same procedure is made for the *hasPriceValue* property *in the InvoiceLinePrice* class (figure 5 b)).

Furthermore, the *Product* class needs to be expanded in the DG. The *Product* class has two subclasses, *FoodProduct* and *ComputerProduct*. The *ComputerProduct* class has one subclass, the *Game* class. When the class has subclasses, the first step is to create the *subClassName* node that has a squared shape (figure 5 b)). This shape defines that each value of the *subClassName* attribute in the dimension table that corresponds with the subClassName node will be one of the names of subclasses (*FoodProduct*, *ComputerProduct* and *Game* in the example) or the root class (*Product*). The designer can rename the name of the *subClassName* node after creating the column in the relational database. In this example the column will be named *categoryName* (figure 6). After creating a node for all subclasses names and their root class, the algorithm gets all common *dataType* and *object* properties of the root class (*Product*) and all child classes (*FoodProduct*, *ComputerProduct*, *Game*). These properties are *hasProductPrice* and *hasProductName*. These properties have the cardinality 1 therefore two nodes are created in the DG. The node *Product* is the parent node of these nodes. After creating nodes for properties of the *Product* class that are common to all subclasses of *Product* class, the properties that are special just for a subset of all subclasses need to be processed. These properties are *isMultiplayered* and *hasIngredient*. The *isMultiplayered* property has cardinality equal to 1

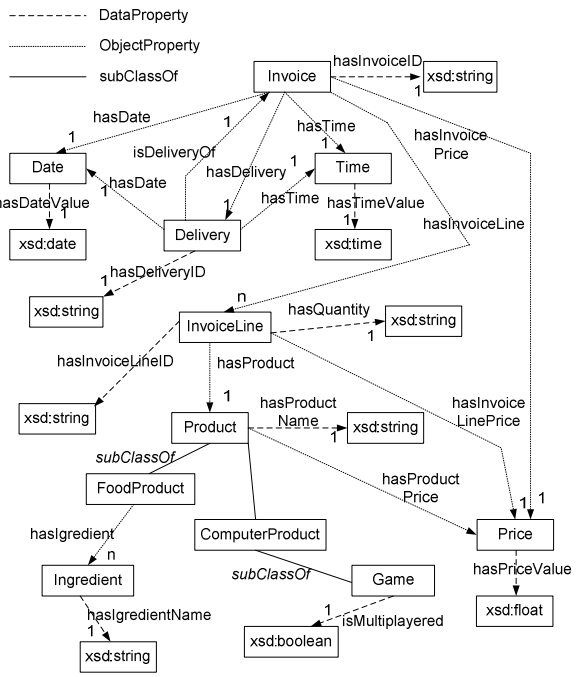


Figure 4 An example of the *Invoice* OWL ontology displayed as directed graph

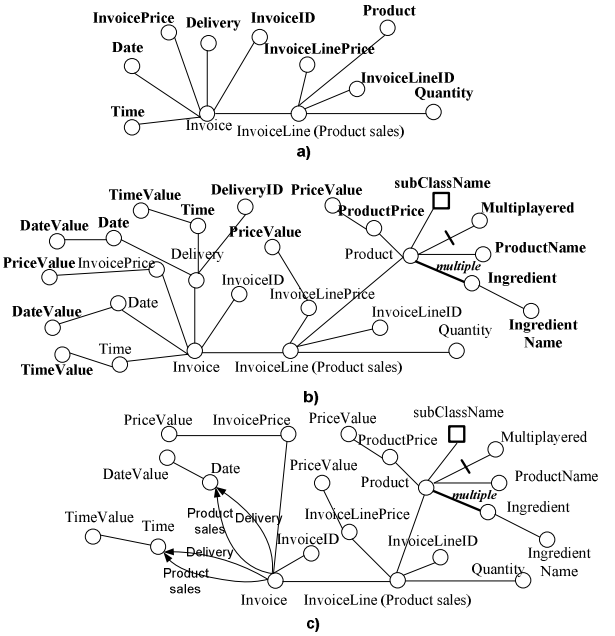


Figure 5 Dependency graph for the transformation of the Invoice OWL ontology

and the node for the range of the property is created in the DG but the arc of that node is marked with dash to indicate that this attribute will have null values in the data warehouse for some products. The *hasIngredient* property has cardinality greater than 1. The *hasIngredient* property represents a many-to-many relationship that the designer decides to include, hence the node for the *range* class of the *hasIngredient* property is created and its arc is labeled as *multiple*. This solution of labeling many-to-many relationship in the DG is adopted from [15] (figure 5 b)).

At the end of the algorithm, the DG may be rearranged as discussed in [16]. Considering the DG in figure 5 c), we observe that there is no need for the *InvoicePrice*. Therefore pruning is made for the subtree of the *InvoicePrice* so the nodes *InvoicePrice* and *PriceValue* are dropped from the DG. Grafting is made for the nodes *Invoice*, *ProductPrice* and *InvoiceLinePrice*. The *Ingredient* and the *DeliveryID* are also dropped from the DG. The final steps of building a multidimensional schema include the choice of dimensions and measures as described in [17]. In our example, *UnitPrice*, *Price* and *Quantity* are chosen as measures, while *DeliveryDate*, *ProductSalesDate (renamed as Date)*, *DeliveryTime*, *ProductSalesTime (renamed as Time)* and *Product* are the dimensions. After including measures in the fact table and creating dimension tables, the logical schema is obtained. The resulting star schema can be seen in Figure 6. The shared hierarchies *Date* and *Time* are represented by only two dimension tables. The star schema has one snowflake structure for the ingredients that the product could have because it represents the many-to-many relationship between the product and the ingredients. In the star schema, the date dimension is enriched by introducing attributes *dayOfWeek* and *month*.

V. CONCLUSION

In this paper a method that transforms OWL ontology into the star schema is proposed. The method semi automatically transforms all elements of the ontology after the designer selects which class in the ontology will be the fact. Before implementing the star schema, the designer selects which data will be included in the data warehouse in order to drop unnecessary data from the data warehouse.

Our method identifies all subclasses of the certain parent class as the same entity like parent class and thus it differs from other methods that transform OWL ontology into the normalized relational database or star schema. It is important because the certain class and its subclasses could become one dimension in the star schema. In this way, all individuals of parent class and all of its subclasses will be stored together although they have some different attributes that are defined only for some subclasses. The main contribution of our work is the direct transformation of the OWL ontology into the DG. A star schema is implemented from the obtained DG.

In the future work the attention will be paid on the ranges of the property that are composed of more classes. The solution for the symmetric property should also be improved. Symmetric property is like a recursive structure and it is difficult to define when the algorithm stops while obtaining the data from that property. Also, recursion in general will be exminated. Finally, the implementation of the tool that uses the method described in this paper will be made.

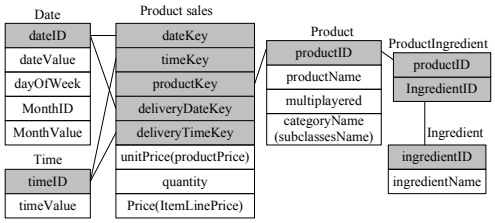


Figure 6 The star schema for the Invoice ontology example

REFERENCES

[1] World Wide Web Consortium (W3C), W3C Semantic Web Activity, Web resource "http://www.w3.org/2001/sw/", retrieved February 4, 2013.

[2] M. Uschold, M. Grüniger, "Ontologies: Principles, Methods and Applications", The Knowledge Engineering Review, Vol 11(2), pp 93–155, 1996.

[3] G. Antoniou, and F. van Harmelen, "Web ontology language: Owl", In A Semantic Web Primer, pages 110−150, MIT Press, 2004.

[4] W. H. Inmon, Building the Data Warehouse, 3rd edition, John Wiley, New York, 2002.

[5] R. Kimball, and M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd edition, Wiley, New York, 2002.

[6] R. Kimball, L. Reeves, M. Ross, and W. Thronthwaite, The Data Warehouse Lifecycle Toolkit, Wiley, New York, 1998.

[7] X. Liu, DataWarehousing Technologies for Large-scale and Right-time Data, dissertation, Faculty of Engineering and Science at Aalborg University, Denmark, 2012.

[8] E. Vysniauskas, and L. Nemuraite, "Transforming ontology representation from OWL to relational database," Information Technology and Control, vol. 35A, no. 3, pp. 333-343, 2006.

[9] E. Vysniauskas, and L. Nemuraite, Mapping of OWL ontology concepts to RDB schemas, Proceedings of the 15th International Conference on Information and Software Technologies, pp. 317–327, Lithuania 2009.

[10] I. Astrova, N. Korda, and A. Kalja, Storing OWL Ontologies in SQL Relational Databases, Engineering and Technology 23, pp. 167–172, 2007

[11] V. Nebot, and R. Berlanga, Building data warehouses with semantic data, Proceedings of the EDBT/ICDT Workshops, 2010

[12] World Wide Web Consortium (W3C), OWL Web Ontology Language Overview, Web resurce "http://www.w3.org/TR/owl-features/", retrieved February 4, 2013.

[13] e-Račun, Web resource "http://edocument.foi.hr/ontologies/ubl2_0.owl" , retrieved February 4, 2013.

[14] GoodRelations: The Professional Web Vocabulary for E-Commerce, Web resurce "http://www.heppnetz.de/ontologies/goodrelations/v1.html", retrieved February 4, 2013.

[15] B. Vrdoljak, M. Banek, and S. Rizzi, Designing Web Warehouses from XMLSchemas, Data Warehousing and Knowledge Discovery, 5th International Conference DaWak, 2003

[16] M. Golfarelli, D. Maio, and S. Rizzi, The Dimensional Fact Model: a Conceptual Model for Data Warehouses, International Journal of Cooperative Information Systems, vol. 7, 1998.

[17] M. Golfarelli, D. Maio, and S. Rizzi, Conceptual design of data warehouses from E/R schemes, Proc. HICSS-31, vol. VII, Kona, Hawaii, pp. 334-343, 1998.

[18] M. Golfarelli, and S. Rizzi, Data warehouse design: modern principles and methodologies, McGraw-Hill, 2009

Using Big Data and Sentiment Analysis in Product Evaluation

Lada Banić, Ana Mihanović, Marko Brakus
Poslovna Inteligencija d.o.o., Zagreb, Croatia
{lada.banic, ana.mihanovic, marko.brakus}@inteligencija.com

Abstract - When purchasing a product for the first time one usually needs to choose among several products with similar characteristics. Companies use to promote their brands and products pointing out good characteristics avoiding to mention the poor ones. The best way to choose the most suitable product is to rely upon the opinions of others. The system to be described here collects opinions about hotels from the web, evaluates them, aggregates these evaluations and offers cumulative, easy-to-understand information. Generated information is intended for the possible prospective customer, but also for the hotel managers providing them with additional guidance in future business development.

I. INTRODUCTION

Rapid growth of online information concerning services and products has been accompanied by users' comments about these items. The amount of reviews has been increasing at a great speed making the web more and more subjective and opinionated. The Web visitors' comments cover almost all areas, as they are present not only on specialized review sites, but also on most of the published news and blog pages.

Independent unbiased consumer reviews are known to be the most credible sources of product or service information and people tend to rely primarily on them when making a decision about a purchase. According to [1] about 80 percent of users of TripAdvisor write travel reviews, 20 percent of visitors rely on other users' reviews when planning a trip; looking into other users' comments and travel blogs is the most popular online activity. Analysis of such texts is a more productive way of collecting user information than the traditional structured data collection by surveys where people are usually unwilling to take the time to answer presupposed questions. Conversely, sentiment analysis [2] "listens" to published opinions and answers "the unknown". However, due to its size, this online form of a word of mouth is difficult to grasp fully. Reading the reviews only partly might induce a biased opinion. This calls for application of algorithmic methods of analysis of a large number of reviews.

When working with the web content concerning online reviews, blogs, forums etc, one deals with huge amounts of unstructured data in the attempt to extract information. To be successful one needs those data to be structured so that the necessary information becomes available. When extracted, the information needs to be aggregated and presented to the interested party(s) in an understandable form.

During the process of unstructured data collection, information extraction or in this case sentiment extraction, aggregation of gathered information and presentation to the interested party(s) one is dealing with several innovative issues:

- BigData (storing and analyzing large amounts of unstructured data)
- text mining (deriving information from text)
- sentiment analysis (finding out opinions from text)

A. *BigData*

The term BigData [3] stands for the process of extracting actionable data from various, often nontraditional data sources. These sources may include structured data such as databases, and in this case unstructured data like HTML, as well as social data and images.

Unstructured data (or unstructured information) refers to the information that either does not have a pre-defined data model and/or does not fit well into relational tables. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts.

The unstructured source data are then structured up to a point and pushed down to a structured format, which is then stored in a database for further manipulation.

The amount of data being collected is more than traditional computer infrastructures can handle, exceeding the capacities of databases, storage, networks and everything in between. Extracting actionable intelligence from Big Data requires handling large amounts of various data and processing them very quickly. Major issues in Big Data processing is that data inputs must be consistent and clean.

IBM describes these new demands operating with Big Data across four dimensions: volume, velocity, variety and veracity, all of them overwhelming

B. *Text Mining and Text Analytics*

Text mining is analysis of data contained in a natural language text. The application of text mining techniques to solve business problems is called text analytics. It is an interdisciplinary field which combines information collection, data mining, statistics and computational linguistics.

The goal of text mining [8] is to derive high-quality information from the text. This is typically done through recognizing the patterns in data. In other words, the purpose of text mining is to process unstructured information and to extract meaningful numeric indices from it. Generally speaking, text mining 'turns text into numbers'.

The numeric indices make the information contained in the text accessible to further analysis or to further data mining (statistical and machine learning) algorithms.

By text mining, different analyses are possible: of words and of clusters of words within documents, of similarities between documents, of the relation of documents to other variables of interest to e.g. data mining projects etc.

Text analysis involves the following processes:

- information retrieval
- Natural language processing
- named entity recognition
- recognition of pattern identified entities (features such as telephone numbers, e-mail addresses, quantities (with units)
- coreference: identification of noun phrases and other terms that refer to the same object
- relationship, fact, and event extraction: identification of associations among entities and other information in text
- sentiment analysis
- Quantitative text analysis

C. *Sentiment Analysis*

Sentiment analysis [2] is one of the applications of the text mining techniques.

An important aspect of our information-gathering is to find out what other people think. Even before the World Wide Web has become widely spread, people asked friends for opinion on different subjects in order to make a better and wiser decision. The availability of the world-wide-web and text-mining techniques allows us, on a much wider scale, to find out opinions of other people who are neither our personal acquaintances nor well-known professional critics. In addition to individuals businesses also seek to identify and capture the substance of the "word of mouth", that is, the information consumers exchange with one another. Their aim is to manage the impact which this information, along with its consequent e-reputation, can have on their products and brands, and to take it into account when developing a strategy or improving business operation.

II. RELATED WORK

As online user reviews are interesting for both the consumers and producers, this area spawned a lot of work. Datasets and corpora are publicly available for sentiment analysis in the mainstream domain of product reviews, whereas touristic domain is less frequent.

Content-based focused web crawling using Apache Nutch is described in [3]. Generally, focused crawling approaches assume that webpages on a given topic are more likely to link to those on the same topic.

While manual creation of several review recognizers is possible for a few manually chosen web sites, it is impractical for crawling at a larger scale. An advanced method of review extraction in [4] deals with the impracticality of manually creating a review wrapper for each specific review web site. The method relies on DOM tree consistencies and similarities and shows encouraging performance.

A system for collecting user comments about hotels from the web and creating classified and structured overviews of such comments is presented in [5]. The system crawls pages from certain web sites, extracts the reviews by screen scrapping, detects the language and splits the reviews into segments. It then detects sentiment polarity by using statistical polarity detection based on n-grams combined with information extraction methods with domain-specific dictionaries. Benchmarks demonstrated F1 measure of above 80%.

High accuracy of opinion mining applied on textual data from travel review sites is demonstrated in [7]. Emphasis was put on feature extraction and burst detection. The underlying techniques can detect hotel features and possible reasons for changes in opinion as well as characterize geographical areas based on hotel reviews.

III. PRODUCT EVALUATION SYSTEM OVERVIEW

The system presented here has been being developed as part of the FAIR project, which is carried out by three partners: Testntrust from France, Beia from Romania and Poslovna inteligencija from Croatia.

The FAIR project encompasses customer satisfaction ratings, brand ratings, social networking, connecting brands with their customers, exchanging and sharing relevant information.

The first product chosen for sentiment analysis and evaluation was a hotel.

The project is concerned with collecting hotel reviews, storing them, and analyzing their sentiment and aggregating analysis results into single hotel-based estimation. The system of review collection handles crawling, extracting the hotel reviews and storing them for analysis. The review dataset obtained is subjected to text

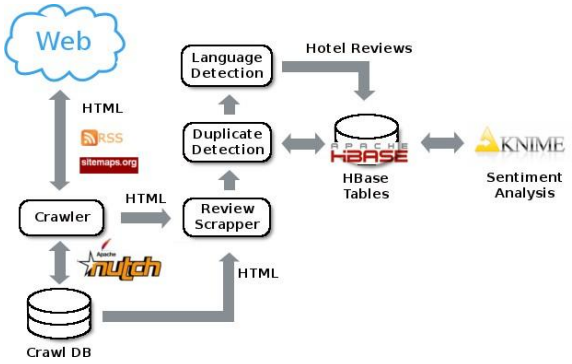


Figure 1. Product Evaluation System configuration

mining and sentiment analysis resulting in evaluation of every single review. Review evaluations are aggregated on the hotel level in order to get a cumulative estimate for each hotel.

A. Data (Hotel Reviews) Collection

In gathering hotel review data [9], focus was set on the web sites specialized for travelers' reviews. The notable ones were tripadvisor.com, hotels.com, laterooms.com, booking.com, the tripadvisor.com being the most prolific. Review pages were obtained by crawling the web sites for hotels and reviews, using the sitemaps.org format and reading RSS feeds. The most efficient solution is offered by sitemaps.org, an XML file format specifying a map of the web site using an updatable list of URLs. Similar and more frequent solutions were RSS feeds published on the web site. Both sitemaps.org XML index files and the RSS feeds were checked periodically for newly added or updated URLs. However, in a large majority of cases, such indices of site's interesting and newly updated web pages were not available and one had to rely on the web site crawling.

A web crawler is a program that traverses the web site starting from a given set of initial URLs and follows the links matching a given pattern to a certain depth. An ideal crawler for the purpose of quickly downloading only the pages containing reviews and checking if they are updated, would be distributed or at least parallel, incremental and focused. To update a set of downloaded pages it is preferable to apply incremental crawling rather than to restart crawling. In focused crawling the space of crawled pages is narrowed by the use of a classifier, which decides whether a page is interesting or not. This can be a simple URL pattern to match. Parallel crawling is performed by running multiple processes simultaneously, to crawl web sites in a reasonable amount of time. For the aforementioned reasons and for the purpose of work planned in the future, the general-purpose web crawler offered by Apache Nutch [13], an open source web-search software project was used.

All hotel review sites have their own review page layout usually containing several reviews for a hotel. Although it is not very scalable to have a manually created scrapping template for each site, currently, we applied the DOM (Document Object Model) based scrapping method to extract review and hotel information from the HTML pages using the templates defined by sets of XPath and regular expressions. More automated methods of review scrapping appropriate for large-scale crawls are described in [10].

The extracted review information comprises the name (nickname) of the reviewer, the date of posting and the review text. The hotel and review information extracted from HTML was further processed to check the language of the text, as most of the sites contain multilingual reviews. Language detection of the review text was done by means of the Apache Tika [12] .Apache Tika is a content analysis toolkit, which among others, contains language detection. Language detection classifies the given text into one of the language classes. Classification is performed by comparing the n-gram profile of the given text to previously created n-gram profiles of the languages. N-grams are N characters long chunks of consecutive text created in a sliding-window manner. N-gram profile is a frequency histogram of all the n-grams of the given text. Comparison of profiles is done using the geometric distance and the chosen default n-gram length is 3.

Special attention was paid to extracting and normalizing a unique hotel name out of the title of the hotel review, as the same hotel can be differently titled across the web sites.

Duplicate reviews were a result of repetition of the review on different pages inside the web site and of its copying across different sites. Duplicate detection was realized from coarse to fine, meaning that the URL name was first checked against the table of the pages visited earlier and the extracted review (its hash value) was then looked up in the table of all reviews.

To make the system extendable to products and services other than hotels and to scale to ever expanding unstructured data, Apache Hadoop [14], an open source implementation of the MapReduce framework was used as a distributed execution and storage environment. The Hadoop is popular for developing large-scale data-intensive applications. The Hadoop ecosystem comprises components such as HBase, Hive and Pig for storing, querying and analyzing data, respectively. Hotels and hotel reviews are stored in HBase tables. HBase [15] is a distributed NoSQL database, which efficiently holds unstructured data in large tables and can be concurrently and randomly accessed.

B. Dataset Description

A dataset of hotel reviews collected from tripadvisor.com. was used. Tripadvisor.com, a travel web site, offers information on a big diversity of touristic sites including hotels, restaurants, museums etc. Most of the information on this site is user generated. Users can log in and post a review for any touristic site. Also, hotel managers have a possibility to reply to any review.

Reviews for a given hotel are usually displayed in a list of five per page. Hotel information is displayed at the top of the page. When extracted from a web page, the web page content and the extracted information were structured in the following way:

1. For each extracted review a key indicator was created before the review was stored into a database
2. The hotel details extracted from the web site content were: hotel name, country, city and street.
3. Hotel reviews details included: text of the review, date of the review, language of the text.

Hotel details were extracted in the mentioned way so that each hotel could be uniquely identified in the process of aggregation of evaluation results. Namely, if aggregations had been performed based on the hotel name only, evaluations of different hotels carrying the same name would have been aggregated, producing wrong evaluation results and supplying the end users with misleading information.

The review date allowed creation of aggregated evaluation for a defined period of time and follow up of

the time related changes in evaluation results. The language attribute allowed the application of a proper dictionary in evaluation of downloaded review, for the English dictionary applied to a French review would fail to produce results.

In this phase of the process only the hotel reviews in English were downloaded.

C. Text Analytics and Sentiment Analysis

Text analytics and sentiment analysis were performed by means of the open-source software KNIME [13].

KNIME is a user-friendly graphical workbench for the entire analysis process: data access, data transformation, initial investigation, powerful predictive analytics, visualization and reporting. The open integration platform provides over 1000 modules (nodes), including those of the KNIME community and its extensive partner network.

By means of KNIME a sentiment analysis stream, consisting of the following major steps, was created: retrieving data from the database, dictionary development and implementation, review scoring.

1) Retrieving Data from the Database

Connectivity to the database was accomplished with the (in-house developed) „HBase reader“ node, which took a few simple parameters to connect to the database, to the host machine IP and to the port. Data were retrieved from the database in packages of 30 000 records. Only records that were not previously evaluated were retrieved, assuring that the reviews from the database were evaluated only once, speeding up the whole process.

2) Dictionary Development

Dictionary development and implementation A new dictionary containing words and phrases used in evaluation of hotel reviews was developed and implemented. It was developed based on the words and phrases found in a sample of downloaded Internet hotel reviews.

Composing a dictionary is precise and time-consuming task. It was necessary to include the terms and phrases from the current hotel reviews which could contribute to the evaluation, and to multiply them and modify them in a way in which they might occur in other users' reviews, following grammar rules but including also slang expressions and abbreviations used in everyday speech.

Hotel evaluation was planned to be carried out according to four different categories: tidiness, service, atmosphere and general category. Every term in the dictionary was associated with only one of the four categories and given a grade in the range from 1 to 5, 1 referring to bad, and 5 referring to excellent.

The category tidiness was meant to evaluate the cleanness of accommodation including room, bathroom and hotel in general, the category service dealt with the affability of the staff, the category atmosphere served for evaluation of hotel location, noise level and similar characteristics, whereas the category general included all terms suitable for hotel evaluation that could not be included in the previously mentioned categories.

Dictionary for KNIME is made with a simple text editor. It consisted of several columns: one column referred to the term of the phrase, the other columns referred to the above mentioned categories for evaluation, and only one of them contained the grade for the respective term or phrase.

The term or phrase within dictionary where word or the sequence of words to be recognized in an unstructured review text.

3) Review Scoring

When loaded from the HBase into KNIME reviews are treated as single documents.

Review scoring is performend in following way:

1. Every word from document is tagged according to specification of terms and phrases in the dictionary (keeping the information about the review it belongs to)

2. Document is transformed into 'Bag of words', and words not specified within applied dictionary are excluded.

3. Every term or phrase within dictionary has assignedd category and mark, so recognized terms and phrases within review are grouped accoridng to categories specified within dictionary, counted and average mark for every category is calculated

The results, i.e. average grade for each evaluated review for each category was written back to the Hadoop database.

As a result of Hadoop aggregation of evaluated reviews for each hotel the average grade for every specified category on the level of single hotel is calculated.

The results, i.e. average grade for each evaluated review for each category was written back to the Hadoop database.

As a result of Hadoop aggregation of evaluated reviews for each hotel the average grade for every specified category on the level of single hotel is calculated.

Evaluation of reviews was performed periodically as was aggregation allowing a timely follow up of changes.

IV. Development of Evaluation Method

Different evaluation systems can be implemented in the process of semantic analysis. The choice of the system will depend on the product of evaluation, its characteristics, on the end users of such cumulative information and on the level of information to be extracted.

In our first evaluation system we evaluated each term or phrase as negative and positive. Cumulative information about the number of positive and negative terms or phrases was generated on the level of the hotel and final evaluation of the hotel was generated in the following way:

- more than 70 percent of terms are positive – grade 3 (referring to very good)
- between 70 percent and 30 percent positive terms – grade 2 (referring to average)
- less than 30 percent positive terms – grade 1 (referring to bad)

The second evaluation system involved evaluation of terms and phrases with the help of grades from 1 to 5, where 1 referred to bad and 5 referred to excellent. Each term or phrase recognized in the review was evaluated according to the specification in the dictionary. Average grade for each single review was obtained. The total grade for each hotel was calculated as average grade of all reviews aggregated on hotel level.

As evaluation system with only one total average grade did not provide enough valuable information about hotels, its development was taken one step further by specification of evaluation categories.

Four different categories were introduced: *location*, *service*, *atmosphere* and *general* (terms that did not fit in any of the specified categories).

Every term was categorized before ti was graded with marks in the range from 1 to 5. For every review all four categories were specified separately. The total grade for every hotel was determined as average grade of all reviews aggregated by category on the hotel level.

V. RESULTS

Reviews were collected using the sitemaps.org format from tripadvisor.com web site.

A total of 105 294 reviews were downloaded for a total of 3403 different hotels.

The developed dictionary comprised all together 1211 terms and phrases extracted from hotel reviews that expressed customer's specific opinion about hotels. In dictionary development, emphasis was on expressions because the same term within different expressions could have a positive and a negative meaning.

An overview of a number of terms and expressions by category, included in the dictionary, is given in Table I.

The use of dictionary resulted in evaluation of 40562 reviews in the tidiness category, of 50425 reviews in the service category, of 31437 reviews in the atmosphere category and of 85883 reviews in the general category.

Results of review evaluations by category with percentages calculated with respect to total number of downloaded reviews are shown in Table II.

The reason for low percentages of evaluated reviews lies with the dictionary which is still under development. For a more complete and reliable review evaluation more efforts should be made to enhance development of dictionary.

Hotel evaluation is result of aggregation of review evaluations by category on the hotel level is shown in Table III. It is clear that the coverage of evaluations on the hotel level is much higher than on the review level. This is because single reviews which were evaluated according to one or two different categories, when aggregated on the hotel level cover in total more categories then single reviews.

Table IV. shows the number of terms and phrases for every category evaluated with a certain grade, implemented into the dictionary and then searched for within the reviews structuring reviews and creating bases for reviews evaluation.

Statistics of grades appearing for the categories in the review evaluation are given in Figure 2. From the graph presentation, it is obvious that the terms and phrases most often found in the review evaluation are those that are graded within the category "general" with excellent mark, within the category "service" with a very good grade and within the categories "general" and "tidiness" with grades very good and excellent, respectively.

According to the statistics of terms and phrases belonging to the different categories and grades, shown in Table IV, it was expected for excellent grade to have the highest number of "hits" in the review evaluation, for most of the terms and phrases in the dictionary are graded as excellent, what is truth for category "General". However, as shown on figure 2, terms belonging to

TABLE I.. NUMBER OF TERMS AND PHRASES IN THE DICTIONARY BY EVALUATION CATEGORY.

TABLE II. THE NUMBER OF REVIEWS EVALUATED BY CATEGORY, AND PERCENTAGES CALCULATED WITH RESPECT TO TOTAL NUMBER OF DOWNLOADED REVIEWS.

Number of downloaded reviews	*Tidiness*	*Service*	*Atmosphere*	*General*
105294	40562	50425	31437	85883
	38,52%	47,89%	29,86%	81,56%

TABLE III. RESULTS OF THE HOTEL EVALUATIONS BY CATEGORY WITH PERCENTAGES CALCULATED WITH RESPECT TO THE TOTAL NUMBER OF HOTELS

Number of evalauted hotels	*Tidiness*	*Service*	*Atmosphere*	*General*
3403	2729	2927	2469	3267
	80,19%	86,01%	72,55%	96,00%

TABLE IV. NUMBER OF TERMS AND PHRASES FOR EVERY CATEGORY EVALUATED WITH A CERTAIN GRADE AND IMPLEMENTED INTO THE DICTIONARY..

Category Grade	*Tidiness*	*Service*	*Atmosphere*	*General*	*Total by grade*
1	50	98	29	56	233
2	60	59	40	44	203
3	9	20	25	42	96
4	76	72	47	45	240
5	108	131	89	111	439
Total by category	303	380	230	298	

category "Service" graded with very good mark have high number of hits despite the fact that number of such terms implemented in dictionary is almost half of number of terms implemented for the same category, but graded as excellent that have significantly smaller number of hits.

All this findings show us the importance of the dictionary development. The more time spent in the extraction of the terms and phrases from the existing reviews and categorization and grading of those terms, the better, more consistent and accurate the results of the review evaluation will be.

The aggregated evaluation results done on the hotel level show to the end users the current situation and position of the hotel on the market. However, of great importance is the analysis of change of customer's perception of the hotel with time. Since every review has a posting date specified, analysis of change of average grade for all the categories can be followed through time. In the figure 3. analysis for the hotel 'Al Duca di Venezia' is presented.

Since dictionary is not fully developed and only a limited number of reviews for evaluation was collected, results are not fully reliable, but they do clearly show the drop in evaluation of categories atmosphere and general at the end of 2011, while category service has slide but steady fall from the beginning of 2010 to the end of 2012.

VI. CONCLUSION

The best way to decide about the most suitable product is to rely upon the opinions of others. With the growing availability and popularity of sources such as online review sites and personal blogs as well as with the ever advancing information technology which enable efficient processing of large scale unstructured data, new opportunities arise for finding and understanding different opinions and for facilitating the decision making process. Systems for sentiment analysis of products and brands can be developed. Such systems can transform vast amounts of unstructured data into an aggregated structured information. However, certain issues that come with processing large amounts of unstructured data in order to structure them have to be considered. One issue is data quality which deals not only with the recognition of the required terms and phases but also with the originality of the collected data. The later will depend on review authors but also on review multiplication which may generate misleading evaluations. Specificity of product also needs to be considered especially when developing a categorical

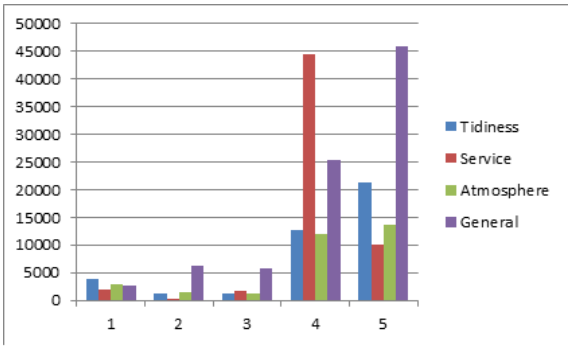


Figure 2. Grades acording to categories in review evaluation

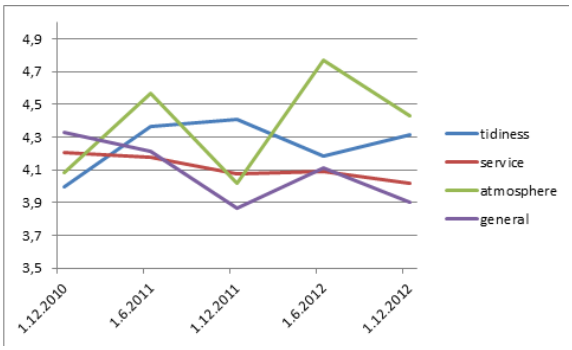


Figure 3. Average grade for all the categories followed through time for 'Al Duca di Venezia' hotel.

evaluation system which needs to be adjusted to the product type. The terms and expressions contained in the dictionary should also be suited to the product type because of different terms and phrases, language forms and slang expressions which refer to different product types. In the future sentiment analysis system will bring new knowledge to the individuals as well as to businesses people. They are going to make managers more aware of the end users' perception of their products but also their competitors' products. In the environment of saturated market and recession aggregated information based on a vast amount of data will become invaluable for planning future business strategy planning and development.

[1] U. Gretzel, K. H. Yoo, M. Purify, Online Travel Review Study. Role & Impact Of Online Travel Reviews. Research Report. Laboratory For Intelligent Systems In Tourism. Texas A&M University, 2007.

[2] B. Pang, L. Lee, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2, 2008

[3] O. Medelyan, S. Schulz, J. Paetzold, M. Poprat, and K. Marko. "Language specific and topic focused web crawling." In Proceedings of the Language Resources Conference LREC. 2006.

[4] W. Liu, Y. Hualiang, and X. Jiangu. "Automatically extracting user reviews from forum sites." Computers & Mathematics with Applications 62, no. 7 (2011): 2779-2792.

[5] W. Kasper, and M. Vela. "Sentiment Analysis for Hotel Reviews." In Proceedings of the Computational Linguistics-Applications Conference. 2011.

[6] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz. "Big data, analytics and the path from insights to value." MIT sloan management review 52, no. 2, 2011, pp 21-32

[7] E. Bjørkelund, T. H. Burnett, and K. Nørvåg. "A study of opinion mining and visualization of hotel reviews." In Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, pp. 229-238. ACM, 2012.

[8] R. Feldman and J. Sanger. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, 2006.

[9] M. McGlohon, N. Glance, and Z. Reiter. "Star quality: Aggregating reviews to rank products and merchants." In International Conference on Weblogs and Social Media, 2010.

[10] L. Wei, Y. Hualiang, X. Jianguo, Automatically extracting user reviews from forum sites, Computers & Mathematics in Natural Computation and Knowledge Discovery, Volume 62, Issue 7, October 2011, pp 2779–2792

[11] http://nutch.apache.org/

[12] http://tika.apache.org

[13] http://knime.org/

[14] http://hadoop.apache.org/

[15] http://hbase.apache.org/

Model of the business intelligence system for credit risk analysis

Tamara Gazdić, M.Sc [*], Ljiljana Kašćelan, PhD [**]
[*] Central Bank of Montenegro/Chief Economist and advisors, Podgorica, Montenegro
[**] University of Montenegro/Faculty of Economics, Podgorica, Montenegro
ljiljak@ac.me

Abstract - The subject of the research of this paper is modeling of the business intelligence system in area of credit risk analysis on the basis of data contained in the credit bureau. The information contained in a credit bureau give a specific view of the banking sector, based on volume of credit activity, current indebtedness of citizens and companies in banks and in other credit institutions and can be a good input in providing signals for monetary and fiscal policy creators. For the regulators, such as central banks, it is extremely important to control the credit risk exposure and monitor any possible deterioration in order to take adequate measures to prevent adverse situations. Therefore, during the research the basic decision making models for credit risk analysis from this perspective were identified and, in accordance with that, an appropriate data mart and OLAP models were developed. The developed system was tested on data from the case study company, The Central Bank of Montenegro, which is owner of credit registry in Montenegro. The results were very good and significant improvement of the performance of business processes was identified.

I. INTRODUCTION

Today, information is a key resource for all enterprises and from the moment when information is created till the moment when is destroyed, technology plays a key role.

Enterprises and their executives, as in [1], strive to:

- maintain quality information to support business decisions,
- generate business value from IT-enabled investments, i.e., achieve strategic goals and realize business benefits through effective and innovative use of IT,
- achieve operational excellence through reliable and efficient application of technology.

Business intelligence systems represent one of the technological solutions that can best respond to these business requirements.

The subject of this research is modeling of business intelligence system in the area of credit risk analysis on the basis of data contained in the credit bureau, as in [3], which includes:

- identification of the basic decision-making processes connected with the evaluation and monitoring of credit risk,
- identification of data sources for identified processes,
- design of data mart for supporting identified decision-making processes,
- defining the concept and methodology for developing defined data mart by using business intelligence tools,
- implementation of OLAP model for supporting identified decision-making processes.

Developed system was tested on real data stored in the IT system that is used for supporting business processes within credit registry in Central Bank of Montenegro. The obtained results pointed out that contribution of developed system was in significant improvement of quality and performance of identified business processes.

II. IDENTIFICATION OF THE DECISION-MAKING MODELS FOR CREDIT RISK ANALYSIS

Pool of data and information contained in credit registry can be a good input in providing signals for monetary and fiscal policy creators in area of credit risk analysis. In controlling the credit risk exposure, early warning identification is important in order to take adequate measures to prevent adverse situations and react already in prevention phase. In this way, in addition to systemic stability at the aggregate level (macro contribution to the economic system's financial stability), there is micro contribution to individual institution's stability through the improvement of business performances of commercial banks.

This is especially important for the banking systems, such as Montenegrin banking system, in which the credit risk is the dominant.

When deciding on the credit risk analysis, the most important data refers to debt balance. To wit, each credit means repayment of an agreed amount in an agreed manner and in an agreed schedule. Tracking the repayment of debt by monitoring the remaining debt (debt

balance) and its trending from different perspectives gives the idea of credit risk at the system level and at the individual institution's level.

The following decision-making models are important in the credit risk analysis process.

- *The analysis of the share of institutions approving loans in total claims* – The division of debt balance by institutions granting debts to sectors (banks, microcredit financial institutions (MFIs), the Investment and Development Fund (IDF)) in the economy gives an overview of concentration of entities granting different types of loans. Banks are the only institutions in the system accepting deposits and, consequently, they are subject to stricter risk control rules. Safety and soundness of the banking sector in the country contributes largely to safety and soundness of the entire financial system. Therefore, it is of utmost importance to monitor the growth and development of each of these subsystems in order to timely assess the level of risks that each of these institutions generates and, if appropriate, tailor measures in case of vulnerabilities identified in the financial stability.
- *The analysis of structure of claims by sectors* – Claims (loans) by industries provide the analytical basis for regulators to identify and rank risks according to this classification of sources. This is of utmost importance for creating strategic directions in streamlining economic policy. Moreover, regulators like the CBCG get the information on the direction which they should pursue and direct their policy with a view to preserving financial stability.
- *The credit portfolio risk analysis* – Extremely important information also include that on assets classification, i.e. credit claims, as the primary indicator of a debtor's risk profile, and/or the entire bank's debt profile. To wit, most analysts agree that assets' quality, as the most important item, is a statistically important for forecasting a bank's insolvency. A higher level of non-performing assets in a bank may lead to its bankruptcy and to reducing the banking system's efficiency. The build-up of non-performing loans (NPLs) in banks' balance sheets is one of the main symptoms in a banking crisis. If NPLs are on an ongoing uptrend, resources get locked in non-profit sectors, thus preventing economic growth, ultimately aggravating the country's economic efficiency. Therefore, it is of utmost importance to monitor these indicators and take proper measures for controlling credit risk, especially in case of systemically important banks which sizeable market share can jeopardize financial stability.
- *The analysis of structure of claims by purpose* – Distribution of claims by purpose provides information on the type of consumption dominating in the system. A high concentration of loans granted for final consumption may signalise the situation of demand overheating and creation of price bubbles in the market, which is unsustainable in a long term. Overheated demand and bubble burst in the next phase jeopardize financial stability. The current crisis has revealed these trends, therefore it is extremely important to identify these trends and, consequently, take appropriate measures.

III. DATA MART IMPLEMENTATION

A. *Defining the requirement for the data mart*

First thing in process of implementation of the data mart is to define the business and technical requirements.

Basic request is that the data stored in data mart should be at a meaningful level of detail for the kind of analysis that the end users want to perform, and should be presented in the business terms that they understand. The expectation is that the analysis of the data in a data mart will lead to more informed business decisions.

In accordance with that and decision-making processes identified in previous section, data mart for credit risk analysis should provide development of such OLAP models that will provide:

- the analysis of the share of institutions approving loans in total claims individually and by type of credit institution and time,
- the analysis of structure of claims by sectors and category of users within the sector by type of credit institution, time and purpose of debit,
- the credit portfolio risk analysis by type of debit base and time.

On the other side, there are several technical requirements that should be met:

- data mart must be realized in such way that can be included in integral data warehouse,
- it should be developed by using some of standard data warehouse development tools and
- it should enable integration with standard OLAP tools.

In addition to this, because we are using data that contain some information with high level of confidentiality such as loan debt of physical and legal persons, it is extremely important to have possibility to implement adequate security measures and access right management.

Also, in this phase of data mart design, data sources, the kind of data cleansing needed, and the frequency with which data should be refreshed are identified.

B. *Data mart design - selection and specification of business facts and dimensions*

For representation of fact, dimensions, and the relationships between them in data mart application will be used the star schema. Typically, it contains a dimension of time and is optimized for access and analysis. Fact table

is in the centre of star schema connected by surrogate keys with dimension tables. Dimension tables are denormalized with surrogate keys that enable storing of dimension history. This dimensional model will provide ROLAP type of analytical processing. Also, data mart that will be realized will use data from the relational database in OLTP source.

After defining the fact and dimensions, it is important to determine the appropriate granularity for the data in the data mart and estimate the resource requirements to provide the requested level of granularity. In practice, based on the costs, at this stage is making decision whether or not to support this level of granularity.

In accordance with the aforementioned, star schema of the designed dimensional module with the requirements recognized in the previous section, consist of one fact table and dimension tables: *Datum*, *Kategorija rizika*, *Kategorija korisnika*, *Kreditne institucije*, *Osnov* i *Namjena*.

Fact table *Krediti* (KREDITI_F), will contain information about indebtedness of legal persons and natural persons and about repayment of all credit obligations belonging to these entities. Attributes of this table, in addition to surrogate keys of dimension tables, contain information about the rest of the claims, the total debt and weighted effective interest rate for particular indebtedness.

Data in dimension tables are hierarchically organized in the following way:

- Time dimension, Date (DATUM_D), reflects the hierarchy of time: *year → quarter → month → day.*
- Dimension of risk (KATEGORIJA_RIZIKA_D) represents classification of loans by individual classification groups according to the likelihood of loss. This dimension corresponds to the hierarchy of type: *group of risk → subgroup of risk.*
- Dimension of user category (KATEGORIJA_KORISNIKA_D) reflect sectorial affiliation of user indebtedness. This dimension corresponds to the hierarchy of type*: sector → category of user.*
- Dimension of credit institution (KREDITNA_INSTITUCIJA_D) represent categorization by type of credit institution that approves indebtedness. This dimension corresponds to the hierarchy of type: *type of credit institution → credit institution.*
- Dimensions of debit base (OSNOV_D) and purpose of debit (NAMJENA_D) contain information about base for approving user indebtedness and purpose of approved indebtedness. These dimensions do not have defined hierarchy.

C. *Data mart realization*

For data mart realization will be used Oracle Warehouse Builder (OWB). Oracle Warehouse Builder, as in [5], is a single, comprehensive tool for all aspects of data management. Warehouse Builder leverages the Oracle Database to transform data into high-quality information. It provides data quality, data auditing, fully integrated relational and dimensional modeling, and full life cycle management of data and metadata. Warehouse Builder enables you to:

- create data warehouses,
- migrate data from legacy systems,
- consolidate data from disparate data sources,
- clean and transform data to provide quality information and
- design and manage corporate metadata.

It also providing standard extraction, transformation and loading (ETL). The process of realisation of data mart is carried out in three stages: stage of defining, generating and data uploading, as in [8].

The defining stage includes:

- defining of targeted warehouse schema - defining of multidimensional models and aggregations,
- defining of data sources and
- defining of mapping operations i.e. transformation and data uploading - ETL procedure.

In the generating stage, scripts for creating warehouse schema and ETL procedures are generated and executed. Through the execution of ETL procedures the initial data uploading is performed.

In Figure 1 is shown star schema for the data mart that is designed in the previous section and is defined by using OWB.

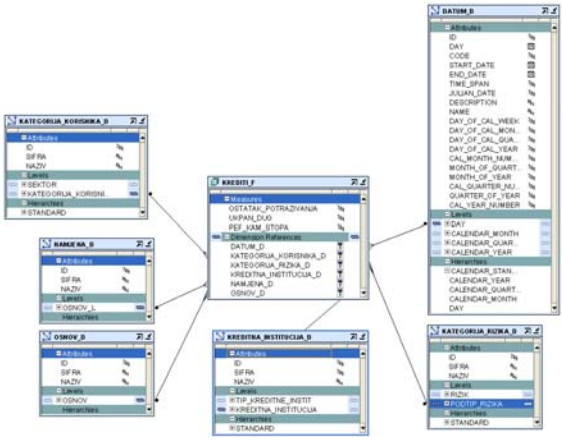


Figure 1. Star schema

IV. REALIZATION OF DECISION MAKING MODEL

For realization of decision making models (OLAP models) for credit risk analysis that are based on the analysis described in the section two, was used Oracle Business Intelligence Discoverer.

Oracle Business Intelligence Discoverer, as in [7], is an intuitive ad-hoc query, reporting, analysis, and Web publishing toolset that gives business users immediate access to information in databases. Using any standard Web browser, they have secure and immediate access to data from both relational and multidimensional data sources. Oracle Business Intelligence Discoverer provides a business view to hide the complexity of the underlying data structures, enabling end users to focus on solving business problems.

Users' requests for information from the database are in the form of worksheets.

When a user creates or opens a worksheet, Discoverer:

- converts the worksheet into the corresponding SQL statements (e.g. by converting folder names and item names to table names and column names respectively),
- sends the SQL statements to the database,
- displays the result set that is returned from the database.

It comprises a number of integrated components that work together with the Oracle database to give a complete and integrated Business Intelligence solution.

In this research, for realization of OLAP models that will support decision-making processes identified in the section three, were used two components of this tool: Oracle Business Intelligence Discoverer Desktop and Oracle Business Intelligence Discoverer Administrator. Figure 2 shows one of OLAP models defined by using this tools.

V. CASE STUDY

For the analysis within case study was used data contained in the Credit Registry whose owner is Central Bank of Montenegro.

The Central Bank of Montenegro (CBM), as in [3], is the supreme institution of the monetary system of Montenegro. By the Constitution of Montenegro it is defined as "an independent organization, responsible for monetary and financial stability and functioning of the banking system." In fulfilling the responsibilities assigned by the Constitution, the Bank among other things: supervises the banking system, performs and supervises interbank payment operations domestic and international, acts as fiscal agent, banker and advisor of the State and produces regular macro-economic analyses giving recommendations for Economic Policy to the Government.

In 2005 Central Bank of Montenegro develop a credit biro (Credit Registry), in order to provide, among other things:

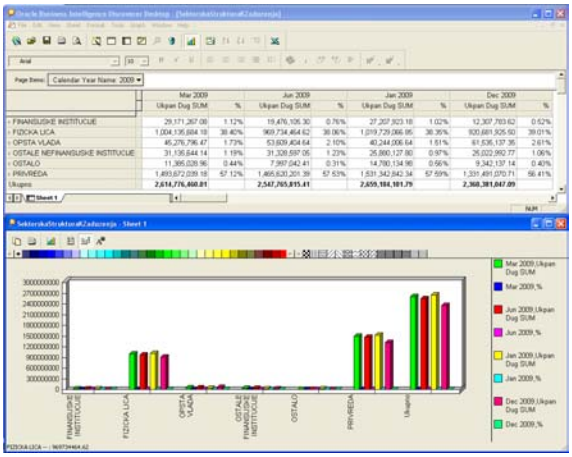


Figure 2. Workbook for indebtedness analysis in the system by sectors and time

- more complete, precise and dynamic assessment and monitoring of credit risk in banks and aggregated at the system level, monitoring the possible transfer of credit risk between banks,
- better monitoring of concentration of credit risk and business operations with related parties,
- enhance credit risk management in the banking sector and facilitate the implementation of the so-called "dynamic provisioning determination."

A Central Bank Registry is defined as a repository or data bank of records of legal persons and natural person's demographics and the payment history/manner or repayment of all credit obligations belonging to these entities and individuals. These records are supplied by financial institutions that are supervised by the Central Bank of Montenegro and include all commercial banks and microcredit financial institutions (MFI) and since 2010 Investment and Development Fund.

Source data used for designing data mart in this research represent business data stored in CBM's Credit Registry for years 2009 and 2010. Application that support business processes within credit registry was developed by using Oracle RDBMS and Oracle development tools so, data were placed in source form in Oracle relation database.

Initial data uploading into dimension tables is realized by using *Control Center Manager*, component of *Oracle 10g Warehouse Builder,* and took a few seconds. Loading data into a fact table (little less than seven million five hundred thousand records) took a bit longer - a couple of minutes.

Initial importing of data into workbooks developed over the fact table KREDITI_F, a about fifteen seconds. Operations drill-down and pivot, for the number of records, executed immediately while the operation collapse (drill-up) in some cases lasted slightly longer (a few seconds).

Several interesting analysis, as in [3], realized through OLAP models over case study data, are presented in the text below.

In Figure 3 is shown share of institutions approving loans in total claims state as of end-2010.

The analysis shows that most of the debt (almost 98 %) goes to the bank claims. Namely, over 90 % of the debt to households, businesses, government and other sectors were approved by the banking institutions. It means that Montenegrin economy is bank based with the banking system as dominant sector in the financial system. Therefore, from stability in this sector largely depends the stability of the entire system. Because of that it is extremely important to timely assess the level of risks that each of these institutions generates and, if appropriate, tailor measures in case of vulnerabilities identified in the financial stability.

Having in mind that the highest percentage of debit in the system goes to the banks, it is interesting to look at the structure of these debts by sectors. The Figure 4 provides an overview of structure of claims by sectors. These reviews are related to the active bank claims on which banks are required to provide reserves for losses depending on the risk level of each placement.

As we can see, in total debt at the banking system, individuals are the second largest category. Therefore, it is interesting to analyse the structure of their debt by purpose. Workbook for this analysis is shown in the Figure 5.

What can be seen is the decrease of the total indebtedness of natural persons per quarters of year 2010, which can mean that they regularly returning undertaken obligations, but also an indication of the reduced bank lending. Also, debt of natural persons for refinancing obligations to other banks recorded a growth in comparison to the previous quarter (114,19 %) which can mean that citizens further borrowing in order to repay previously conducted commitments.

In the Figure 6 is shown status and structure of active claims by certain categories of risk in relation to the data that credit institutions have displayed at the credit registry at the end-2010.

Considering the importance of claims in the system for loans in Figure 7 is shown workbook for the analysis of the structure of loans by category of risk, separately for each subtype of risk on a quarterly basis in 2010.

Analysis of the quality of ordinary loans that in the highest percentage share in total loan structure indicates a

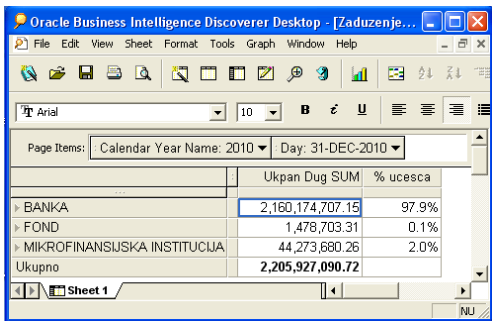


Figure 3. Share of institutions approving loans in total claims state

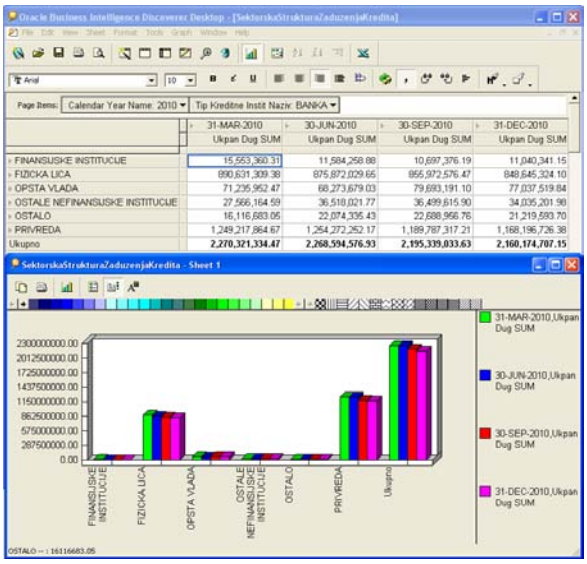


Figure 4. Structure of claims by sector

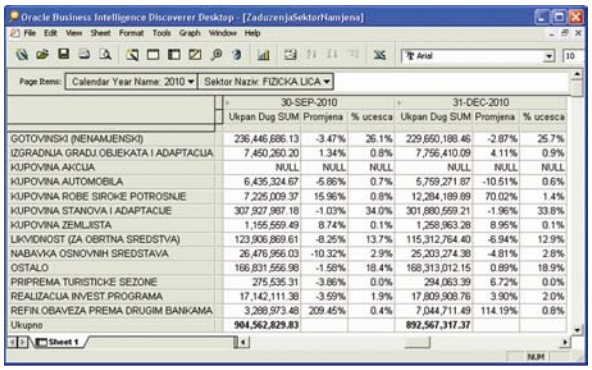


Figure 5. Distribution of claims for natural persons by purpose on the system level

high percentage of share for higher categories of risks. The share of non-performing loans (loans classified in the highest risk categories "C", "D" and "E") amounted to 17,1 %. For some categories (C1 and C2) is, compared to the previous quarter, evident increasing trend (over 200 %). Also, compared to the previous quarter, it is evident decline in the share of ordinary credit risk categories "D" and "E" of about 50 %, which is a good trend.

VI. CONCLUSION

This paper presents the development of a business intelligence model for credit risk analysis based on the data reported to the credit registry.

The process of data mart design included following:

- identification of main decision-making processes in for credit risk analysis,
- definition of data sources for identified processes,
- projection of data mart design for the support to identified decision-making processes,

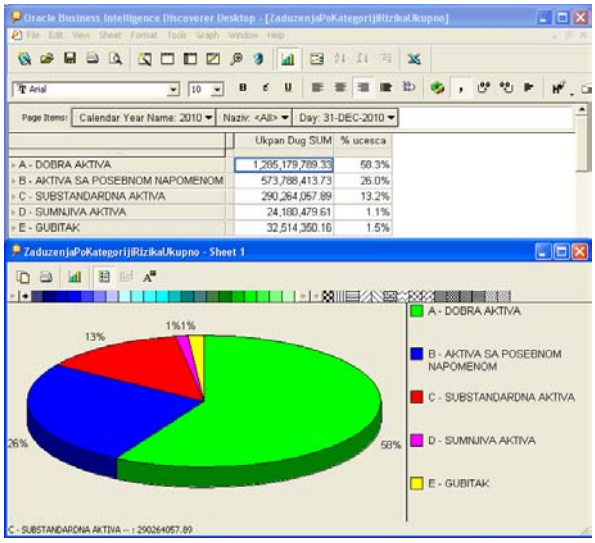


Figure 6. Distribution of claims in the system by category of risk

- definition of concepts and methodologies for the realisation of this data mart using "business intelligence" tool and
- realisation of defined targets of data mart, i.e. OLAP model;

The results of test performed for the developed model on the data available to the CBM and their comparison with the credit registry reporting subsystem that is currently in usage in the CBM's information system which is not based on business intelligence systems, provide undeniable proof of the correctness of application of this concept.

Conclusion is that implementation of such systems which are based on contemporary achievements and technologies in the areas of Data Warehousing, Data Mining and On-Line Analytical Processing (OLAP) are of immeasurable importance for supporting management functions in the Central Bank, but also for all the users of its services. Having in mind importance of such institutions for the whole society it is clear how big contribution of implementation of such systems can be.

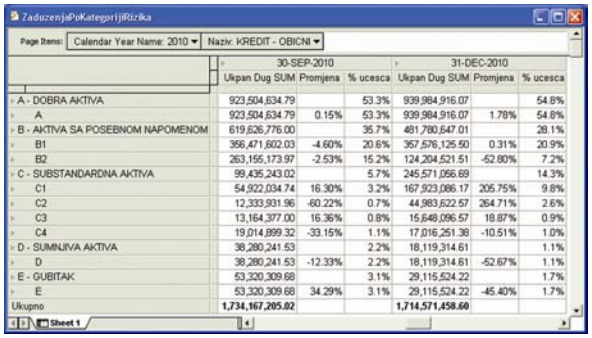


Figure 7. Classification of regular credit by risk categories on a quarterly basis in year 2010

Further research could be focused on the possibilities of integration of data mart defined in this manner into uniform data warehouse and expansion of possibilities, i.e. higher number of analysis which can include information about way of debt repayment, collaterals, security instruments that are used etc.

REFERENCES

[1] "COBIT 5", A Business Framework for the Governance and Management of Enterprise IT, ISACA, 2012.

[2] Kašćelan, Lj., "Creation of OLAP model in Oracle Discoverer ", INFO-M, No. 13/2005, p. 18-22, 2005

[3] Gazdić T.,: "Model of the business intelligence system for credit risk analysis", Master's Thesis, Faculty of Economics, Podgorica, 2012

[4] Panian Ž, Klepac G: "Business intelligence", MASMEDIA, Zagreb, 2003

[5] Oracle Warehouse Builder user's guide 10g Release 2 (10.2.0.2), PDF version, http://docs.oracle.com/cd/B31080_01/doc/owb.102/b28223.pdf , Apr 2009

[6] Oracle Business Intelligence Discoverer Desktop User's Guide, 10g Release 2 (10.1.2.1) PDF version, http://docs.oracle.com/pdf/B13917_03.pdf , July 2005

[7] Oracle Business Intelligence Discoverer Administration Guide, Release 10g, PDF version, http://docs.oracle.com/cd/B25016_08/doc/dl/bi/B13916_04/B13916_04.pdf, July 2005

[8] Kašćelan, Lj. : "Oracle Warehouse Builder and stages in development of data warehouse" , SYMOPIS 2004 - Conference Proceedings, 2004

The integral OLAP-model of the emergency risk estimation in the case of Krasnoyarsk region

A.V. Korobko, T.G. Penkova, V.V. Nicheporchuk, A.S. Mihalev
Institute of Computational Modeling of the Siberian Branch of the Russian Academy of Science, Krasnoyarsk, Russia
Siberian Federal University, Krasnoyarsk, Russia
lynx@icm.krasn.ru, penkova_t@icm.krasn.ru, valera@icm.krasn.ru, spaming1@yandex.ru

The paper focuses on implementation of the integral OLAP-model constructing method for the emergency risk estimation in the case of Krasnoyarsk region. Emergency, meteorological, hydrological, seismological, radiation, flood, housing and communal situations are observed. The process of an analytical query constructing support based on integral OLAP-model is described. Discovered nonstandard analytical dependencies between monitoring data and correspond cube-concept of the integral OLAP-model are depicted. The paper is aimed at representing of the integral OLAP-model advantages and its potential.

I. INTRODUCTION

One of the most important aspects of territory development is the efficiency of the emergency risk management. The developed and developing counties spend significant part of the budget for emergency prevention, preparedness and response. Widely spread instrumental monitoring systems allow us to accumulate and to process monitoring data for crisis management and for minimization of the emergency risk. To prevent negative events, informational decision support should be based on statistical data, expert knowledge and powerful processing tool. The integral OLAP-model combines opportunities of the OLAP technology and benefits of the Formal Concept Analysis. And it helps decision maker to get operative data in multidimensional view from primary sources in different combinations according to database structure and expert knowledge.

The paper focuses on implementation of the integral OLAP-model constructing method for the emergency risk estimation in the case of Krasnoyarsk region. Emergency, meteorological, hydrological, seismological, radiation, flood, housing and communal situations are observed. The process of an analytical query constructing support based on integral OLAP-model is described. Discovered nonstandard analytical dependencies between monitoring data and correspond cube-concept of the integral OLAP-model are depicted. The paper is aimed at representing the integral OLAP-model advantages and its potential.

The paper is structured as follows. Section II briefly describes the method of constructing integral OLAP-model of domain. Section III presents the original approach to identifying analyzed object based on data base structure, metadata and functional dependencies. Section IV demonstrates implementation of integral OLAP-model constructing method for emergencies monitoring of the region. Section V represents the paper conclusions and some issues for future research.

II. THE METHOD OF CONSTRUCTING INTEGRAL OLAP-MODEL

The method of constructing integral OLAP-model is based on integration of OLAP (On-line Analytical Processing) and FCA (Formal Concept Analysis) technologies [1,2]. The method consists in a formal concept analysis of measures and dimensions based on an expert knowledge about their comparability. As a result, the integral OLAP-model is represented as a concept lattice of multidimensional cubes.

The method of constructing integral OLAP-model consists of four main stages.

At the first stage a set of analyzed objects is identified. According to the multidimensional data model the analyzed objects are divided into two sets: the set of measures $F = \{f_1, f_2, \ldots, f_m\}$ and the set of dimensions $D = \{d_1, d_2, \ldots, d_n\}$.

At the second stage a formal context is formed. In accordance with FCA the formal context K is defined as triplet (F, D, R), which consists of set F, set D and relation $R \subseteq F \times D$. The elements of set F are measures; the elements of set D are dimensions. A binary relation R is the relation of comparability, $(f_i, d_j) \in R$ if the i-th measure can be processed with the j-th dimension. The formal context can be represented as a binary matrix where the measures are rows, the dimensions are columns and relation of comparability is a cross at intersection between a row and a column.

At the next stage the formal cube-concepts are generated based on formal context. The pair (A, B) is a cube-concept, where A is a set of equidimensional measures, which are processed with all dimensions of B. The formal cube-concept is a multidimensional cube, which is complete with respect to addition of equidimensional measures and compatible dimensions.

At the final stage the concept lattice of multidimensional cubes is constructed. The set of all formal concepts is ordered by the subcube-supercube relation. For two cube-concepts (A_1, B_1) and (A_2, B_2) this order is formalized as: $(A_1, B_1) \leq (A_2, B_2)$: $\Leftrightarrow A_1 \subseteq A_2$ ($\Leftrightarrow B_2 \subseteq B_1$). (A_1, B_1) is a subcube and (A_2, B_2) is a supercube. It means that measures of a parent cube include measures

of a child cube and dimensions of the child cube include dimensions of the parent cube. The concept lattice of multidimensional cubes is an image of the conceptual OLAP-model of the domain.

One of the most important ways of the conceptual OLAP-modeling evolution is developing approaches to analyzed objects identification on the ground of primary data base structure. Tables of different types can produce different objects for OLAP-model according to table structure (its fields). Analyzing functional dependencies chains provides us with strict method of comparability relationship detection between dimensions and measures. Then we can invoke expert for context revision and verification. It takes us opportunity to improve reliability of the final integral OLAP-model and to reduce the time of model constructing.

III. IDENTIFING ANALYZED OBJECTS IN THE DATA WAREHOUSE OF EMERGENCIES MONITORING

As a primary data base for illustrating suggested method we use existing data warehouse of emergencies monitoring for Krasnoyarsk region in the paper. In accordance with classical Ralph Kimball architecture, the considered DW has number of specific data layers [3,4]. Each of them is intended to storage data with special structure and access mode. Study the main data layer – stationary storage. It consists of two divisions: dimension tables and fact tables. In the first we can keep classificatory, nomenclatures and etc., which is changed rarely, but is used frequently. In terms of multidimensional model, this division represents tables of analytical aspects – dimensions. As usually, dimension table has structure as follow (Fig.1):

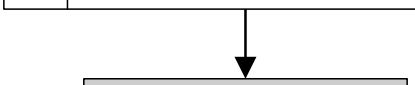
Dimension table	
PK	**Object identifier**
FK1	Object name Collateral references Service fields Object measures

Figure 1. The dimension table content

In that regard, each dimension table can give us "object name" as a dimension for our OLAP-model. And object identifier can be properly used for connection between dimension table and fact table during constructing an analytical query based on integral OLAP-model. Moreover, the table has a number of references on additional dimension tables. It takes us an opportunity to construct some kind of ordering relationship (or hierarchy) for dimension tables and as consequence, for OLAP-model dimensions. In the following, this tables feature can help us to form analytical context. Because, a measure can be processed with a dimension which has other relational dimensions, it means that this measure can be processed with child dimensions too. The object attributes produce measures for integral model.

Consider an obvious example of this approach – the table "Hydrological stations" (Fig.2). Here the field "Hydrological station" produces eponymously-named dimension. There are five dimension's numeric attributes in the table: Limit level for hydro station, Critical level for hydro station, Zero water level, Distance from hydro station to stream firth and Hydrological station altitude. These fields can be treated as measures (numeric characteristics of the analyzing process) in the our integral multidimensional analytical model. Geographical coordinates (longitude and latitude) can be used for aggregation primary facts at higher level of abstraction. Thus, the "Longitude" and "Latitude" should be added to integral model as dimensions. Also, there are several service fields in the table. For the dimension tables service fields cannot produce dimensions, but should be used during an analytical query constructing to provide user with conformal data.

Hydrological stations	
PK	**ID hydrological station**
FK1	Hydrological station **ID water object** Limit level for hydro station Critical level for hydro station Zero water level Distance from hydro station to stream firth Hydrological station altitude Geo longitudes Geo latitudes Begin date Begin reason Close date Close reason



Water objects	
PK	**ID water object**
FK1 **FK2** **FK3** **FK4**	Water object **ID water object type** **ID hydroeconomic area** **ID river basin** **ID river sub-basin** Begin date Begin reason Close date Close reason

Figure 2. The structure of dimension tables: "Hydrological stations" and "Water objects"

The field "ID water object" gives us an additional dimension – "Water object". In its turn, "Water object" dimension has a number of lateral references to other dimension tables (a number of foreign keys). Respectively, we even can get a dimension hierarchy (Fig.3).

This result gives us an opportunity to make constructing context easier. For example, the measure "Critical level for hydro station" is comparable with "Hydrological station" dimension. Technically, the measure and the dimension can be a part of one OLAP-cube – can be processed together. But for the hierarchy we know that the measure can be processed with such

dimensions as: “Water objects”, “Water object types”, “Hydroeconomic areas”, “River basins” and “River sub-basins”. Probably, dimensions hierarchy will be useful for concepts forming process. Take into consideration, the relation between dimensions is not symmetric. And if a measure can be processed with child dimension, the converse is not true.

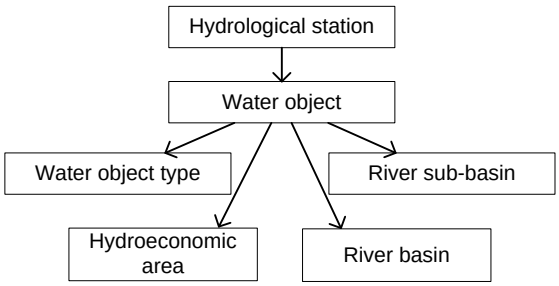


Figure 3. The ”Hydrological station” dimension hierarcy

In the second division of the stationary data warehouse storage we keep data of the instrumental monitoring, which is renewed frequently and is used frequently too. In terms of multidimensional model, this division represents tables of numeric characteristics – measures (facts). As usually, fact table has structure as follow (Fig.4):

Fact table	
PK **PK**	**Dimension identifier** **Date of measurement**
FK1	**Collateral references** Measurement data Description

Figure 4. The fact table content

The fact tables has obligatory field – “Dimension identifier” as a reference to dimension table, which is necessary for an analytic query constructing and the fact aggregation. The field “Date of measurement” is a part of composite primary key and provides multidimensional model with the time dimensions. The collateral references are foreign keys – connections with other dimension tables. They can be treated as additional dimensions for compatible processing. The time dimension is a keystone of the activity-based OLAP cubes. But for accentual inner hierarchy of the time, instrumental monitoring data can be processed on different aggregation levels: year, quartile, month, week and day. However, in some fact tables the time dimension is represent as numerical field, which stores only year or period. In future investigations it might be possible to use this finding to construct composite hierarchical dimensions. The descriptive fields are a special case among the fact table fields. In structure it looks like dimension, but originally it is an attribute of the observed object. Thus, it cannot be included into multidimensional model, because it cannot be comparable processed with other analyzed object. But it should be stored separately to be added in analytical result if it’s needed.

Consider one more example – the fact table “Hydro station data” (Fig.5).

Hydro station data	
PK,FK1 **PK**	**ID hydrological station** **Date of measurement**
FK2	**ID ice event type** Water level respectively hydro station Ice events describtion

Figure 5. The structure of the fact table ”Hydro station data”

The field “ID hydrological station” is a reference to main dimensional table “Hydrological stations”. It is needed for an analytical query constructing. The collateral references are represented by “ID ice event type” field which gives us additional comparable dimension for the measure. “Water level respectively hydro station” is a primary data of instrumental monitoring got from sensors – the goal of the analytical processing. There is one descriptive field in the fact table – “Ice event description”. While the fact table references to the dimensional table “Hydrological stations”, the measure “Water level respectively hydro station” can be processed with all dimensions from Fig.3.

The results of this study show that investigating data base structure gives us an opportunity to construct preliminary analytical context of the integral OLAP-model. It can be used by expert to finalize context and to prepare it for OLAP-model constructing.

IV. THE CONCEPTUAL OLAP-MODEL FOR EMERGENCY RISK ESTIMATION

The emergencies monitoring of the region involves the number of important domains, such as meteorology, hydrology, seismology, radiation, spring flood, housing and communal services [5,6]. Each of these domains has a reflection in the monitoring data warehouse.

There are 32 dimension tables and 65 fact tables in stationary storage. All of them are carefully studied with an authority tool – a data warehouse analyzer. The program finds 407 fields of the dimension tables and 488 fields of the fact tables. On the ground of the metadata and foreign keys, the perception about analyzed objects comparability is made. In cooperation with expert, 85 measures (Table 1) and 53 dimensions (Table 2) are identified.

Table 1.Measures of the emergencies monitoring OLAP-model

MID	Measure name
m1	Wind azimuth
m2	Wind azimuth from fire line to locality
m3	Atmosphere pressure
m4	Lateral inflow of HPP
m5	Meteorological station altitude
m6	Altitude of hydraulic construction dam
m7	Seismological station altitude
m8	Hydrological station altitude

m9	Epicentrum depth
m10	DTC
m11	Dam length
m12	Flooded transportation lines
m13	Rainfall intensity
m14	Sensor count
m15	Count of chemical hazard substances
m16	Count of hydraulic construction
m17	Flooded industrial objects
m18	Flooded buildings
m19	Count of landscape fire
m20	Number of persons with misbalanced living conditions
m21	Rainfall
m22	Flood count
m23	Number of dead persons during flood
m24	Number of dead persons during extraordinary situation
m25	Number of dead persons during emergency
m26	Number of emergency victims
m27	Number of victims during extraordinary situation
m28	Count of extraordinary situatuion
m29	Emergency count
m30	Ammonia concentration
m31	Danger substance concentration
m32	Chlorine concentration
m33	Critical level for hydro station
m34	Ice events on stream
m35	Ice events on water storage
m36	Limit level for hydro station
m37	Seismic event magnitude
m38	Maximal number of victims for emergency scale
m39	Maximal allowed level of HPP uncharge
m40	Maximal damage for emergency scale
m41	Flood level
m42	Maximal damage for grassland during flood
m43	Maximal damage for farmland during flood
m44	Minimal number of victims for emergency scale
m44	Minimal number of victims for emergency scale
m45	Minimal allowed level of HPP uncharge
m46	Minimal damage for emergency scale
m47	Seismic event energy
m48	Suboptimal HPP uncharge
m49	Suboptimal HPP volume
m50	Suboptimal HPP headwater
m51	Suboptimal HPP tail-water
m52	Suboptimal HPP inflow level
m53	Zero water level
m54	Flood frequency
m55	Flood damage
m56	Water storage volume of hydraulic construction
m57	Volume of the biggest container danger substance
m58	HPP uncharge volume
m59	Water radioactivity
m60	Danger HPP uncharge
m61	Danger HPP volume
m62	Danger HPP headwater
m63	Danger HPP tail-water
m64	Danger HPP inflow
m65	Degree of air saturation
m66	Inflow square of hydro station
m67	HPP water storage square
m68	Flood square
m69	Landscape fire square at the end
m70	Landscape fire square at the begin
m71	Weather phenomena description
m72	Threshold toxodose
m73	Landscape fire duration
m74	Flood duration
m75	Distance from hydro station to stream firth
m76	Distance from landscape fire line to locality
m77	Distance from hydraulic construction to stream firth
m78	Wind speed
m79	Lethal toxodose
m80	Average level of the HPP water storage
m81	Total resources of chemical hazard substances in dangerous object
m82	Air temperature
m83	Headwater in absolute terms
m84	Water level respectively hydro station
m85	Radiation level

Table 2. Dimensions of the emergencies monitoring OLAP-model

DID	**Dimension name**
d1	Chemical hazard substances
d2	Address
d3	Ice event types
d4	Water objects
d5	Hydroeconomic area
d6	Time-Year
d7	Time-Date
d8	Geo longitudes
d9	Geo latitudes
d10	Hydrological events
d11	Hydrological station
d12	Hydro power plant
d13	Groups of weather forecast areas
d14	Hydraulic constructions
d15	Sensor
d16	Physical quantities
d17	Financial sources during flood
d18	Extroardinary situation sourses
d19	Explosion hazard classes
d20	Burning classes
d21	Landscape fires
d22	Emergency scales
d23	Meteorological station
d24	Substance molar mass
d25	Territory
d26	Control plot
d27	Reporting office

d28	Floods
d29	Weather phenomena
d30	C horizon
d31	Hydro-meteorological observatories
d32	Flood sources
d33	River basins
d34	River sub-basins
d35	Seismological stations
d36	Monitoring networks
d37	Seismological station networks
d38	Hydraulic construction states
d39	Water object types
d40	Sensor types
d41	Landscape fire types
d42	Reporting office types
d43	Hydrological station types
d44	Emergency types
d45	Address types
d46	Hydraulic construction types
d47	Sensor types
d48	Seismic event types
d49	Safety levels of hydraulic construction
d50	Dangerous chemical objects
d51	Chemical features
d52	Extroordinary situation
d53	Emergency

The graph of the dimensions relationships consists of 53 vertexes (respectively) and 60 edges. There are 2 isolated vertexes in the graph: “Ice event types” and “C horizon”. It means that they have no relations with other dimensions. The most common dimensions “Geo longitudes” and “Geo latitudes” can be processed with 81 measures.

With help of the original concept generator, 50 cube-concepts are found. The potency of the cube-concept dimension set takes on value from 0 to 53 and the potency of the cube-concept measure set takes on value from 0 to 81. Let’s consider a demonstrative example of the analytical query constructing and its results. The conceptual OLAP-model includes all the possible combinations of analyzed objects and allows the user to discover the unexpected analytical dependencies.

When user selects measures and dimensions, assigned for analysis, the conceptual OLAP-model determines the set of cubes that correspond to user's query and form a sub-lattice. The concept lattice features allow user to determine the additional measures and dimensions to be processed with selected objects. The cube-concept of sub-lattice located on the upper level (supreme) contains maximum set of the additional measures and the cube-concept located on the lowest level (infimum) contains maximum set of the additional dimensions. So, measures set of the supreme cube-concept of the sub-lattice is a maximal set of additional measures to be add to the user query. Dimensions set of the infimum cube-concept of the sub-lattice is a maximal set of additional dimensions to be add to the user query.

The pair (X,Y) is a user analytical query, where X is a set of chosen measures and Y is a set of chosen dimensions. To support constructing user query we define a set of additional measures X’ and a set of additional dimensions Y’ which can be added in the query for each step of query forming. Consider the example of the user analytical query forming step by step.

Step 1. X = {m84}, Y = {d8,d9} →
X’ = {m1, m2, m3, m4, m5, m6, m7, m8, m9, m10, m11, m12, m13, m14, m15, m16, m17, m18, m19, m20, m21, m22, m23, m24, m25, m26, m27, m28, m29, m30, m31, m32, m33, m34, m35, m36, m37, m39, m41, m42, m43, m45, m47, m48, m49, m50, m51, m52, m53, m54, m55, m56, m57, m58, m59, m60, m61, m62, m63, m64, m65, m66, m67, m68, m69, m70, m71, m72, m73, m74, m75, m76, m77, m78, m79, m80, m81, m82, m83, m84, m85},
Y’ = {d2, d4, d5, d6, d7, d8, d9, d11, d25, d27, d31, d33, d34, d39, d43, d45};
The sub-lattice – Fig.6a;

Step 2. Y=Y∪“Territory” →
X’ = {m1, m2, m3, m4, m5, m6, m8, m11, m12, m13, m14, m15, m16, m17, m18, m19, m20, m21, m22, m23, m24, m26, m28, m30, m31, m32, m33, m34, m35, m36, m39, m41, m42, m43, m45, m48, m49, m50, m51, m52, m53, m54, m55, m56, m57, m58, m59, m60, m61, m62, m63, m64, m65, m66, m67, m68, m69, m70, m71, m72, m73, m74, m75, m76, m77, m78, m79, m80, m81, m82, m83, m84, m85},
Y’ is the same;
the sub-lattice – Fig.6b;

Step 3. X=X∪“Ice events on stream” →
X’ is the same,
Y’ is the same;
the sub-lattice – Fig.6b;

Step 4. X=X∪“Ice events on water storage” →
X’ is the same,
Y’={d4, d5, d6, d7, d8, d9, d25, d31, d33, d34, d39, d43};
the sub-lattice – Fig.6c;

Step 5. Y=Y∪“River basins” →
X’ = {m4, m6, m8, m11, m12, m16, m17, m18, m20, m22, m23, m33, m34, m35, m36, m39, m41, m42, m43, m45, m48, m49, m50, m51, m52, m53, m54, m55, m56, m58, m60, m61, m62, m63, m64, m66, m67, m68, m74, m75, m77, m80, m83, m84},
Y’ is the same;
the sub-lattice – Fig.6d;

Step 6. X=X∪“HPP uncharge volume” →
X’ is the same,
Y’={d4, d5, d6, d8, d9, d25, d31, d33, d34, d39, d43};
the sub-lattice – Fig.6e;

Step 7. Y=Y∪“Time-Year” →
X’ = {m6, m11, m12, m16, m17, m18, m20, m22, m23, m34, m35, m39, m41, m42, m43, m45,

m54, m55, m56, m58, m67, m68, m74, m77, m84},
Y' is the same;
the sub-lattice – Fig.6f;

Step 8. *X*=*X*∪" HPP water storage square" →
X' is the same,
Y'={d4, d5, d6, d8, d9, d25, d33, d34, d39};
the sub-lattice – Fig.6g;

Step 9. *Y*=*Y*∪" Water objects" →
X' = {m6, m11, m16, m34, m35, m39, m45, m56, m58, m67, m77, m84},
Y' is the same;
the sub-lattice – Fig.6h;

Commonly, on the first step, the user query is empty and the sets of the additional analyzed objects are equal to entire set F and D respectively. A user analytical query corresponds with sub-lattice of the integral OLAP-model. In our example the first step query is ({m84},{d8,d9}). It can be committed by 13 cube-concepts (fig.6a). The next step we add a dimension or a measure to the user query, update corresponded sub-lattice (the filled nodes of the sub-lattice on fig.6) and calculate a set of additional measures X' and a set of additional dimensions Y'. In its turn, the last query corresponds with only one cube-concept which gives us the sets of additional analyzed objects (fig.6h).

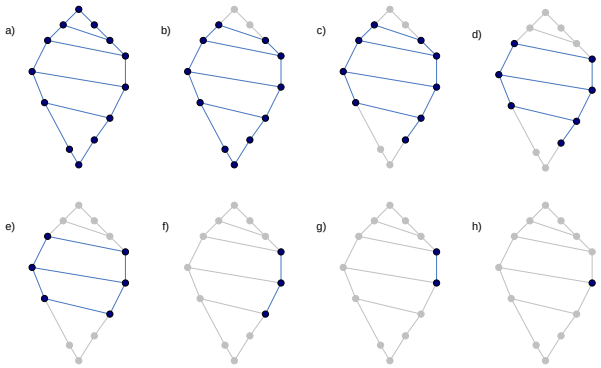


Figure 6. The sub-lattices for user query steps

Implementing the suggested method allows us to construct unusual analytical query and process unexpected analyzed object together. Thus, we have got a rare analytical result where we can compare HPP uncharged volume and water level on the nearest hydro station for chosen water object.

V. CONCLUSION

This paper has presented an approach to analytical decision making support based on integral OLAP-model of the domain. The method of constructing integral OLAP-model of domain as a concept lattice of multidimensional cubes allows user to discover the unexpected analytical dependencies based on expert knowledge about analyzed objects structure and their comparability. The effectiveness of this method was confirmed successfully by implementation for emergencies monitoring.

This finding experiment detects a number of issues to be solved during suggested method embodiments: formulation strict criteria for identifying analyzed object, development of the approach for descriptive field treatment and usage aggregation function during analytical query committing. The future research will be connected with integration of the used programming tools and evolution of the existed method.

ACKNOWLEDGMENT

This paper was supported by a grant of the Fundamental investigations of the Russian Academy of Science for 2012-2014 (No 15.3) and by a grant of the Russian Fond of Fundamental Researchers for 2012-2013 (No 12-07-31143).

REFERENCES

[1] E.F. Codd "Providing OLAP to user-analysts: An IT mandate", Codd and Associates, 1993.

[2] Penkova T. Korobko A. Method of constructing the integral OLAP-model based on formal concept analysis // Frontiers in Artificial Intelligence and Applications. – IOS Press, ISSN 0922-6389 . – 2012. – V. 243. – pp. 219-227, doi:10.3233/978-1-61499-105-2-21.

[3] Kimball R. The data warehouse toolkit: practical techniques for building dimensional data warehouse //New York, itd: Wiley. – 1996.

[4] Nozhenkova L.F., Nicheporchuk V.V., K Badmaeva.V., Penkova T.G., Korobko A.V., Evsukov A.A., Nozhenkov A.I., Markov A.A., Morozov R.V., Esavkin S.E. The system for emergencies monitoring data consolidation and analysis in the Krasnoyarsk region// Safety and emergencies problems. – 2012. – №4. – P. 63-73.

[5] Penkova T., Nicheporchuk V., Korobko A. Emergency situations monitoring using OLAP technology // Proc. 35th International Convention. The conference «Business Intelligence Systems (miproBIS)», 2012. – pp. 1941-1946.

[6] Osipov V.I. Natural hazard at the turn of the 21st century // Vestnik RAS. V. 71. №4. P. 291-302.

Automatizacija generiranja prezentacijskog sloja skladišta

S. Pavlek [1], M. Sorić [2]

[1] Koios savjetovanje d.o.o., Zagreb, stjepan.pavlek@koios.hr

[2] Koios savjetovanje d.o.o., Zagreb, mirko.soric@koios.hr

Nakon implementacije kompleksnijih skladišta podataka uglavnom se otvara niz projekata koji se bave korištenjem podataka u skladištu. Podaci se iz centralnog sloja skladišta učitavaju u prezentacijski sloj, koji je modelom prilagođen izvještajnim alatima ili određenom specijaliziranom analitičkom softveru. Naručitelji sustava mogu biti (neugodno) iznenađeni situacijom u kojoj se opet govori o razvoju ETL-a, umjesto da se podatke u skladištu jednostavno – koristi. Ideja vodilja pri razvoju sustava opisanog u radu bila je nastojanje da se put od poslovne specifikacije do gotovih ETL procedura skrati u najmanjoj mogućoj mjeri.

Specifikacije za punjenje prezentacijskih slojeva pregledno se daju u tabličnom obliku: jedne pored drugih navode se tablice i kolone potrebne u prezentacijskom sloju te tablice i kolone iz kojih se podaci trebaju uzeti, uz opise eventualnih transformacija. Takva tablica razumljiva je i laka za korištenje (održavanje) poslovnim korisnicima, a dovoljno tehnički precizna.

Informacije iz takve specifikacije u minimalno izmijenjenom obliku koriste se u razvijenom sustavu za automatsko generiranje strukture ciljnih tablica te ETL procedura potrebnih za njihovo punjenje. Time je preskočen korak ručne izrade ETL-a, a sve kasnije promjene u specifikaciji se u sustavu realiziraju automatski i trenutačno.

UVOD

Učitavanje podataka u prezentacijski sloj manje je složen proces od učitavanja podataka u samo skladište. Zadatak učitavanja u prezentacijski sloj sastoji se od poštivanja poslovnog modela te poštivanja strukture ulaznog sučelja alata za poslovnu inteligenciju ili drugih aplikacija kojima su podaci namijenjeni.

Po Kimball-u [1], modeliranje prezentacijskog sloja skladišta podataka počinje definiranjem poslovnog procesa nakon kojeg slijedi identifikacija potrebnih informacija i podataka u skladištu. Riječ je o modeliranju od vrha prema dnu (eng. *top to bottom*). Prirodno sučelje za definiciju učitavanja trebalo bi za taj pristup biti što sličnije dokumentu u kojemu je opisana poslovna specifikacija koja se traži od aplikacije. Takav dokument mogao bi biti sličan tablici u tabličnom kalkulatoru na Slici 1. U tablici su navedene odredišne tablice u prezentacijskom sloju i njihove kolone te paralelno izvorišne tablice u skladištu podataka.

Naveden je i izraz, odnosno transformacija, kojom se trebaju puniti polja odredišne tablice. Taj izraz nije nužno jednostavno preslikavanje izvorišne kolone i može sadržavati konstante, konkatenaciju, SQL funkcije, SQL analitičke funkcije i slične transformacije.

	A	B	C	E	F	G	H
1	Definicija odredišta			Definicija izvora			Komentar
2	Odredišna tablica	Odredišna kolona	Tip podatka	Izvorišna shema	Izvorišna tablica	Izraz kojim se puni kolona odredišne tablice iz izvorišne tablice	
3	CUSTOMERS	CUSTOMER_ID	VARCHAR2(100)	DWH	KLIJENTI	KLIJENT_ID	Uvjet dohvata iz tablice: DAT_TO = '31.12.9999'
4	CUSTOMERS	CUSTOMER_STATUS	VARCHAR2(10)	konstanta	KLIJENTI	'A'	Konstanta
5	CUSTOMERS	DATE_OF_BIRTH	DATE	DWH	KLIJENTI	to_date(substr(JMBG,1,2) || '.' || substr(JMBG,3,2) || '.' || decode(substr(JMBG,5,1), '0', '2', '1') || substr(JMBG,5,3), 'dd.mm.yyyy')	Datum rođenja dobivamo iz JMBG-a
6	CUSTOMERS	INSERT_DATE	DATE		KLIJENTI	SYSDATE	
7	TRANSACTIONS	TRANSACTION_ID	NUMBER	DWH	TRANSAKCIJE	TRANID	Uvjet dohvata iz tablice (transakcije sa jučerašnjim datumom obrade): DATUM_OBRADE = TRIM(SYSDATE) – 1
8	TRANSACTIONS	PAYER_ACCOUNT	VARCHAR2(100)	DWH	TRANSAKCIJE	PLATITELJ_VBDI || '-' || PLATITELJ_BR_RAC	
9	TRANSACTIONS	PAYEE_ACCOUNT	VARCHAR2(100)	DWH	TRANSAKCIJE	PRIMATELJ_VBDI || '-' || PRIMATELJ_BR_RAC	
10	TRANSACTIONS	AMOUNT	NUMBER(16,2)	DWH	TRANSAKCIJE	IZNOS	
11	TRANSACTIONS	CURRENCY	VARCHAR2(3)	DWH	TRANSAKCIJE	TRIM(VALUTA)	
12	TRANSACTIONS	WORKING_DAT	DATE	DWH	TRANSAKCIJE	DATUM_OBRADE	

Slika 1. Primjer poslovne specifikacije za učitavanje podataka u prezentacijski sloj skladišta u tabličnom kalkulatoru

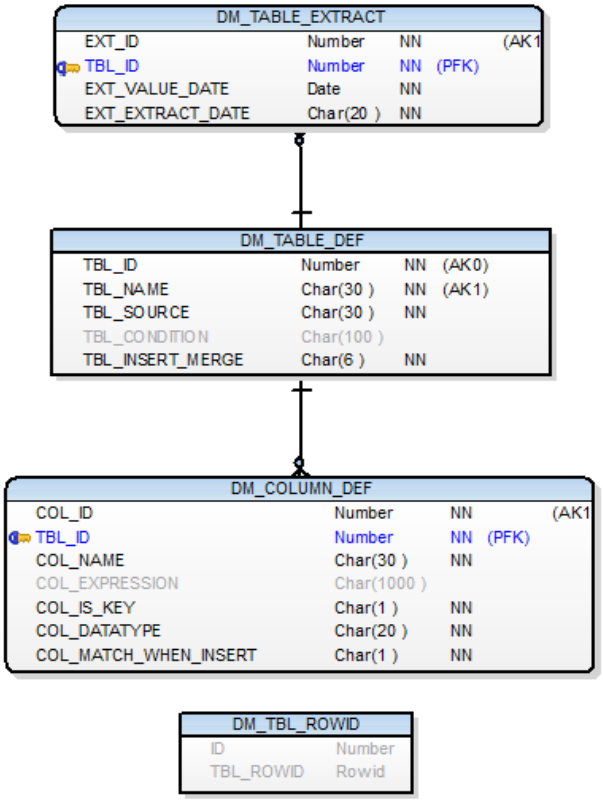
No, taj izraz nije niti suviše kompliciran – ne podrazumijeva zahtjevnije transformacije, na primjer, spajanje na druge tablice. Ograničena kompleksnost ovog izraza pretpostavka je na kojoj se ovaj pristup temelji. Takvu manju kompleksnost očekujemo prilikom punjenja prezentacijskog sloja skladišta. Ukoliko navedeni uvjet nije zadovoljen, vjerojatno je primjerenije pristupiti razvoju ETL procedura na klasičan način, koristeći općenitiji alat.

No, ukoliko je uvijet ograničene kompleksnosti transformacija zadovoljen, ručni razvoj procedura za punjenje podataka u prezentacijski sloj skladišta postaje posao koji bi se mogao automatizirati.

Osnovna ideja je razviti tehničko sučelje za definiranje učitavanja koje u najmanjoj mogućoj mjeri mijenja poslovnu specifikaciju. ETL procese želimo generirati automatski, generatorom koda, te automatizirati što više pratećih radnji pri razvoju sučelja: dodavanje tablica, kolona, izmjene specifikacije... Platforma na kojoj je sustav implementiran je baza podataka *Oracle 11gR2*.

IZVEDBA

A. *Metatablice*



Slika 2. Model metatablica

Definiraju se četiri metatablice u kojima će biti sadržana pravila preslikavanja podataka i koje će služiti za kontrolu učitavanja (Slika 2.):

- U dvije metatablice definirano je uparivanje izvorišnih i odredišnih kolona i tablica na sličan način kao u poslovnoj specifikaciji koja je navedena kao primjer u uvodu (Slika 1.). To su tablice DM_TABLE_DEF (Tablica I.) i DM_COLUMN_DEF (Tablica II.). Struktura je slična repozitoriju metapodataka (eng. *Data Dictionary*) u bazi podataka *Oracle* [2].

TABLICA I. METATABLICA DM_TABLE_DEF

Ime atributa	*Opis*
TBL_ID	Surogatni ključ tablice
TBL_NAME	Ime odredišne tablice
TBL_SOURCE	Izvor učitavanja definiran u bazi podataka. Izvor može biti tablica, pregled (eng. *view*), sinonim ili tablična funkcija
TBL_CONDITION	Uvjet dohvata podataka s izvora koji se umeće u WHERE dio izraza za dohvat podataka
TBL_INSERT_MERGE	Način ućitavanja odredišnje tablice: „MERGE“ ili „INSERT“

- Tablica DM_TABLE_EXTRACT sadrži podatke o učitavanjima. Predviđeno je periodičko punjenje odredišnih tablica (dnevno, mjesečno, tjedno, itd.). Ova tablica sadrži osnovne podatke o učitavanjima (datum učitavanja, datum na koji podaci vrijede). U svakoj odredišnoj tablici umetnut je strani ključ (EXT_ID) koji se referencira na ovu tablicu.
- Tablica DM_TBL_ROWID služi za optimizaciju učitavanja i detaljno je opisana kasnije.

Poslovna specifikacija prikazana na Slici 1. preslikava se u tablice DM_TABLE_DEF (Tablica III.) i DM_COLUMN_DEF (Tablica IV.) tako da se odredišne tablice sa uvjetom dohvata definiraju u tablici DM_TABLE_DEF a atributi tablica, tipovi podataka i izrazi za dohvat atributa u tablici DM_COLUMN_DEF.

TABLICA II. METATABLICA DM_COLUMN_DEF

Ime atributa	*Opis*
COL_ID	Surogatni ključ tablice
TBL_ID	Vanjski ključ na tablicu DM_TABLE_DEF, Govori kojoj tablici atribut pripada
COL_NAME	Ime atributa
COL_EXPRESSION	Izraz kojim se puni atribut
COL_IS_KEY	Označava je li atribut dio poslovnog ključa odredišne tablice (za vrijednost „Y“) ili ne („N“)
COL_DATATYPE	Tip podatka atributa.
COL_MATCH_WHEN_INSERT	Ako atribut ima vrijednost „Y“ prate se izmjene atrebuta na izvoru, ako je vrijednost „N“ izmjena atributa na izvoru pri učitavanju neće rezultirati ažuriranjem odredišne tablice

B. *Programski paket DM_ENGINE*

Sve funkcionalnosti sustava definirane su u paketu DM_ENGINE. Definirane su procedure za:

- Generiranje odredišne tablice prema metapodacima (argument je ime odredišne tablice).
- Učitavanje podataka iz izvorišne u odredišnu tablicu (argument je ime odredišne tablice).
- Učitavanje svih tablica odjednom.

D. Učitavanje podataka u odredišne tablice

Dizajn skladišta podataka kako ga je definirao Kimball [1] ima dvije glavne vrste tablica: činjenične i dimenzijske. Činjenične tablice tipično sadrže podatke kao što su bankovne transakcije, podatke o prodajnim volumenima i slično. Činjenične tablice se uglavnom ne ažuriraju nego se u njih peroidički umeću novi reci u

TABLICA III. PRIMJER DEFINICIJE ODREDIŠNIH TABLICA U METATABLICI DM_TABLE_DEF

TBL_ID	TBL_NAME	TBL_SOURCE	TBL_CONDITION	TBL_INSERT_MERGE
1	CUSTOMERS	KLIJENTI	DAT_TO = TO_DATE('31.12.9999','DD.MM.YYYY')	MERGE
2	TRANSACTIONS	TRANSAKCIJE	TRIM(DATUM_OBRADE) = TRIM(SYSDATE) - 1	INSERT

C. Genereriranje odredišne tablice

Izvršavanje procedure za kreiranje odredišne tablice teče sljedećim tokom:

- U repozitoriju metapodataka baze podataka provjerava se je li već kreirana tablica s definiranim imenom. Ako nije, generira se izraz za kreiranje tablice koristeći ime tablice, ime atributa i tipove podataka atributa definiranih u DM_TABLE_DEF i DM_COLUMN_DEF. Izraz se po generiranju izvršava.
- Ako tablica već postoji preskače se kreiranje tablice. Provjerava se, također u repozitoriju metapodataka *DBMS-a*, postoje li svi atributi koji su navedeni u definiciji. Ako neki od atributa fizički ne postoji u odredišnoj tablici generira se izraz koji mijenja odredišnu tablicu (izraz *ALTER TABLE*) i dodaje taj atribut. Atributi koji fizički postoje u tablici, a ne postoje u definiciji se ne uklanjaju iz tablice.

skladu sa poslovnim procesom koji se prati. Posjeduju strane ključeve na dimenzijske tablice koje se ne ažuriraju periodički nego po potrebi, odnosno kada se dogodi promjena u izvorišnim tablicama koju je potrebno učitati. Postoji više načina čuvanja povijesti izmjena dimenzijskih tablica. Jedan od pristupa je *SCD* (eng. *Slowly changing dimension*) tip II [4] u kojem se pri izmjeni atributa postojećeg retka u dimenzijskoj tablici umeće novi redak s aktualnim podacima (jedina razlika u odnosu na postojeći redak u tablici je atribut dimenzije koji je izmijenjen). Redak koji je prethodno postojao u tablici pomoćnim se atributom označi kao nevažeći. Na taj način se čuva povijest izmjena dimenzijske tablice.

Sukladno toj podjeli, u sustavu je moguće generirati dva tipa upita za učitavanje podataka: upit za umetanje podataka (eng. *insert*) i upit za spajanje podataka (eng. *merge* ili *upsert*). Informacija o tome koji upit se generira za pojedinu tablicu nalazi se u metatablici definicije DM_TABLE_DEF u atributu TBL_INSERT_MERGE (Tablica I.). Za vrijednost atributa „INSERT“ generira se

TABLICA Iç. DEFINICIJA TABLICA IZ PRIMJERA U METATABLICI DM_COLUMN_DEF

COL_ID	TBL_ID	COL_NAME	COL_EXPRESSION	COL_DATATYPE	COL_IS_KEY	COL_MATCH_WHEN_INSERT
1	1	CUSTOMER_ID	KLIJENT_ID	VARCHAR2(100)	Y	Y
2	1	CUSTOMER_STATUS	'A'	VARCHAR2(1)	N	Y
4	1	DATE_OF_BIRTH	to_date(substr(JMBG,1,2) || '.' || substr(JMBG,3,2) || '.' || decode(substr(JMBG,5,1), '0', '2', '1') || substr(JMBG,5,3), 'dd.mm.yyyy')	DATE	N	Y
5	1	INSERT_DATE	SYSDATE	DATE	N	N
6	2	TRANSACTION_ID	TRANID	NUMBER	Y	Y
7	2	PAYER_ACCOUNT	PLATITELJ_VBDI || '-' || PLATITELJ_BR_RAC	VARCHAR2(100)	N	Y
8	2	PAYEE_ACCOUNT	PRIMATELJ_VBDI || '-' || PRIMATELJ_BR_RAC	VARCHAR2(100)	N	Y
9	2	AMOUNT	IZNOS	NUMBER(16,2)	N	Y
10	2	CURRENCY	TRIM(VALUTA)	VARCHAR2(3)	N	Y
11	2	WORKING_DAT	DATUM_OBRADE	Y	N	Y

izraz za umetanje podataka, a za izraz „MERGE“ generira se izraz za spajanje podataka.

- Učitavanje odredišne tablice umetanjem podataka

Kada je vrijednost atributa TBL_INSERT_MERGE u metatablici DM_TABLE_DEF jednaka „INSERT“ podaci se u nju učitavaju *SQL* izrazom *INSERT INTO*.

Primjer tako definirane odredišne tablice dan je u Tablici III. i Tablici IV. To je tablica TRANSACTIONS koja se učitava iz izvorišne tablice TRANSAKCIJE.

Sintaksa izraza *INSERT INTO* definirana je na sljedeći način [5]:

```
INSERT INTO <ime odredišne tablice>
(<lista atributa odredišne tablice>)
SELECT <lista izraza za punjenje odredišne tablice>
FROM <ime izvorišne tablice>
WHERE <uvjet dohvata>
```

Izraz se kreira na sljedeći način:

- na mjesto *<ime odredišne tablice>* umeće se vrijednost atributa TBL_NAME metatablice DM_TABLE_DEF (Tablica I.)

- na mjesto *<lista atributa odredišne tablice>* umeću se imena atributa odredišne tablice definirani u atributu COL_NAME metatablice DM_COLUMN_DEF i sistemski atributi ID, EXT_ID i IS_CURRENT.

- na mjesto *<lista izraza za punjenje odredišne tablice>* umeću se izrazi definirani u atributu COL_EXPRESSION metatablice DM_COLUMN_DEF, sistemska kolona ID puni se automatski generiranom sekvencom, IS_CURRENT konstantom „Y“ a EXT_ID ključem koji pokazuje na redak u tablici DM_TABLE_EXTRACT (Slika 2.).

- Izraz za spajanje podataka

Ovaj izraz nešto je kompliciraniji. Mora zadovoljiti SCD tip II. To znači sljedeće [4]:

- Redak s novim ključem definiranim u metatablici koji se umeće u odredišnu tablicu ima vrijednost atributa IS_CURRENT jednaku „Y“.
- Redak koji više ne postoji na izvoru poprima novu vrijednost atributa IS_CURRENT jednaku „N“.
- Redak kojemu se na izvoru izmijenila vrijednost atributa koji u metatablici DM_COLUMN_DEF ima definirano svojstvo COL_MATCH_WHEN_INSERT jednako „Y“ umeće se u odredišnu tablicu s vrijednošću atributa IS_CURRENT jednakoj „Y“. Atribut IS_CURRENT u staroj verziji retka ažurira se na „N“.

Da bi dobili ovakvo ponašanje pri učitavanju koristi se jedna pomoćna tablica (DM_TBL_ROWID, na Slici 2.) i izraz za višestruko umetanje redaka (eng. *Multi-table insert*) [6]. Izraz se generira koristeći podatke iz metatablica na sličan način kao i izraz za umetanje podataka ali poštujući sintaksu izraza za višestruko umetanje redaka. Provjerava se razlika slike izvorišnih podataka (po uvjetu dohvata) i slika aktualnih podataka u odredišnoj tablici (tj. redaka kojima je vrijednost atributa IS_CURRENT jednaka “Y”). Ako postoji razlika izraz umeće retke u dvije tablice:

- U odredišnu tablicu umeću se reci koji postoje u slici izvorišnih podataka, a ne postoje u slici aktualnih podataka odredišne tablice ili im se podudaraju poslovni ključevi a ne podudaraju im se atributi kojima se prate promjene;
- U tablicu DM_TBL_ROWID umeću se parovi atributa ID i pseudoatributa ROWID [7] odredišne tablice za retke koji postoje u slici odredišne tablice a ne postoje u slici izvorišnih podataka ili im se podudaraju poslovni ključevi a ne podudaraju im se atributi kojima se prate promjene.

TABLICA V. ODREDIŠNA TABLICA CUSTOMERS

Ime atributa	***Opis***
ID	Surogatni ključ tablice
EXT_ID	Vanjski ključ na tablicu DM_TABLE_EXTRACT. Omogućava praćenje povijesti učitavanja svakog učitanog redka.
IS_CURRENT	Oznaka važećeg retka. Atribut može imati vrijednost 'Y' i 'N'.
CUSTOMER_ID	Atributi generirani iz definicije
CUSTOMER_STATUS	
DATE_OF_BIRTH	
INSERT_DATE	

Nakon umetanja podataka odredišna tablica se ažurira na način da se svim recima čiji su parovi (ID, ROWID) umetnuti u tablicu DM_TBL_ROWID ažurira atribut IS_CURRENT na “N”. Time su navedeni reci uklonjeni iz slike aktualnih podataka. Prije svakog pokretanja učitavanja brišu se svi podaci iz tablice DM_TBL_ID. Operacije umetanja podataka u odredišnu tablicu, umetanja podataka u pomoćnu tablicu i ažuriranje odredišne tablice obavljaju se u jednoj transakciji baze podataka. Ažuriranje traje vrlo brzo bez obzira na veličinu odredišne tablice jer se recima pristupa preko pseudoatributa ROWID, koji je zapis fizičke adrese retka u tablici, što je najbrži mogući način [7].

Opisani način učitavanja podataka postavlja određena ograničenja za izraze koji se mogu pojaviti u transformacijama atributa i u selekciji:

- izraz transformacije ne smije sadržavati zareze osim ako zarezi nisu dio učahurenog izraza (npr. *DECODE(...)*)
- transformacije ne smiju sadržavati agregatne funkcije
- nije moguće neposredno učitavati podatke iz dviju ili više spojenih tablica

Drugo i treće ograničenje mogu se zaobići korištenjem *view*-a u kojem se definira željeno spajanje tablica ili agregacija. Tako definirani *view* se postavlja kao izvor u metatablici. DM_TABLE_DEF.

E. *Praćenje povijesti učitavanja*

Svaka odredišna tablica, bez obzira na tip učitavanja, posjeduje atribut koji je vanjski ključ na metatablicu s evidencijom učitavanja (*DM_TABLE_EXTRACT* na Slici 1.). Ta tablica sadrži datum učitavanja, datum na koji podaci vrijede i vanjski ključ na definiciju učitavanja.

Prije svakog pokretanja učitavanja dodaje se novi redak u tablicu s evidencijom učitavanja. Ključ tog retka (EXT_ID) koristi se pri učitavanju podataka u odredišne tablice. Na taj način je moguće čuvati više informacija o učitanom setu podataka.

Sve navedene radnje povezane su u proceduri unutar paketa DM_ENGINE koja pokreće učitavanje podataka za pojedinu tablicu.

Po potrebi model je moguće proširiti statusom učitanog seta podataka, npr. informacijom o dostupnosti seta podataka ako se podaci periodički arhiviraju ili brišu ili informacijom o validaciji i provjeri kvalitete učitanih podataka.

F. *Optimizacija odredišnih tablica*

Obzirom na prirodu odredišnih tablica, struktura omogućuje dva prirodna pristupa optimizacije dohvata podataka baziranim na particioniranje sistemskih atributa:

- Činjenične tablice

  Particioniranje po vanjskom ključu na tablicu za evidenciju učitavanja (EXT_ID) omogućuje dohvat svakog pojedinačno učitanog seta podataka. Pri svakom dohvatu podataka dohvaćaju se samo podaci za željeni vremenski period, bez viška podataka koji se pri dohvatu filtriraju. Također, olakšano je arhiviranje starih podataka: moguće je starije particije koje se više ne koriste premjestiti ili izbrisati bez prekida u dostupnosti podataka.

- Dimenzijske tablice

  Moguće je koristiti particioniranje tablice po vanjskom ključu na metatablicu evidencije učitavanja (EXT_ID), po atributu oznake aktualnosti reda (IS_CURRENT) ili kombinacija particioniranja po ova dva atributa. Takvo particioniranje omogućuje dohvat aktualnih podataka, najnovijih izmjena (razlika u odnosu na prethodno aktualne podatke: *delta*) te arhiviranje prebacivanjem starih podataka u rezervnu kopiju. Za particioniranje tablice po atributu IS_CURRENT mora biti omogućena migracija podataka iz jedne u drugu particiju [8]. Ažuriranje tog atributa je u tom slučaju skuplja operacija jer uključuje fizičko pomicanje podataka na disku. Međutim, ako je količina izmijenjenih podataka pri svakom učitavanju relativno mala a odredišna tablica velika kompromis pri učitavanju se može isplatiti prilikom (čestog) dohvata podataka.

ZAKLJUČAK

Opisani sustav odlikuje jednostavnost konfiguracije i brzina razvoja. Jedini elementi sustava su metatablice i programski paket s procedurama za učitavanje podataka. Upiti za učitavanje podataka pri svakom se učitavanju generiraju iznova. Sustav ne zahtjeva ručno ažuriranje i testiranje koda za pojedino učitavanje u slučaju implementacije izmjena u učitavanju. Takav pristup donosi dvije prednosti:

- Izmjene u poslovnoj specifikaciji koje rezultiraju nužnošću izmjene tehničke specifikacije realiziraju se brzo što smanjuje vrijeme analize rezultata i konačno trajanje implementacije.
- Olakšana je prenosivost sustava s jedne baze podataka na drugu. Instalacija obuhvaća postavljanje programskog paketa DM_ENGINE u željenu bazu i kreiranje metatablica. Prenošenje definicije učitavanja obavlja se kopiranjem podataka u metatablice.

Najveći nedostatak sustava proizlazi iz nemogućnosti neposrednog definiranja izvora iz spojenih ili agregiranih tablica. Takvi izvori mogu se simulirati *view*-ovima koji se u metatablicama navode kao izvori učitavanja. Na taj način je moguće implementirati široki spektar pretakanja podataka ali se istovremeno gubi jednostavnost izmjena implementacije ETL-a kroz izmjene specifikacije.

Takav način predstavlja i određene izazove u implementaciji podrške za praćenje toka podataka – problem mogućnosti automatikom podržane analize utjecaja izvornih podataka na odredišne, odnosno analize porijekla ukoliko se problem promatra u suprotnom smjeru. Da bi se došlo do izvornog atributa više nije dovoljno pregledati metatablicu s definicijom učitavanja nego treba analizirati upite kroz koje se taj atribut dohvaća odnosno integrirati se s odgovarajućim metapodacima ETL alata u kojem je predprocesiranje implementirano.

Model učitavanja i definicija učitavanja pogodni su za proširenja, npr.:

- modul za učitavanje podataka u datoteke
- modul za validaciju podataka i automatsku kvalitativnu analizu

što otvara prostor za daljnje analize i razvoj.

Sustav je testno implementiran u projektu u kojem je cilj bio generirati datoteke za dostavu podataka analitičkoj aplikaciji potrošaču. Model je definirao 13 odredišnih datoteka, od toga četiri velike (klijenti, ugovorni odnosi, veze između klijenata i ugovornih odnosa te transakcije) i 9 manjih datoteka s popratnim i referencijskim tablicama. Generirane datoteke, odnosno tablice koje ih predstavljaju, zajedno su imale oko 600 atributa.

Definicije za generiranje odredišnih tablica iskorištene su i za generiranje koda koji podatke iz baze podataka prenosi u datoteke. Takav pristup bio je optimalan za generiranje datoteka, jer su se specifična pravila za formatiranje datoteka mogla kodirati na jednom mjestu. Funkcionalnost za specifični format datoteka kombinirana je s definicijama odredišnih tablica kako bi se dobilo automatskeo generiranje koda za kreiranje svake pojedinačne datoteke.

Vrijeme potrebno za implementaciju ETL-a na način opisan u ovom radu, što je uključivalo dizajn sustava, modela metapodataka te prve verzije biblioteka za generiranje koda – bilo je otprilike dvostruko veće od procijenjenog vremena potrebnog za implementaciju ETL procedura na klasični način.

U nekoliko mjeseci koji su slijedili stavljanju sustava u pogon, kontinuirano je mijenjana specifikacija učitavanja te su rađene manje dorade u definiciji mapiranja te obuhvatu polja. Implementacija takvih izmjena bila je značajno jednostavnija i brža nego u slučaju postupka izmjene i postavljanja na produkcijsku okolinu ETL procedura stvorenih na klasičan način. Možemo reći da je u dijelu života nakon početnog puštanja u rad sustav pokazao pravu snagu, kroz efikasniju podršku i održavanje.

Kako je sustav razvijen u sklopu relativno malog projekta – model sa samo 13 tablica – time što je već na toj veličini problema pokazao efikasnost, pokazano je da je dizajn u ovom smjeru imperativ u izrazito velikim projektima.

LITERATURA

[1] Kimball, R., Ross, M., *The Data Warehouse Toolkit, Second Edition: The Complete Guide to Dimensional Modeling*, John Wiley & Sons, Inc., 2002.

[2] http://docs.oracle.com/cd/B28359_01/server.111/b28310/tables014.htm, *Oracle® Database Administrator's Guide* , preuzeto u veljači 2013.

[3] http://en.wikipedia.org/wiki/Data_dictionary, preuzeto u veljači 2013.

[4] http://en.wikipedia.org/wiki/Slowly_changing_dimensio, preuzeto u veljači 2013.

[5] http://docs.oracle.com/cd/B13789_01/server.101/b10759/statements_9014.htm#SQLRF01604, *Oracle® Database SQL Reference*, preuzeto u veljači 2013.

[6] http://www.oracle-developer.net/display.php?id=209 *Multitable inserts*, preuzeto u veljači 2013.

[7] http://docs.oracle.com/cd/B19306_01/server.102/b14200/pseudocolumns008.htm, *ROWID Pseudocolumn*, *Oracle® Database SQL Reference,* preuzeto u veljači 2013.

[8] http://docs.oracle.com/cd/B10501_01/server.920/a96521/partiti.htm, *Managing Partitioned Tables and Indexes, Oracle9i Database Administrator's Guide*, preuzeto u veljači 2013.

Upravljanje matičnim podacima u suradnji sa skladištem podataka

Aleksandar Matić, Marin Sekula, Davor Udier
Multicom d.o.o., Zagreb, Hrvatska
aleksandar.matic@multicom.hr, marin.sekula@multicom.hr, davor.udier@multicom.hr

U radu se opisuju karakteristike Multicom sustava za upravljanje matičnim podacima (engl. *Master Data Management,* MDM) te tehnologije korištene za razvoj njegovih komponenata. Osnovni su zadaci implementiranog sustava MDM sinkronizacija podataka o korisnicima među različitim sustavima i kontrola kvalitete tih podataka te odgovaranje na trajno velik broj upita o *master* zapisima korisnika (*Single-Source-Of-Truth* za matične podatke kupaca). Rješenje se sastoji od centralnog sustava MDM, prilagodbe izvornih sustava kako bi mogli komunicirati sa sustavom MDM te komponenata *middleware* na kojima su izloženi servisi. Centralni sustav MDM razvijen je u tehnologiji *Oracle PL/SQL*, a Web sučelje razvijeno je korištenjem tehnologija Java i Javascript. Svaka promjena u izvornim sustavima šalje se kao poruka u centralni sustav. Glavne su faze obrade poruke kontrola kvalitete podataka (mehanizam karantene tj. zadržavanje poruka s nekvalitetnim podacima radi ručnog odobrenja), povezivanje s postojećim *master* zapisom ili stvaranje novog *master* zapisa (*entity resoution* korak), ponovna izgradnja *master* zapisa iz svih povezanih izvornih zapisa (*data integration* korak) te propagacija promjena slanjem poruka izvornim sustavima. Sustav za skladište podataka (engl. *Data Warehouse*, DWH) je povezan s MDM sustavom te pruža dodatnu kvalitetu u smislu nadopune podataka.

I. UVOD

IT okruženje u velikom je broju poslovnih subjekata rascjepkano na način da se isti podaci nalaze i održavaju kroz više sustava. Osim poteškoća u održavanju većeg broja sustava s istom namjenom dolazi do nejasnog preklapanja, odnosno redundancije podataka u postojećim sustavima jer svaki od njih prikuplja podatke za svoje određene potrebe.

Donedavno je takva situacija postojala i u *Hrvatskom Telekomu* u kojem su podaci o kupcima prikupljani kroz više odvojenih sustava na mobilnoj i fiksnoj strani, u prvom redu zbog integracije kompanije, a dodatno i zbog većeg broja sustava koji su implementirani s vlastitom bazom korisnika ciljano za specifične potrebe određenog dijela poslovanja (prodaja s vlastitim *CRM/Order management* aplikacijama, financije – *SAP*, operativni sustavi poput *Billinga* sa svojom bazom korisnika i sl.).

Potreba da se objedine ti podaci tj. da se dobije jedinstveni prikaz kupca sa svom njegovom imovinom i sa svim poveznicama tijekom je vremena postajala sve izraženija pa je na kraju i donesena odluka da se sustavi povežu na način da se informacije o kupcima uzimaju iz centralnog sustava.

Implementirano rješenje za konsolidaciju podataka o HT-ovim kupcima jest *Multicom MDM sustav* koji kroz integraciju s izvornim sustavima prikuplja sve transakcije vezane uz matične podatke, hijerarhijski ih integrira ispod zajedničkog kupca (engl. *master*) te vrši propagaciju pročišćenih podataka u sve spojene sustave.

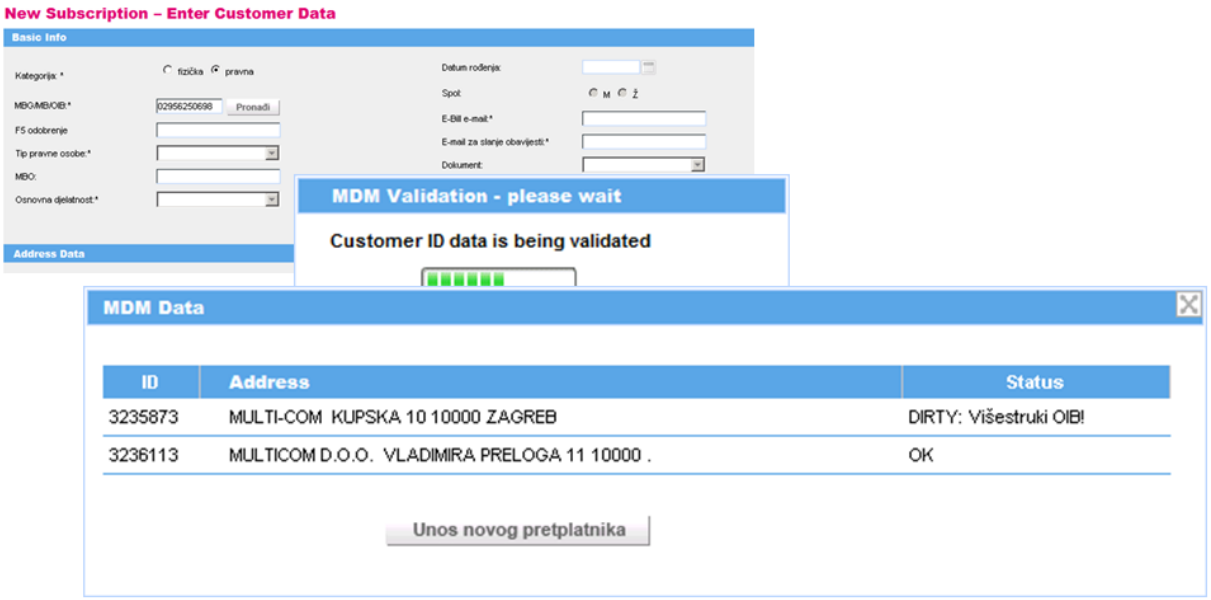
Elementi su rješenja centralni sustav MDM s pripadajućim integracijskim sučeljima, usvojeni jedinstveni model podataka, upravitelji podacima ovlašteni od strane organizacije (engl. *Data Steward*), DWH sustav koji vrši ulogu automatskog upravitelja podacima na način da šalje korektivne instrukcije preko zasebnog sučelja te jasno definirano vlasništvo i procesi upravljanja podacima [5].

II. CENTRALNI MDM SUSTAV

Sustav Multicom MDM implementiran je u skladu s provjerenim (engl. *best-practice*) tržišnim smjernicama i osigurava jedinstvenost (npr. isti se kupac unosi samo jednom), kvalitetu te dostupnost podataka svim povezanim sustavima.

Sustav MDM zauzima centralno mjesto u IT okolini. Servisi su izloženi organizaciji putem prilagodljivog sučelja (direktno ili preko platforme ESB (engl. *Enterprise Service Bus*) te odrađuje sljedeće funkcionalnosti za povezne informacijske sustave [2]:

- Konsolidacija podataka. Za vrijeme trajanja poslovnog događaja u izvornim sustavima upiti i transakcije šalju se MDM-u u realnom vremenu te se na taj način osigurava jedinstvena i potpuna informacija u trenutku nastanka promjene (Slika 1).
- Automatsko praćenje i detektiranje promjena nad podacima.
- Spajanje, automatsko ispravljanje i deduplikacija podataka.
- Sinkronizacija i propagacija promjena nad podacima u povezane sustave.
- Servise za pretragu i dohvat *master* zapisa (engl. *Single-Source-Of-Truth*) kao osnovu za aplikativnu nadogradnju sustava:



Slika 1. Interakcija izvornog sustava s MDM-om u realnom vremenu

- *SingleCustomerView* – uvid u podatke i usluge kupaca.
- *OneBill* – izdavanje objedinjenog računa za sve usluge pojedinog kupca

A. *Funkcije sustava*

MDM usklađuje višestruke zapise o istom entitetu i kontrolira kvalitetu tih zapisa. Entitet je kupac (engl. *customer*), adresa, korisnički račun (engl. *account*), itd. Višestruki zapisi nalaze se u različitim bazama i/ili u jednoj bazi.

Sustav je baziran na komunikaciji porukama. Izvorni sustav šalje poruku MDM-u prigodom izmjene izvornog zapisa. MDM šalje izvornim sustavima propagacijske poruke za izmjenu izvornih zapisa.

Obrada svake dolazne poruke može se razložiti na nekoliko uzastopnih koraka: odbacivanje nevažnih zapisa, kontrola kvalitete, razrješenje entiteta, integracija podataka, propagacija.

1) *Odbacivanje nevažnih zapisa*

U prvom koraku odbacuju se zapisi u kojima nema znatnih promjena u odnosu na prethodnu verziju istog zapisa u dolaznim porukama. Važne su promjene one kojima se mijenjaju vrijednosti atributa koji su označeni kao važni (što je moguće podesiti), a bilježe se u modelu podataka SCD2 (engl. *Slowly Changing Dimension Type 2*) [3]. Time se smanjuje broj zapisa koji se kasnije moraju obrađivati.

2) *Kontrola kvalitete*

U drugom koraku vrši se kontrola kvalitete: za svaku dolaznu poruku evaluira se skup uvjeta koje poruka mora zadovoljavati kako bi prešla u sljedeću fazu obrade. Moguće je podesiti da se ispunjenjem nekog uvjeta poruka trajno ili privremeno zadrži u „karanteni“. Zadržavanje poruke u karanteni znači obustavljanje obrade poruke dok je upravitelj podataka ručno odobri ili odbije.

Zapisi iz poruka koje prođu kontrolu kvalitete spremaju se u MDM skup tablica (engl. *upsert*) i koriste se u budućnosti za integraciju podataka *master* zapisa i propagaciju.

Moguća su sljedeća podešavanja:

- Podešavanje uvjeta kojima se inicira kontrola kvalitete (ukoliko poruka ne zadovoljava ni jedan od podešenih uvjeta, ona se automatski prihvaća). Sustav podržava unos uvjeta u *PL/SQL* kompatibilnoj sintaksi.
- Za svaki uvjet koji je inicirao kontrolu kvalitete može se podesiti jedna od sljedećih aktivnosti: prihvaćanje, odbacivanje ili zadržavanje poruke u karanteni te odgođeno prihvaćanje ili odgođeno odbacivanje poruke, kao što je prikazano na slici 2, pri čemu se za vrijeme odgode poruka drži u karanteni kako bi je mogli razriješiti upravitelji podatcima.

3) *Razrješenje entiteta (eng.* entity resolution*)*

U trećem koraku (razrješenja entiteta [4]) vrši se povezivanje novih dolaznih zapisa o kupcima s postojećim *master* zapisima na način da se traži *master* zapis koji predstavlja istog fizičkog kupca. Taj se korak preskače za poruke koje se odnose na postojeći izvorni zapis kupca koji je nekom prošlom porukom došao u sustav i za njega se već izvršio korak povezivanja. Povezivanje se odvija u dvije faze: faza blokiranja i faza bodovanja [5]. U fazi blokiranja izvršava se SQL upit kojim se za zadani izvorni zapis dobiva skup (takozvani blok) potencijalnih *master* zapisa, kandidata za povezivanje. U fazi se bodovanja za svaki *master* iz bloka evaluira jedan ili više uvjeta koji pridonose ukupnom rezultatu pozitivnom (bodovna vrijednost ukoliko je uvjet ispunjen) ili negativnom (vrijednost ukoliko uvjet nije ispunjen) težinom [5].

Event (change)	Reason code (condition)	DW Validation, Resolving & Propagation		
		CRM	Billing	ERP
Company name	Legal document proof	Auto	Auto	Auto
Company name	Bankruptcy	Auto	Auto	Manual
Name	Typing error	Auto	Auto	Auto
Last name	Unknown (residential user)	Manual (accept after 48h)	Manual (Accept after 48h)	Auto
Billing address change	Unknown	Manual	Manual	Manual
Billing address change	Signed and verified request	Auto	Auto	Auto

Slika 2. Definiranje pravila kontrole kvalitete

Funkcionalnost razrješenja entiteta koristi se i u funkciji pretraživanja *master* zapisa pa postoji mogućnost podešavanja više različitih pravila povezivanja, iako se u danom trenutku za povezivanje novih dolaznih zapisa uvijek koristi samo jedno podešeno pravilo. Mogućnosti su podešavanja pravila povezivanja (engl. *Matching rules*) sljedeće:

- Moguće je odrediti SQL uvjet spajanja koji se koristi u SQL upitu za pronalaženje bloka.
- PL/SQL uvjeti kojima se boduju *masteri* mogu se podešavati; za svaki se uvjet mogu podesiti dvije različite težine (pozitivna i negativna) koje se pribrajaju ukupnom broju bodova u slučaju da je uvjet ispunjen ili u slučaju da uvjet nije ispunjen.
- Za svako se pravilo podešavaju dva praga:
  - negativni prag koji označava razinu ispod koje se *master* zapis ignorira s dovoljno velikom vjerojatnosti da ne pripada istom entitetu kao zadani izvorni zapis;
  - pozitivni prag koji označava zbroj koji odgovara dovoljno velikoj vjerojatnosti da *master* zapis predstavlja isti entitet kao zadani izvorni zapis.

Mogući ishodi evaluacije skupa pravila povezivanja (npr. na osnovi tipičnih atributa kao što su OIB, naziv, tip kupca) jesu:

- **Automatsko povezivanje** – ako postoji točno jedan *master* zapis s zbrojem iznad pozitivnog praga, onda se izvorni zapis automatski povezuje s njim.
- **Konflikt** - ako postoji više *master* zapisa sa zbrojem iznad pozitivnog praga (više kandidata za automatski povezivanje), onda se izvorni zapis stavlja na popis za ručno razrješenje upravitelju podataka.
- **Novi *master* zapis** - ako nije pronađen ni jedan master zapis iznad negativnog praga, onda se automatski stvara novi (prazni) *master* s kojim se povezuje izvorni zapis.
- **Potencijalno povezivanje** – zbroj bodova između negativnog i pozitivnog praga. Zapisi se stavljaju na popis za ručno razrješenje.

Dok je izvorni zapis na popisu za ručno razrješenje, on još uvijek nije povezan ni s jednim *masterom* i ne vidi se u rezultatima pretraživanja.

4) Izgradnja master *zapisa (eng.* data integration*)*

U koraku integracije vrši se izgradnja ili ponovna izgradnja *master* zapisa pomoću svih vrijednosti izvornih zapisa. Za svaki atribut *master* zapisa moguće je podesiti integracijski operator (funkciju) kojim se dobiva vrijednost atributa *master* zapisa [4]. Izgradnja *master* zapisa vrši se evaluacijom integracijskih operatora za svaku kolonu *master* zapisa. Integracijski operatori imaju određen skup parametara u ovisnosti o tipu operatora. Slijedi opis implementiranih tipova operatora i njihovih parametara:

- ***Best value*** operator: uzima najbolju iz rang-liste poznatih vrijednosti; koristi se npr. za atribut *market segment*, pri čemu se za vrijednost *master* atributa uzima najbolji *market segment* pronađen u izvornim zapisima;
- ***Last update wins*** operator: svi izvorni zapisi poredaju se po zadanom kriteriju (primjer: kolona s datumom stvaranja zapisa) te se uzima vrijednost iz zadnjeg zapisa;
- ***Min*** operator: uzima najmanju pronađenu vrijednost zadanog atributa među izvornim zapisima; primjer: datum aktivacije korisnika;
- ***Max*** operator: uzima najveću pronađenu vrijednost zadanog atributa među izvornim zapisima.

Svi navedeni integracijski operatori imaju još nekoliko zajedničkih parametara: naziv kolone iz izvornog zapisa iz koje se uzimaju vrijednosti te popis uvjeta prioriteta koje izvorni zapisi moraju zadovoljiti kako bi uopće bili uzeti u obzir kao ulaz za zadani integracijski operator (ukoliko postoji bar jedan izvorni zapis koji zadovoljava prvi uvjet prioriteta, onda se kao ulaz u operator uzimaju samo zapisi koji zadovoljavaju taj uvjet; ako ne postoji zapis koji zadovoljava prvi uvjet, onda se na isti način traži zapis koji zadovoljava drugi uvjet prioriteta, itd). Tipični su uvjeti prioriteta koji se koriste: prioritet izvornih sustava, status kupca (aktivni su jači od suspendiranih), status zapisa (sumnjivi zapisi imaju manji prioritet) itd.

5) Propagacija promjena

Zadnji je korak obrade poruke propagacija promjena izvornim sustavima. Vrši se usporedba *master* zapisa i povezanih izvornih zapisa na način da se uspoređuju atributi kojima se inicira propagacija; za izvorne zapise kojima se razlikuje bar jedan takav atribut u odnosu na *master* zapis obavlja se slanje kompletnog *master* zapisa izvornom sustavu. Moguće je podešavanje atributa koji uzrokuju propagaciju.

6) Funkcije za pretraživanje i dohvat podataka

Postoje funkcije za dohvat podataka iz MDM-a po primarnim ključevima, te funkcija za pretraživanje MDM kupaca. Funkcija za pretraživanje implementirana je tako da koristi funkcionalnost razrješenja entiteta kako bi se pronašao zapis (koristi se posebno podešeno pravilo povezivanja, kako je već opisano).

7) Ostale funkcije

Osim navedenih funkcionalnosti postoji i nekoliko funkcija sustava MDM za ručno spajanje više *master* zapisa u jedan i odvajanje izvornog zapisa od *mastera*. Te su funkcije namijenjene upraviteljima podataka (ili

procedurama za automatsko čišćenje podataka, npr. DWH) koji popravljaju neispravno povezane zapise.

Pri ručnom spajanju dvaju *mastera* jedan se od njih označava kao obrisan, a izvorni zapisi koji su prije bili povezani sa tim *master* zapisom sada se povezuju s novim *master* zapisom. Prilikom odvajanja pojedinačnog izvornog zapisa od *mastera* moguće je kreirati novog *mastera* ili povezati izvorni zapis s nekim drugim, postojećim *masterom*.

B. *Model podataka o korisnicima*

Apstraktni model podataka o kupcu, prikazan na slici 3, preslikava se u 4 skupa tablica: ICR (*incoming change record*), INP (*input record*), MDM (tekući MDM podaci), OCR (*outgoing change record*), te dodatni skup tablica sa povijesnim podacima (HST).

MDM skup tablica glavni je skup tablica u kojem se drže podaci o kupcima, što uključuje *master* zapise i izvorne zapise, odnosno samo izvorne zapise za tablice koje nemaju odgovarajuću *master* tablicu (npr. tablica korisničkih računa (*account*), budući da koncept *master accounta* nije smislen).

ICR skup tablica drži izvorne podatke o kupcima iz dolaznih poruka (poruke sadrže dio podataka iz modela kupca); INP skup tablica drži važne izvorne zapise koji su kopirani iz odgovarajućih ICR tablica (nevažni zapisi se odbace na način da se ne izvrši kopiranje u ICR tablice); OCR skup tablica sadrži podatke o kupcima koji se propagiraju izvornim sustavima; u HST tablicama prati se povijest svih zapisa iz MDM skupa, a služi za povrat podataka (engl. *point-in-time recovery*).

Sustav podržava neograničeni broj relacija (poveznica) među entitetima. S relacijama su prirodno povezani entiteti kupca i adrese ili računa i njegove dostavne adrese, za povezivanje kupca s osobama za kontakt (koje mogu biti neovisni MDM kupci), ali je pomoću istog mehanizma moguće kupce dodatno povezivati u hijerarhijske grupe kao što su: udruge, kućanstva, holdinzi, grupe na Facebooku i sl. te na taj način dobiti ispravnu sliku o važnosti svakog kupca i osigurati ispravan pristup.

Svaka je vrsta relacije definirana kao poveznica između entiteta vrste A i vrste B.

C. *Uloga sustava skladištenja podataka*

Multicom MDM sustav ima izloženo i skupno (*batch*) sučelje prema sustavu DWH za čišćenje podataka i poboljšanje kvalitete podataka pomoću atributa kalkuliranih od strane sustava DWH.

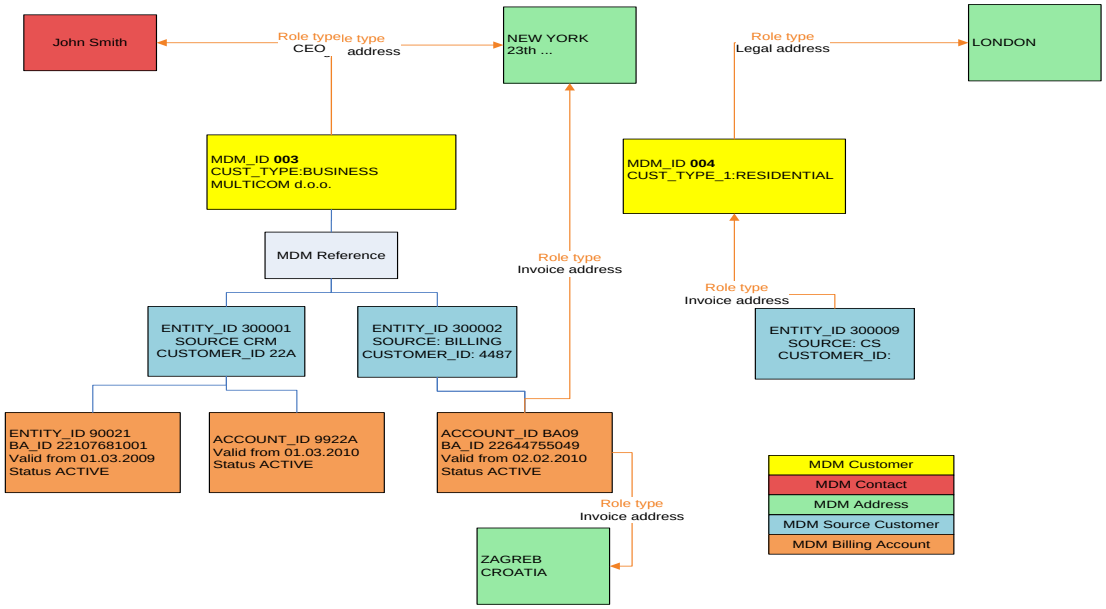
Sustav DWH ima ulogu automatskog upravitelja podataka i na taj način povećava integritet i kvalitetu informacija u MDM-u. Operacije kojima se to postiže jesu:

- Periodički prolaz kroz sve MDM zapise radi provjere integriteta podataka i poveznica korištenjem vlastitih *data quality* (DQ) alata;
- Generiranje poruka za korekciju MDM podataka u *batch* načinu.

Specifična uloga sustava DWH je da primjenom isprobanih DQ pravila, koja po performansama nisu pogodna za izvođenje u realnom vremenu (npr. napredna usporedba naziva – algoritmi permutacije slova, baza sličnih imena, razlika više slova u nazivima i sl.), ispravlja MDM zapise slanjem instrukcija preko zasebnog sučelja.

Sustav DWH ima mogućnosti spajanja i razdvajanja MDM kupaca (*SplitCustomer*, *MergeCustomer* funkcije), a one se koriste ukoliko se kontrolom uoči greška u primjeni MDM pravila povezivanja.

Osim toga sustav DWH se koristi i prilikom izračuna vrijednosti (npr. segmentacija korisnika) koje se ne unose u sustav MDM putem redovnih operativnih transakcija te za popunjavanje zapisa kupaca koje nedostaju u sustavu MDM (u situacijama kada promjene nastaju u sustavima koji nisu povezani s MDM-om).



Slika 3. Primjer definicije podatkovnog modela

DWH sustav, koji ionako ima ulogu obogaćivanja i čišćenja podataka, opisanim mehanizmima može izravno utjecati na operativne izvorne sustave jer se inicirane promjene kroz MDM infrastrukturu propagiraju prema svim povezanim sustavima.

D. *Enterprise rješenje*

Opisano rješenje je implementirano s ciljem da podržava vrlo veliki broj transakcije u realnom vremenu, i isprobano je u produkciji s velikim skupom podataka. Primjer mjerenja performansi, na sustavu reda veličine 50-tak milijuna izvornih zapisa, za prosječno trajanje transakcija u periodu najvećeg prometnog intenziteta (mjerenja su provedena u periodu od 13:00-14:00 sati tijekom jednog radnog tjedna; više od 20.000 transakcija u satu) je prikazan na slici 4.

Detalji o serveru i IT okruženju na kojem su mjerenja provedena, mogu se dobiti na upit.

Stabilnost rješenja je potvrđena u neprekidnom četveromjesečnom produkcijskom radu bez ijednog ispada.

Sigurnosne komponente sustava putem korisničkih aplikativnih uloga te korištenjem mehanizma Oracle VPD osiguravaju da pojedini korisnici mogu vidjeti i ispravljati samo pridijeljene segmente podataka (konfigurabilan pristup podacima na osnovi vrijednosti određenih korisničkih atributa).

Komponente *audit* i *loging* sustava osiguravaju da se svaka promjena podataka može identificirati bilo da je uzrokovana transakcijom iz povezanih sustava, bilo da je odrađena preko MDM korisničkog sučelja.

	DURATION (ms)				
FUNCTION	min	max	percentile_ disc_98	median	avg
create_customer	8	8073	165	39	59
create_billing_account	9	277	70	13	25
create_service_account	9	130	32	12	14
create_address	9	16912	181	36	64
create_contact	9	754	152	45	56
create_relation	13	24	24	14	16
get_customer_tree	1	1117	47	32	21
search_customer	4	180	35	6	12

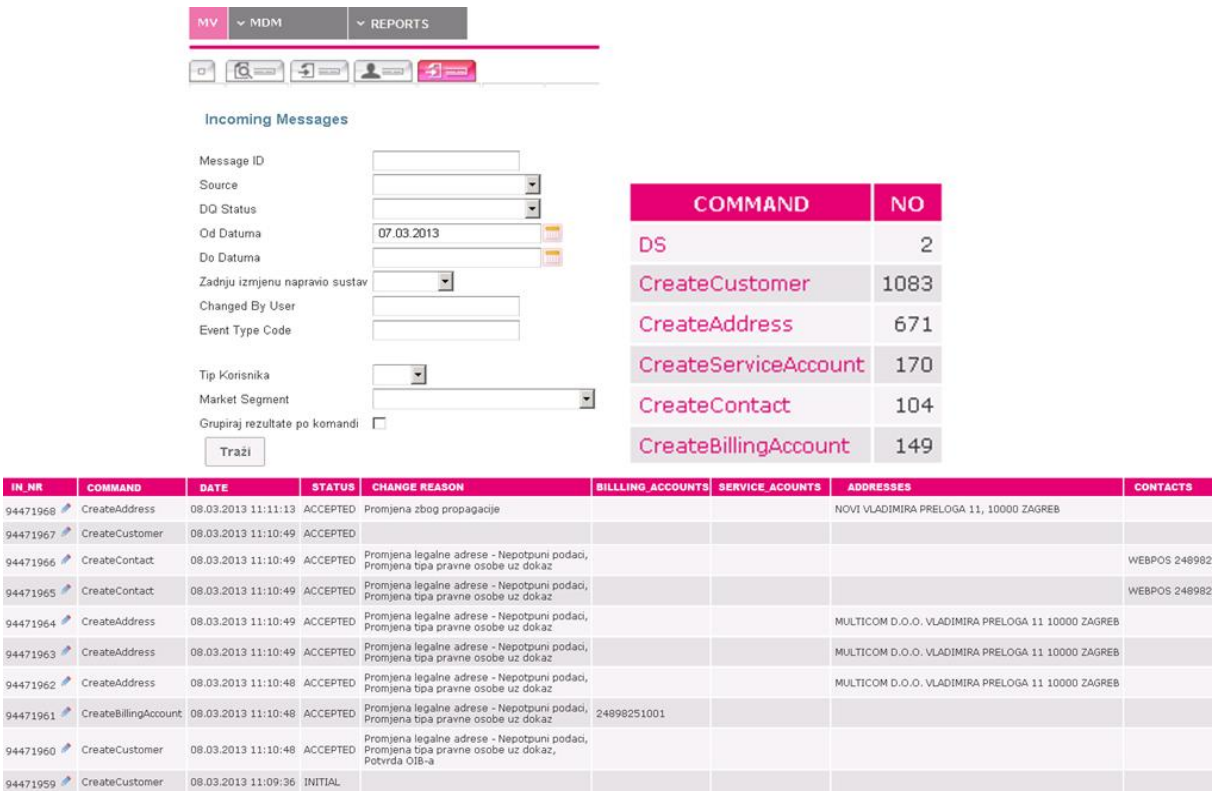
Slika 4. Statistika trajanja transakcija *(peek hour)*

Implementirani mehanizam *Point-in-time recovery* osigurava povrat povijesnih zapisa preko kronološkog popisa promjena nad odabranim objektom iz podatkovnog modela.

III. KORISNIČKO SUČELJE

Za potrebe nadzora i čišćenja podataka razvijeno je intuitivno web korisničko sučelje (aplikacija), prikazano na slici 5, koje prvenstveno koriste dedicirani upravitelji podataka za nadzor i ručno razrješenje poruka u karanteni.

Aplikacija sadrži funkcionalnosti za pojedinačne i masovne korekcije podataka, kao i niz izvještaja za nadzor i provjeru ispravnosti rada sustava.



Slika 5. Pregled dolaznih poruka kroz aplikaciju

Primarna je uloga aplikacije da omogući pristup karanteni i da se ručno prihvaćanje ili odbijanje sumnjivih promjena vrši kroz aplikaciju, kao što je prikazano na slici 6, ali dodatno je osiguran niz funkcionalnosti u smislu osiguranja neometanog rada *data strewarda* na poslovima kontrole i korekcije podataka:

- Mogućnost vraćanja (engl. *point-in-time recovery*) podataka svakog entiteta u prijašnje stanje. napravljene promjene uz propagaciju starih vrijednosti u spojene sustave (i u sustav uzročnika promjene);
- Pregled svih dolaznih i odlaznih poruka sa statusima obrade u MDM-u i u povezanom sustavu (konfigurabilni skup filtara za suženje liste poruka prilikom pretrage);
- Direktno mijenjanje vrijednosti atributa nad MDM *master* podacima, ili nad podacima u spojenim sustavima (promjene se odmah propagiraju u povezane sustave);
- Spajanje dvaju ili više *master* entiteta u jedan;
- Prebacivanje pojedinih entiteta s jednog *mastera* na drugi prilikom uočene iznimke od definiranih pravila povezivanja;
- Pregled i nadzor korekcijskih instrukcija koje je poslao sustav DWH.

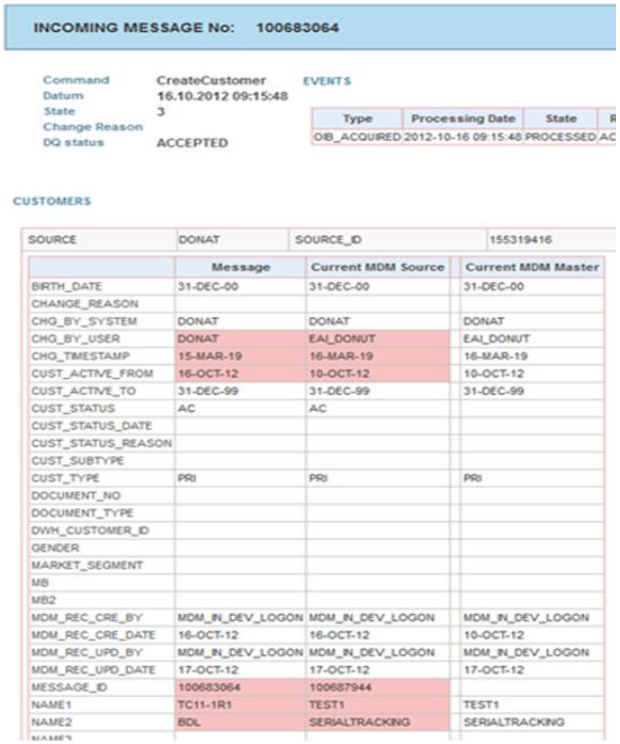
Aplikacija omogućava jedinstveni hijerarhijski pregled *master* zapisa pojedinog kupca s *drill-down* mehanizmom za pregled svih povezanih entiteta. Osiguran je generički dizajn i proširenje formi i izvještaja bez potrebe za gašenjem aplikacije.

Aplikacija je razvijena u standardnim tehnologijama Java (JSF, JSP) i JavaScript (jQuery) i kao takva je prilagođena za izvođenje na svim popularnim aplikacijskim poslužiteljima (WebLogic, Jboss, Tomcat, WebSphere, itd.) odnosno internetskim preglednicima (IE, Firefox, Chrome, Opera, itd.).

IV. ZAKLJUČAK

Sustav MDM i povezani procesi upravljanja podacima iznimno su važni u tvrtkama i institucijama koje pružaju veliki broj usluga i proizvoda različitim kanalima i organizacijskim jedinicama (holdinzi, javna uprava, banke, osiguravajuća društva), a koje žele ostvariti objedinjeni uvid u podatke o klijentima, partnerima i proizvodima s ciljem efikasnog imenovanja njihovih potreba (prodaja, usluge), povećanja zadovoljstva klijenata (brža i efikasnija usluga, kraće rješavanje problema i zahtjeva, manje žalbi itd.) te mogućnosti egzaktnog izračuna profitabilnosti istih (telekomunikacije, financije, javni sektor).

Multicom MDM rješenje objedinjava sve tražene uloge koje se očekuju od sustava *enterprise Master Data Management.*



INCOMING MESSAGE No: 100683064

Command	CreateCustomer
Datum	16.10.2012 09:15:48
State	3
Change Reason	
DQ status	ACCEPTED

EVENTS

Type	Processing Date	State
OIB_ACQUIRED	2012-10-16 09:15:48	PROCESSED

CUSTOMERS

SOURCE	DONAT	SOURCE_ID	155319416

	Message	Current MDM Source	Current MDM Master
BIRTH_DATE	31-DEC-00	31-DEC-00	31-DEC-00
CHANGE_REASON			
CHG_BY_SYSTEM	DONAT	DONAT	DONAT
CHG_BY_USER	DONAT	EAI_DONUT	EAI_DONUT
CHG_TIMESTAMP	15-MAR-19	16-MAR-19	16-MAR-19
CUST_ACTIVE_FROM	16-OCT-12	10-OCT-12	10-OCT-12
CUST_ACTIVE_TO	31-DEC-99	31-DEC-99	31-DEC-99
CUST_STATUS	AC	AC	
CUST_STATUS_DATE			
CUST_STATUS_REASON			
CUST_SUBTYPE			
CUST_TYPE	PRI	PRI	PRI
DOCUMENT_NO			
DOCUMENT_TYPE			
DWH_CUSTOMER_ID			
GENDER			
MARKET_SEGMENT			
MB			
MB2			
MDM_REC_CRE_BY	MDM_IN_DEV_LOGON	MDM_IN_DEV_LOGON	MDM_IN_DEV_LOGON
MDM_REC_CRE_DATE	16-OCT-12	16-OCT-12	10-OCT-12
MDM_REC_UPD_BY	MDM_IN_DEV_LOGON	MDM_IN_DEV_LOGON	MDM_IN_DEV_LOGON
MDM_REC_UPD_DATE	17-OCT-12	17-OCT-12	17-OCT-12
MESSAGE_ID	100683064	100687944	
NAME1	TC11-1R1	TEST1	TEST1
NAME2	BDL	SERIALTRACKING	SERIALTRACKING

Slika 6. Analiza promjene s označenim razlikama atributa

Pozicionira se na području sličnih sustava koji osiguravaju kvalitetu podatka kroz niz bitnih prednosti kao što su: visok stupanj fleksibilnosti prilikom definicije podatkovnog modela, brojne opcije integracije s izvornim IT sustavima, prilagodljivost tržištu u regiji (gramatika, palatali, računovodstvena pravila i sl.), vrhunske performanse izvođenja u transakcijski intenzivnom okruženju (izvođenje transakcija u memoriji, vlastiti mehanizam za obradu događaja), jednostavno uklapanje u korporativne procesne mehanizme korištenjem prilagodljivih pravila za kvalitetu podataka i integraciju te modeliranjem poslovnih procesa (eng. *workflow engine*).

LITERATURA

[1] Roger Wolter and Kirk Haselden, The What, Why, and How of Master Data Management, 2006 *h*ttp://msdn.microsoft.com/en-us/library/bb190163.aspx

[2] Marty Moseley: Integrating Master Data Management into the Data Warehouse, 2013, http://content.arma.org/imm/FeaturesWebExclusives/FeatureWebExclusiveIntegratingMasterData.aspx

[3] R. Kimball and M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. New York, NY: John Wiley & Sons, 2002.

[4] John R. Talburt, Entity Resolution and Information Quality. Burlington, MA: Morgan Kaufmann, 2010.

[5] I. Fellegi and A. Sunter, A theory for record linkage. Journal of the American Statistical Association, vol. 64, pp.1183-1210, December 1969.

[6] Gartner, October 2012, Magic Quadrant for Master Data Management of Customer Data Solutions

Suvremena rješenja poslovne inteligencije zasnovana na konceptu računarstva u oblaku

Ivana Sekula, Mladen Franić
Hrvatski Telekom d.d./Sektor za informacijsko-komunikacijska rješenja
za poslovne korisnike (ICT rješenja), Zagreb, Hrvatska
Hrvatski Telekom d.d./Odjel za korporativnu sigurnost, Zagreb, Hrvatska
Ivana.Sekula@t.ht.hr, Mladen.Franic@t.ht.hr

Sažetak - Računarstvo u oblaku (engl. cloud computing) predstavlja novi model poslovanja koji uključuje čitav niz koncepata kao što su virtualizacija, dizajn distribuiranih aplikacija i poduzetno IT-upravljanje koji omogućuju fleksibilniji pristup razmještanju i skaliranju aplikacija posredstvom internetski povezanih poslužitelja. Uz pomoć web preglednika i posebnih klijenata, ove su usluge prilagodljive, a korisnici plaćaju samo ono što koriste. Računarstvo u oblaku predstavlja budućnost poslovne produktivnosti. U ovome će se radu razmotriti kako poslovna inteligencija (Business Intelligence) može utjecati na poslovnu produktivnost u kontekstu primjene rješenja iz domene računarstva u oblaku. Primjena poslovne inteligencije izdašan je financijski trošak s obzirom na veliku količinu podataka koju je potrebno analizirati i obraditi kako bi se utvrdili trendovi u poslovanju te značajne procesorske resurse koje je potrebno osigurati s tim u vezi. Cilj ovog rada je pokazati kako se poslovna inteligencija može ponuditi po modernom skalabilnom SaaS (Software as a Service) konceptu, koji se u posljednje vrijeme nametnuo kao jednostavno i održivo rješenje za organizacije opterećene kompleksnom softverskom infrastrukturom i istovremenim ograničenim financijskim sredstvima. Ovaj pristup u poslovnom okružju može omogućiti značajnu uštedu troškova zadržavajući pritom rješenja iz domene poslovne inteligencije te pristup aktualnim aplikativnim rješenjima s pozitivnom reperkusijom na konkurentnost. Ovaj rad ima za cilj ukazati na novu kvalitetu koju uporaba SaaS koncepta u poslovnoj inteligenciji donosi u poslovno okružje, kako takva rješenja izgledaju te koji su izazovi koje njihova implementacija donosi sa sobom.

I. UVOD

U današnjem poslovnom okruženju s obzirom na tržišne i ekonomske uvjete poslovanja, pravovremena i cjelovita informacija od ključnog je značaja za održavanje profitabilnosti i konkurentnosti. Imajući pravi uvid u zbivanja u vlastitom poslovnom okruženju, praćenjem i analizom tržišnih trendova postiže se mogućnost poduzimanja neophodnih mjera za brzo identificiranje poslovnih mogućnosti s ciljem postizanja maksimalne učinkovitosti. Kako bismo bili sigurni da poslovni procesi raspolažu znanjem o svojem poslovnom okruženju, u svrhu potpore strateškog planiranja nameće se ideja korištenja poslovne inteligencije (Business Intelligence) za produkciju i obradu informacija o poslovnom okruženju.

Za bolji uvid u poslovanje, a s ciljem provođenja odluka i akcija, na raspoložive podatke važno je primijeniti metode podatkovne analize, što omogućuje biranje ispravne strategije za postizanje boljih ciljnih rezultata (prodajnih, marketinških itd.), identifikaciju mogućih poteškoća i postizanje veće dinamičnosti poslovnih procesa što s druge strane može zadovoljiti današnje izazove po pitanju pravovremenog djelovanja.

Obrada velike količine podataka izoliranih iz operativnih baza i spremljenih u posebne baze odnosno skladišta podataka (Data Warehouse) zahtijeva značajnu procesorsku snagu i prostor za pohranu. Navedeni resursi ključni su za obnašanje poslovnih aktivnosti na dnevnoj bazi, a posebne korporativne organizacijske jedinice zadužene za održavanje informacijsko-komunikacijske infrastrukture odnosno tehnologije (IT), nemaju mogućnost osigurati dodatnu procesorsku snagu za navedenu namjenu izvan vremena redovnog rada. U tom slučaju koncept računarstva u oblaku (cloud computing) može se uspješno primijeniti kao skalabilno rješenje u ključnim periodima kada postoji potreba za obradom podataka, a s tim u vezi trošak infrastrukture svodi se na intervale korištenja iste.

Istraživanje sažeto u ovome radu ima za cilj razjasniti kako bi se koncept računarstva u oblaku mogao iskoristiti za pružanje rješenja poslovne inteligencije zasnovanih na primjeni suvremenog SaaS (Software as a Service) koncepta te će biti ponuđeni odgovori na sljedeća pitanja:

- Kako izgleda i kako funkcionira tipično rješenje poslovne inteligencije zasnovano na primjeni SaaS koncepta;
- Koji isporučitelji rješenja poslovne inteligencije pružaju komercijalno dostupnu uslugu putem Interneta;
- Koje su prednosti primjene rješenja poslovne inteligencije zasnovanog na SaaS konceptu;
- Koji su izazovi u primjeni rješenja poslovne inteligencije zasnovanog na SaaS konceptu.

II. POSTIZANJE PRAVOG UVIDA U STANJE POSLOVANJA

Primjena poslovne inteligencije u svrhu poboljšanja poslovnih procesa donosi sa sobom složena rješenja koja

objedinjuju analizu i kombiniranu primjenu izvorišnih podataka izvlačeći pritom korisne informacije iz tih podataka i pružajući ključni uvid u utvrđene činjenice, a u svrhu informiranja i provođenja inteligentnih poslovnih odluka. Ova rješenja uključuju skladištenje podataka, dubinsku analizu podataka (data mining) i sustave za analizu trendova [1].

Dohodovna strana i učinak poslovanja poboljšavaju se primjenom poslovne inteligencije u svrhu utvrđivanja ključnih podataka na temelju kojih bi se donosile odluke i provodile akcije. Poslovna inteligencija objedinjuje proizvode, tehnologiju i metode koji pomažu u poboljšanju poslovne učinkovitosti a što povećava prihode i/ili smanjuje rashode [2].

Jedan od najvećih nedostataka s kojim se korporacije danas susreću svodi se na problem učinkovitog upravljanja informacijama. S tim u vezi, u većim je tvrtkama značajan broj visoko pozicioniranih menadžera koji smatraju da su im potrebne učinkovitije mogućnosti upravljanja informacijama te da te mogućnosti trebaju napredovati. Implementacijom rješenja poslovne inteligencije, poslovanje dobiva mogućnost pristupa novim podacima u svrhu boljeg informiranja pri donošenju strateških odluka te se ovim pristupom ostvaruju značajne uštede u troškovima poslovanja budući da se organizacijski resursi učinkovitije dodjeljuju [3].

Da bi se vidjelo u kojoj je mjeri poslovna inteligencija korisna u primjeni u poslovanju, razmotrit će se primjer iz kojega je vidljivo kako poslovna inteligencija može unaprijediti upravljanje logistikom u poslovanju koje je suočeno s povećanjem složenosti u upravljanju poslovima logistike. To povećanje složenosti u upravljanju logistikom dovelo je mnoge korporacije do izdvajanja logistike iz temeljnih djelatnosti (outsourcing) te oslanjanja na rješenja trećih strana kao davatelja usluga upravljanja logistikom.

Za učinkovito upravljanje opskrbnim lancima svojih klijenata, treće strane koriste alate za poslovnu inteligenciju, poput podatkovnog skladištenja i OLAP (Online Analytical Processing) alata kako bi se vršila konstantna analiza podataka prikupljenih iz različitih izvora i pretvorba istih u informaciju temeljem koje je moguće provoditi akcije. Pružajući unificirani pogled na opskrbni lanac, ti alati pomažu trećim stranama da poboljšaju usluge poput upravljanja transportom, skladištenjem i upravljanja imovinom i time klijentima pružaju informacije specifične za njihov opskrbni lanac povećavajući tako njihovu tržišnu konkurentnost [4].

Korištenjem poslovne inteligencije, korporacije imaju mogućnost zadržati profitabilne klijente i privući one druge, njima slične, kreirajući marketinške programe koji odražavaju individualnost klijenata i njihove preferencije. Primjena poslovne inteligencije omogućuje korporacijama nadalje stjecanje uvida u promjene ponašanja klijenata što može izazvati određene akcije, a to s druge strane korporacijama daje mogućnost povećanja prodaje po klijentu, predviđanjem koji bi od klijenata mogao biti najvjerojatniji kupac određenog proizvoda ili usluge. Ovo se postiže podešavanjem cjenovnih strategija kako bi se efikasno nadmašila konkurencija i maksimizirala dobit, uz primjenu proaktivnog pristupa u nadgledanju, utvrđivanju i predviđanju potreba klijenata čime se postiže prednost u odnosu na pasivni pristup čekanja da se kod klijenta dogodi neka nepovoljna situacija. Poslovna inteligencija omogućuje preraspodjelu ograničenih resursa (financijskih, ljudskih i tehničkih) na optimalan način [5].

Prilikom korištenja poslovne inteligencije, korporacije se susreću sa sljedećim izazovima [6]:

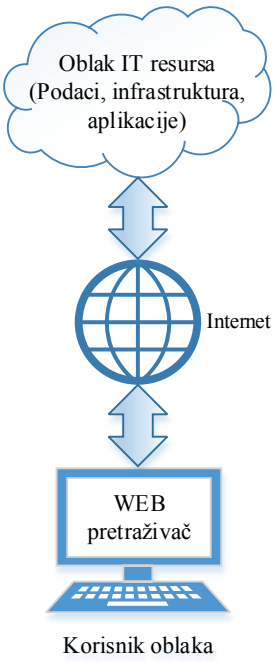
- Za korporacije postoji nužnost smanjivanja troškova za kapitalna ulaganja i operativnih troškova.
- Nedostatak vještina za učinkovito stvaranje i održavanje rješenja iz domene poslovne analitike, koja poslovanju nadodaju novu vrijednost.
- Današnje poslovne aktivnosti generiraju veliku količinu podataka iz različitih izvora, uključujući sustave za planiranje resursa (ERP-Enterprise resource planning), rješenja za upravljanje odnosima s klijentima (CRM-Customer relationship management) itd.
- Brzina generiranja podataka vrlo je visoka, zbog čega podaci pristižu na dnevnoj i satnoj bazi te u stvarnom vremenu što iziskuje snažnu procesorsku snagu opreme kako bi su mogli identificirati određeni trendovi. Iz navedenog razloga obrada podataka ponekad se može izvršiti izvan redovnog radnog vremena kako ne bi bilo utjecaja na normalno odvijanje poslovnih aktivnosti tijekom radnog vremena.
- U svrhu analize prodajnih rezultata, raspoloživosti opskrbnih lanaca, analitičke obrade klijentskih zahtjeva, analize proizvoda itd., poslovne aktivnosti zahtijevaju rješenja na zahtjev.
- U cilju kontrole troškova infrastrukture, osigurati odgovarajuću infrastrukturu u vrijeme kada se očekuje vršno opterećenje iste može biti dosta zahtjevan zadatak.
- Visoki troškovi implementacije i održavanja rješenja poslovne analitike.

Poslovna inteligencija nudi tvrtkama učinkovito rješenje za analizu poslovnih podataka te omogućuje da se donose poslovne odluke na osnovu prikupljenih znanja u cilju povećanja prihoda i učinkovitosti poslovnih procesa. S druge strane, implementacija i održavanje rješenja poslovne inteligencije skupo je samo po sebi te zahtijeva moćnu infrastrukturu koja je ponekad dostupna samo izvan redovnog radnog vremena.

Ekonomska kriza posljednjih godina izvršila je značajan utjecaj na poslovne odluke u smjeru smanjenja troškova poslovanja iako je tendencija i nadalje zadržati potrebne resurse i tehnologiju čime bi se pravovremeno mogle donositi ispravne poslovne odluke na osnovu prethodno identificiranih prilika i proaktivnog pristupa upravljanju rizicima u poslovanju. To je prostor u kojemu koncept računarstva u oblaku (cloud computing) može tvrtkama ponuditi učinkovito rješenje za smanjenje troškova ostavljajući pritom dostup do najboljih rješenja za svaku pojedinu namjenu.

III. KORIŠTENJE KONCEPTA RAČUNARSTVA U OBLAKU ZA POSTIZANJE POSLOVNE UČINKOVITOSTI

Primjena koncepta računarstva u oblaku daje korporacijama mogućnost korištenja informacijskih tehnologija posredstvom Interneta što je dinamički skalabilan pristup u odnosu na treće strane kao davatelje IT usluga (slika 1.). Računarstvo u oblaku potaknuto je ekonomijom razmjera (economy of scale) s glavnim obilježjem postojanja tzv. pool-a virtualizirane i dinamički skalabilne procesorske snage, memorijskih resursa za pohranu podataka, platformi i servisa, a što je dostupno na zahtjev vanjskim klijentima korisnicima posredstvom Interneta. Navedeni pristup predstavlja novu paradigmu za korištenje informacijskih tehnologija u korporativnom okruženju s primjenom načela poslovnog modela s unaprijed plaćenom uslugom za nabavku, isporuku i podršku infrastrukturnih, platformskih, aplikacijskih i poslovno procesnih mogućnosti. U tom smislu u korporativnom se okruženju gubi potreba za investiranjem u specifična znanja i stručnost kao i u stjecanje vještina za upravljanjem informacijskim tehnologijama koje imaju svoju podršku u oblaku [7].



Slika 1. Računarstvo u oblaku

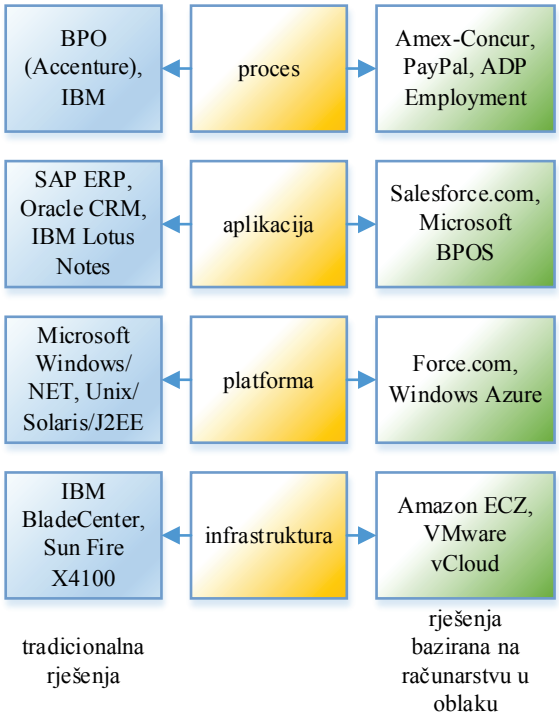
Računarstvo u oblaku može se prikazati modelom složaja od četiri sloja od kojih je tzv. poslovni oblak (Industry Cloud) na vrhu tog složaja (slika 2.).

Poslovni oblak je vanjski servis za potporu i davanje mogućnosti za upravljanje poslovnim procesima utemeljen na Internet povezivosti. Primjer takvih servisa je upravljanje zahtjevima, upravljanje troškovima ili nabavom te se procesi ovdje procjenjuju po transakciji [8].

Sljedeći sloj u složajnom modelu računarstva u oblaku je aplikacijski oblak (desktop i kolaboracija, SaaS). Aplikacijski oblak pruža usluge od nabave do ERP sustava i sustava za upravljanje sadržajem. Te se aplikacije izvršavaju na infrastrukturi trećih strana te su korporacije pretplaćene na te servise na osnovu broja korisnika istih [8].

Treći je sloj platformski oblak koji ima ulogu omogućiti dobavljačima aplikativnih rješenja brzi razvoj i implementaciju on-line aplikacija koristeći infrastrukturu trećih strana [8].

Posljednji sloj u složajnom modelu računarstva u oblaku je infrastrukturni oblak ili tzv. utility računarstvo (utility computing) koje je na mjestu gdje isporučitelj usluge računarstva u oblaku na zahtjev kompanijama-klijentima daje tzv. „sirovu“ infrastrukturu, poput procesorske snage, mrežne propusnosti i memorijskog kapaciteta za pohranu podataka [8].



Slika 2. Složajni model računarstva u oblaku

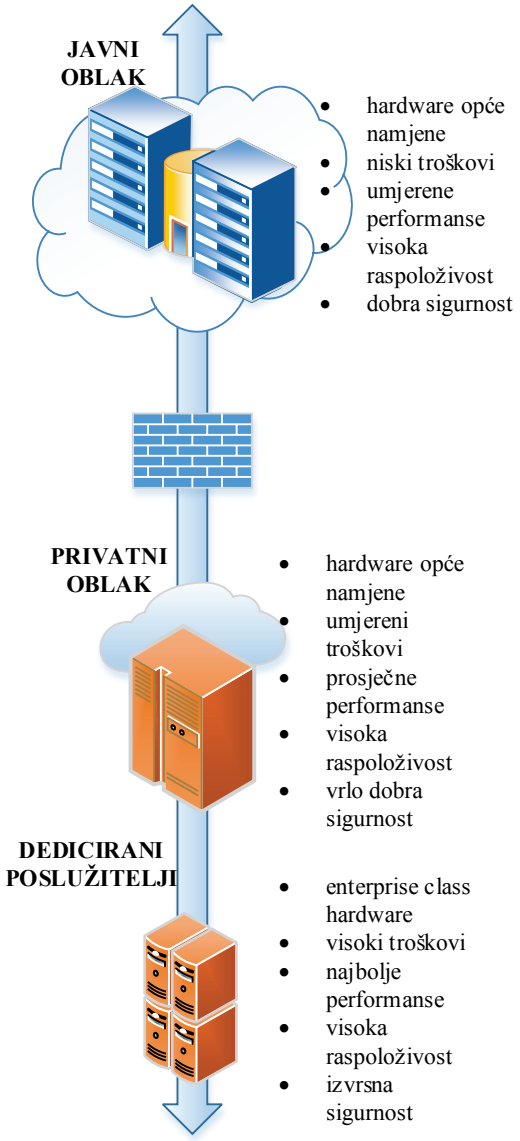
Računarstvo u oblaku je paradigma koja raznim aplikativnim rješenjima omogućuje smještaj bilo u privatni ili javni oblak. Programsko rješenje za aplikacije iz oblaka u načelu je koncipirano da uz minimalne promjene i podešavanja bude potpuno funkcionalno u radu. Davatelji usluge računarstva u oblaku pokrivaju sve potrebite zahtjeve za održavanjem normalne funkcionalnosti rada te vode računa o dostupnosti i raspoloživosti infrastrukture [9].

Aplikativna rješenja za poslovne potrebe mogu biti smještena ili korištenjem posebnog namjenskog, dediciranog poslužitelja ili u privatni ili javni oblak. Takav model hibridnog oblaka prikazan je na slici 3. Dedicirani poslužitelj je poslužitelj namijenjen pokretanju određenih softverskih aplikacija [10]. Svi resursi takvog

poslužitelja namijenjeni su isključivo za određenu softversku aplikaciju te se ne mogu dodjeljivati drugim aplikacijama ukoliko je za rad istih potrebno npr. više procesorske snage.

Kod migracije poslovnih procesa u privatni oblak postiže se fleksibilnost u poslovanju korištenjem vlastite IT infrastrukture. Potražnjom za resursima može se upravljati dinamički kako bi se udovoljilo potrebama za procesorskom snagom. Prednost ovog pristupa je poboljšanoj sigurnosti te omogućuje tvrtkama da izvuku maksimum od postojeće računalne opreme i mrežne infrastrukture kreiranjem privatnog oblaka (npr. Ubuntu).

Javni oblak pruža servise globalno dostupne od svuda u svijetu i omogućuje poslovanju da proporcionalno potrebama dinamički smanji ili poveća uporabu resursa potrebnih za obavljanje poslova. Infrastruktura javnog oblaka u vlasništvu je poslovnog subjekta koji javno prodaje uslugu korištenja oblaka te on istom upravlja [11].



Slika 3. Hibridni oblak

Usluge i servisi temeljeni na oblaku dostupni su na svim razinama tehnološkog složaja. Tvrtke-klijenti biti će u mogućnosti razvijati zasebne IT modele temeljene isključivo na poslovnim potrebama umjesto na tehnološkim ograničenjima. Takav novi prilagodljivi okvir za informacijske tehnologije mogao bi pojednostaviti upravljanje troškovima, prilagodljivost i unijeti dodatnu agilnost u poslovanje. Računarstvo u oblaku biti će poticaj za nastanak novih inovativnih rješenja u oblaku i povećanje informatizacije sektora informacijskih tehnologija. Ta inovativna rješenja rezultirat će promjenama u poslovnim modelima i omogućit će veći stupanj poslovne učinkovitosti [12].

IV. POSLOVNA INTELIGENCIJA NA ZAHTJEV

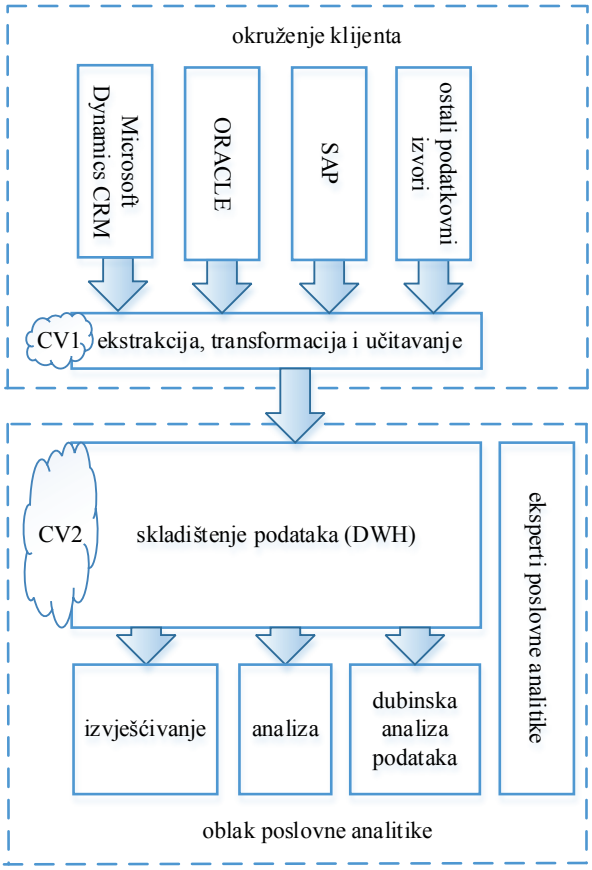
Korištenjem rješenja temeljenih na računarstvu u oblaku, tvrtke-klijenti u mogućnosti su riješiti poteškoće s kojima su bili suočeni kod primjene vlastitih rješenja poslovne inteligencije. Raspolaganje rješenjem poslovne inteligencije na zahtjev omogućit će korporacijama da postignu uštede i iskoriste rješenja na bazi pretplate na ista. To eliminira potrebu za dugotrajnim zahtjevima za investiranjem u potrebna rješenja te poslovanje dobiva trenutni dostup do sustava, omogućavajući tako odjelima za IT podršku poslovanju ubrzanje poslovnih procesa i postizanje nove kvalitete u poslovanju.

Poslovna inteligencija na zahtjev dostavlja se kao SaaS (Software as a Service) rješenje za poslovanje. Softversko rješenje dostavlja se posredstvom Weba na bazi godišnje ili mjesečne naknade za uslugu što ima prednost pred naručivanjem na bazi plaćanja unaprijed. Ovakvo SaaS rješenje ponuđeno od treće strane posredstvom Interneta može pružati veću operativnu i financijsku fleksibilnost i istovremeno smanjiti održavanje i podršku tehnologije.

Slika 4. prikazuje kako se rješenje za poslovnu inteligenciju može definirati u okruženju računarstva u oblaku. Transakcijski podaci se na dnevnoj bazi ekstrahiraju iz klijentskog okruženja te se potom transformiraju i učitavaju u poslovne kocke u skladištu podataka.

Alati za skladištenje podataka (DWH) i dubinsku analizu podataka (data mining) u okruženju oblaka koristit će se kako bi poslovanje opskrbili s informacijama i alatima potrebnim za donošenje operativnih i strateških poslovnih odluka. Skladište podataka pruža poslovanju mogućnost učenja ponašanja iz prošlosti i akcija, kako bi se razumjela prošlost poslovanja, odredila trenutna situacija i predvidjelo ili promijenilo moguće događaje u budućnosti.

Softver za poslovnu inteligenciju pruža tzv. „drill down“ značajku što omogućuje korisnicima da odabirom informacije u nekom izvješću dođu do relevantnih detalja o određenom dijelu podataka. Rješenje za poslovnu inteligenciju u oblaku osigurat će standardne izvještaje koji se mogu uporabiti odmah kada poslovanje započne koristiti rješenje u oblaku. Uobičajena poslovna izvješća mogu se objavljivati u centraliziranom, upravljanom i sigurnom okruženju, kojemu se može pristupiti globalno iz bilo kojeg dijela svijeta.



Slika 4. Poslovna inteligencija na zahtjev

OLAP engine u rješenju poslovne inteligencije u oblaku omogućuje brzo i jednostavno izvođenje ad-hoc upita i analize velikih količina složenih podataka u svim aspektima poslovanja. To omogućuje poslovanju da unaprijedi izvješćivanje za sve segmente - prodaju, marketing, upravljanje izvješćima, upravljanje poslovnim procesima, budžetiranje i predviđanje, financijsko izvješćivanje itd.

Važno obilježje poslovne inteligencije u oblaku je tzv. „dashbord“ odnosno kontrolna ploča koja pruža izvršni informacijski pogled na poslovanje s naglaskom na jednostavnosti korištenja. Kontrolna ploča objedinjuje sve informacije u jedinstveno i korisnički prijateljsko sučelje jednostavno za korištenje, a koje omogućuje rukovodećem osoblju odnosno managementu da proaktivno identificiraju problematična područja u poslovanju i poduzmu nužne mjere potrebne za ispravljanje negativnih trendova. Kontrolna ploča pruža vizualnu prezentaciju mjera učinkovitosti i za poslovanje daje fleksibilnost u smjeru donošenja odluka sa što većim stupnjem informiranosti na temelju prikupljenih podataka. Uporaba kontrolne ploče također štedi vrijeme budući da je moguće istovremeno izraditi višestruka izvješća i omogućiti rukovodećem osoblju da u kratkom vremenu dobiju potpunu preglednost u svim aspektima poslovanja.

Davatelj usluge računarstva u oblaku odgovoran je za održavanje ažurnosti softvera, a budući da je softver smješten izvan resursa klijenta, on ne koristi kritične resurse koji utječu na temeljnu djelatnost kompanije-klijenta. Procesorska snaga i resursi za pohranu podataka mogu se koristiti za potporu temeljnoj djelatnosti poslovanja, a ono što ovaj model čini još atraktivnijim jest činjenica da se algoritmi za podatkovnu i dubinsku analizu mogu izvršavati u bilo koje vrijeme dana. Rješenje računarstva u oblaku također je globalno dostupno iz bilo kojeg dijela svijeta.

V. UTJECAJ NA POSLOVANJE

Dobavljači SaaS rješenja poput američke tvrtke Panorama (http://www.panorama.com/) [13] koji su uveli analitiku i izvještavanje u tvrtku Google osmislili su SaaS rješenje poslovne inteligencije koje donosi sa sobom robusniji OLAP, analitiku i izvještavanje. Panorama također pruža određene prednosti za kompanije-klijente koje pojednostavljuju uporabu SaaS rješenja poslovne inteligencije na način da se usluga implementira samo jednom što olakšava rad odjelima za IT podršku s obzirom da nema potrebe za kontinuiranim ažuriranjem i održavanjem desktop aplikacija poslovne inteligencije.

Uz primjenu SaaS rješenja poslovne inteligencije postiže se veća fleksibilnost u dobivanju sustava poslovne inteligencije na principu korištenja po potrebi. To omogućuje poslovanju jednostavnu nadogradnju na noviju reviziju SaaS rješenja sa potrebnim funkcionalnostima jednom kada budu dostupni bolji tehnološki uvjeti [14]. Dobavljač SaaS rješenja odgovoran je za održavanje ažurnosti konfiguracijskih postavki i zadnje aktualne verzije softvera koji se koristi što oslobađa klijenta korisnika rješenja od navedenih poslova [15].

Drugi primjer isporučitelja SaaS rješenja poslovne inteligencije jest američka tvrtka RightScale (https://www.rightscale.com/). Klijentima nude solidna rješenja uz pouzdan rad i dobre performanse. Prednost ovih rješenja je u niskim troškovima bez žrtvovanja performansi ili funkcionalnosti te vidljivi rezultati odmah u početku primjene.

Računarstvo u oblaku omogućuje povećanje ili smanjenje aktivnosti ovisno o kapacitetu potrebnom za odvijanje poslovnih aktivnosti, a ta se karakteristika isporučitelju usluge plaća postupno [16]. Korištenjem SaaS rješenja poslovne inteligencije pomiče se teret financiranja poslovanja iz segmenta nadolazećih kapitalnih izdataka prema kontinuiranim operativnim troškovima. Uz to iz poslovanja se izdvaja za troškove usluga samo kada su iste potrebne što je bolji pristup nego li planiranje i izdvajanje visokih troškova unaprijed [17]. S poslovnim modelom baziranim na korištenju usluga i servisa u oblaku smanjeni su financijski rizici za poslovanje budući da bi se iz poslovanja plaćali samo troškovi onih usluga koje se koriste, a raskid ugovora o pružanju usluga moguće je realizirati u željeno vrijeme.

Ključni izazov s kojim se suočava poslovanje kada se koristi SaaS rješenjem za usluge poslovne inteligencije jest u činjenici da je migracija velike količine podataka u oblak financijski izdašan trošak te je preporuka koristiti

prijenosne medije za pohranu kojima bi se podaci efikasnije mogli dostaviti davatelju usluge [18]. Dodatni izazov jest i sigurnost te je s tim u vezi preporuka da povjerljivi podaci ostanu u domeni isključive dostupnosti kompanije-klijenta a da se oni podaci koji se migriraju na pohranu u oblak transferiraju i pohrane uz primjenu metoda kripto zaštite kako bi bili zadovoljeni sigurnosni zahtjevi koji s te strane mogu postojati [18].

VI. ZAKLJUČAK

Kompanije koje pravovremeno reagiraju u uvjetima kontinuiranih promjena na tržištu bit će konkurentne u dinamičnom globalnom gospodarstvu. To će poslovanju omogućiti da ima svijest o tržišnim trendovima te da može udovoljavati potrebama svojih korisnika te će omogućiti prepoznavanje prilika u pravom trenutku i brže lansiranje novih proizvoda na tržište.

Poslovna inteligencija daje poslovanju moćan instrument za analizu izvornih podataka i izvlačenje korisnih informacija iz transakcijskih izvora podataka koje pružaju ključne uvide u korisnički segment. To poslovanju daje sposobnost donošenja inteligentnih poslovnih odluka na bazi prethodno prikupljenih i obrađenih informacija. Poslovanje također može korištenjem poslovne inteligencije stjecati uvid u promjene ponašanja okoline koje bi mogle izazvati određene radnje ili angažmane korisnika. Implementiranje i podrška poslovne inteligencije u poslovanju ipak iziskuje značajna kapitalna ulaganja i potrebu analize velike količine podataka kako bi se identificirali trendovi u poslovanju. Potrebna je i značajna procesorska snaga sa značajnim reperkusijama na poslovne resurse.

Usluge računarstva u oblaku opisane su u ovom radu kao prilagodljive. Računarstvo u oblaku omogućuje skalabilnost u vremenskom periodu vršnog opterećenja i smanjenje troškova u uvjetima manje potrebe za korištenjem. Oblaci su dizajnirani da budu dinamički skalabilni kada je potreba za resursima veća, tako da se programska podrška može izvršavati na bilo kojem raspoloživom poslužitelju s viškom kapaciteta. Računarstvo u oblaku mogu koristiti klijenti-kompanije za rješavanje poteškoća koje proizlaze uporabom postojećih rješenja poslovne inteligencije na vlastitoj infrastrukturi [19].

Alatima za dubinsku analizu podataka (data mining) u SaaS rješenju poslovne inteligencije može se osigurati brže i točnije donošenje strateških poslovnih odluka. Proučavanjem ponašanja u prošlosti, poslovanje stječe poboljšano razumijevanje sadašnje situacije i bolje je opremljeno mogućnostima predviđanja odnosno utjecanja na promjene. To će omogućiti poslovanju da usvoji ispravnu strategiju za postizanje poslovnih ciljeva (npr. povećanje prodajnih rezultata).

Korištenje SaaS rješenja poslovne inteligencije donosi mjerljivu korist poslovanju, ali postoje i neki izazovi koje bi poslovanje trebalo uzeti u obzir prije korištenja te vrste usluge. Migracija velike količine podataka u oblak može biti skup zahvat a pored toga sigurnost je također upitna u slučajevima korištenja rješenja u oblaku. Unatoč tome, navedene se poteškoće mogu riješiti te je potrebno oprezno isplanirati sve aspekte poslovanja prije migracije u oblak. SaaS rješenja poslovne inteligencije pružaju fleksibilniji model bolje usklađen s ciljevima poslovanja. To omogućuje poslovanju da razvija pojedinačne IT modele temeljene isključivo na poslovnim potrebama, a ne na tehnološkim ograničenjima. Poslovanje mora pronaći balans između rizika i koristi od migracije na računarstvo u oblaku. Kada je posrijedi hrvatsko tržište poslovne inteligencije, ovo je područje još uvijek velika nepoznanica u našem gospodarstvu. U odnosu na SAD, gdje udio poduzeća koja imaju razvijene sustave za poslovnu inteligenciju prelazi 90 %, u Hrvatskoj se taj udio kreće oko 8 do 10 % svih poduzeća, s time da tu prevladavaju banke, osiguravajuće kuće i koncerni, dok su za mala poduzeća ovakvi sustavi još uvijek skupi.

LITERATURA

[1] R. Bell and G. Todd, Information Management Services: Overview of Services. Accenture, s.l., April 2007.

[2] S. Williams and N. Williams, The Profit Impact of Business Intelligence, Morgan Kaufmann Publishers, Massachusetts, 2007.

[3] L. Chisholm and A. Human, „Old Mutual's new SAP commissions system first of a kind in Africa, largest in the world", in ITWeb, http://www.itweb.co.za/office/sap/PressRelease.php?StoryID=160299 , 2006.

[4] W. Chung, H. Chen, and F. Jay, Business Intelligence Explorer: A Knowledge Map Framework for Discovering Business Intelligence on the Web, Department of Management Information Systems, Eller College of Business and Public Administration, The University of Arizona, Tucson, AZ., 2002.

[5] N. Maini, Leveraging Siebel Analytics to Enhance Sales in Pharmaceutical Industry. Accenture, s.l., 2010.

[6] S. LaValle, Business analytics and optimization for the intelligent enterprise, IBM Institute for Business Value, IBM Global Services, Route 100, Somers, New York, 2009.

[7] I. Foster, Y. Zhao, I. Raicu, and S. Lu, Cloud Computing and Grid Computing 360-Degree Compared, Department of Computer Science, University of Chicago, Chicago, November 2008.

[8] P. Daugherty and S. Raj, „Cloud Computing and Software as a Service", in Accenture, http://www.accenture.com/cloud/, 2008.

[9] A. T. Velte, T. J. Velte, and R. Elsenpeter, Cloud Computing: A Practical Approach, McGraw-Hill/Osborne, San Francisco, 2010.

[10] EduCause, 7 Things you should know about virtualization, http://net.educause.edu/ir/library/pdf/EST0904.pdf , 2009.

[11] Microsoft, Forecast: Improved economy in the cloud, www.slideshare.net/govloop/cloud-services-whitepaper , 2010.

[12] K. S. Swaminathan, P. Daugherty, and J. F. Tobolski, „What the Enterprise Needs to Know about Cloud Computing", Accenture Technology Labs, http://www.accenture.com/accenturetechlabs/ , 27 April 2009.

[13] T. Sullivan, „BI steps into the cloud" in InfoWorld, http://www.infoworld.com/d/cloud-computing/bi-steps-cloud-085, 2008.

[14] H. Li, J. Sedayao, J. Hahn-Steichen, E. Jimison, C. Spence, and S. Chahal, Developing an Enterprise Cloud Computing Strategy. Intel Information Technology, Computer Manufacturing, Cloud Computing, USA, 2009.

[15] T. Schadler, Should Your Email Live In The Cloud ? A Comparative Cost Analysis, Forrester Research, Inc., USA, 2009.

[16] D. Raths, Cloud Computing: Public-Sector Opportunities Emerge, from http://www.govtech.com/gt/387269 , 2008.

[17] T. Schadler, Talking To Your CFO About Cloud Computing, Forrester Research, Inc., USA, 2008.

[18] A. All, „Pros and Cons of Business Intelligence in the Cloud", IT Business Edge, http://www.itbusinessedge.com/cm/blogs/all/pros-and-cons-of-business-intelligence-in-the-cloud/?cs=35060 , 2009.

[19] A. Stipić and T. Bronzin, How Cloud Computing Is (not) Changing the Way We Do BI, 35th International Convention MIPRO 2012, Opatija, Croatia, 21-25 May 2012.

Končar MIS – snažan alat za mudra poduzeća

Anamarija Frančić, Damir Ferenčak, Drago Cmuk
Končar-Inem/PJ ICT sustavi, Zagreb, Hrvatska
{afancic, dferencak, dcmuk}@koncar-inem.hr

Sažetak - Menadžment srednjih i velikih poduzeća često je pozvan donositi odluke koje ovise o velikoj količini dostupnih i pravilno organiziranih informacija. Bez obzira radi li se o turbulentnom tržištu ili naglim zakonskim promjenama - problem je sličan. Takve se potrebe najčešće javljaju upravo u kriznim vremenima, a samo poduzeća pripremljena za njih mogu ih prevladati. Usred takvih okolnosti Končar - Elektroindustrija d.d. odlučila je investirati u provedbu novog Menadžerskog informacijskog sustava (MIS) sa sljedećim prednostima: automatizirano prikupljanje podataka iz različitih poslovnih sustava, pristup detaljnim informacijama (bolja drill-down sposobnost), visoka fleksibilnost u generiranju izvješća i jednostavnija prilagodba zakonski definiranih izvješća propisima.

Informacije koje MIS stavlja na raspolaganje korisnicima pomažu na strateškim i operativnim razinama odlučivanja. Ti se podaci temelje na bruto bilanci, troškovima računu dobiti i gubitka, novčanom toku prodaji, ugovorenim i zaključenim poslovima, te utječu na sve upravljačke razine. MIS osigurava pristup svim za to važnim podacima – stanju, povijesti i srednjoročnim planovima. Veliki broj pokazatelja uspješnosti omogućuje upravljanje na svim razinama. MIS je trenutno u završnoj fazi razvoja i implementacije od strane Končar - Elektronika i informatika d.d.

I. UVOD

U današnjim tržišnim okolnostima neprijeporna je činjenica kako sve veći dio poslovanja ovisi o potpori informacijskih sustava [1, 2, 3]. Time poslovna vrijednost informacijskih sustava stalno raste i čini ih nezaobilaznim, ključnim, strateškim poslovnim elementima.

Menadžment srednjih i velikih poduzeća često iziskuje obavljanje radnji koje su ovisne o velikoj količini dostupnih i pravilno organiziranih informacija. Bilo da se nose sa čestim zakonodavnim promjenama ili sa poslovanjem u uvjetima tržišne krize, problem je sličan i zahtijeva brze reakcije. Da bi izveli "kirurški precizan zahvat" potreban je dobar i sveobuhvatan pregled situacije, s povjerenjem u izvješća koje informatički sustav kompanije mora osigurati [3, 4]. Ipak, veliki će se dio uprava, nadzornih odbora i drugih upravljačkih tijela prečesto oslanjati na krute, obično predefinirane vrste izvještaja u diskretiziranim vremenskim intervalima, koji obično boluju od dvije „bolesti“. Prva je svakako prevelika količina podataka, koja u takvoj vrsti izvještavanja omogućuje da se „od šume ne vidi drvo“; a druga je svakako bolest nedovoljne fleksibilnosti. Prvi od tih problema u praksi često ne rješavaju ni napredni BI sustavi, jer zahtijevaju razumijevanje koncepata od kojih članovi uprava nerijetko zaziru (a ponekad i opravdano). Cijela situacija zna kulminirati na sastancima na kojima se donose odluke: ključni ljudi tada često raspolažu svim informacijama – osim onih koje su im zaista potrebne [3,5].

Informacija (naročito povratna) je osnova svake uspješne kontrole u poduzeću, pa odatle i potreba da si menadžment osigura kvalitetne i pravovremene informacije. Kako bi se to postiglo Grupa Končar unapređuje Menadžerski informacijski sustav za menadžment (MIS) .

II. MIS U KONČARU

Grupa Končar danas se sastoji od dvadeset vlasnički povezanih društava i jednog pridruženog društva sa snažnom poslovnom suradnjom među članovima grupe. Društva Grupe Končar do sada su podatke o poslovanju vodila po strukturi koja je bila u skladu sa zakonskim propisima. U međuvremenu se zakonski okvir promijenio. Hrvatska agencija za nadzor financijskih usluga donijela je u travnju 2011. godine *Pravilnik o obliku i sadržaju financijskih izvještaja* za razdoblja tijekom godine i to za društva čije su dionice uvrštene na Zagrebačku burzu. Tim pravilnikom su podaci iz postojećeg sustava postali nedostatni za izradu tromjesečnih i godišnjih konsolidiranih financijskih izvještaja, te je donijeta odluka o nadogradnji postojećeg sustava.

Novo rješenje predviđa da se podaci automatski učitavaju iz Glavne knjige društva i to na razini konta. Takav način prikupljanja podataka omogućit će veću fleksibilnost sustava odnosno imunost na probleme čestih izmjena oblika i sadržaja obrazaca financijskih izvještaja. Financijski izvještaji MIS-a prilagođeni su eksternim izvještajima (GFI-POD, TFI-POD).

Na izradi ove aplikacije radi tim koji se sastoji od grupe stručnjaka iz područja financija i računovodstva, aplikativnih specijalista s dugogodišnjim iskustvom na području poslovnih aplikacija i iskusne grupe web programera. Njihov naporan rad na razini planiranja, razrade funkcionalnih specifikacija ali i napredne arhitekture te suvremenog korisničkog sučelja sa snažnim alatima za fleksibilno izvještavanje rezultirao je razvojem sveobuhvatne aplikacije koja pored osiguravanja informacija uvelike olakšava redovno poslovanje i izvještavanje. Nadalje, ona pojednostavljuje i pripremu za konsolidaciju, koja obuhvaća uskladu internog prometa unutar Grupe, i samu konsolidaciju odnosno eliminaciju internog prometa.

Web arhitektura aplikacije korisnicima omogućava unos informacija i pristup njima na brz i siguran način s bilo kojeg mjesta na kojem postoji standardni preglednik.

Jezgru sustava čine dvije ključne aplikacije – MIS i Konsolidacija.

1. ***Moduli aplikacije MIS:***
   a) Financijsko izvještavanje
   b) Nefinancijsko izvještavanje
   c) Izračunavanje financijskih pokazatelja
2. ***Moduli aplikacije Konsolidacija:***
   a) Usklade
   b) Izvještavanje o uskladama
   c) Usporedbe sa podacima MIS-a
   d) Korekcije
   e) Financijsko izvještavanje
   f) Izračunavanje financijskih pokazatelja
   g) Vlasnička struktura

Međusobna interakcija ova dva sustava prikazana je na Slici 2.

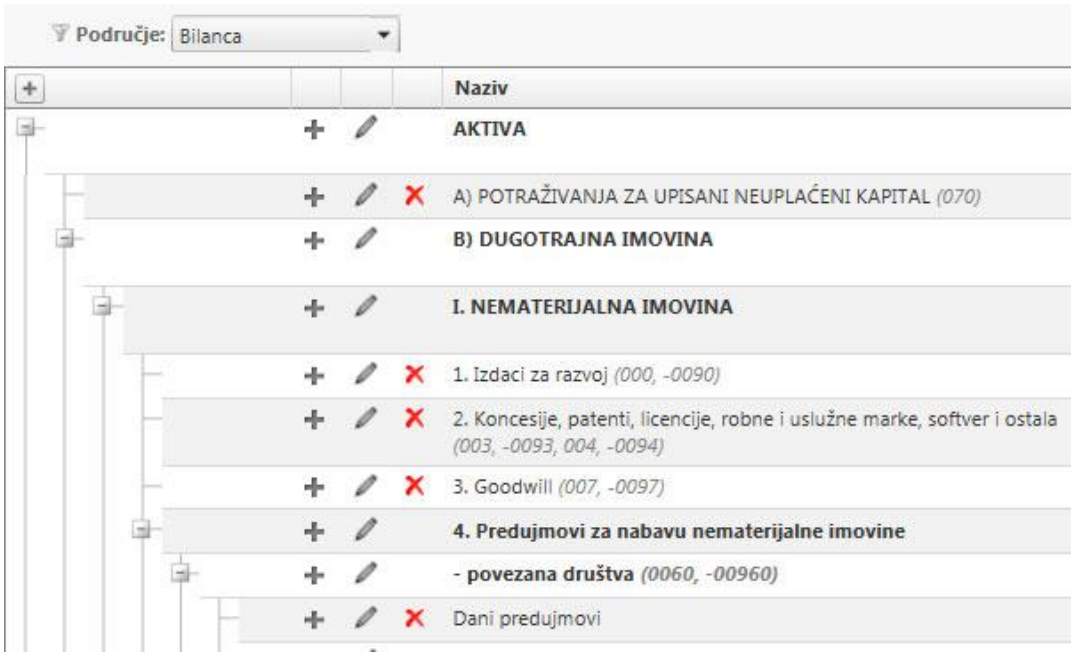
Poslovno izvještavanje na razini neke grupacije izrazito je složen proces čija kompleksnost je proporcionalna stupnju nezavisnosti pojedinog poduzeća unutar grupacije. U Končar grupaciji su sva poduzeća potpuno nezavisna, izuzev vlasničke strukture, i zbog toga postoji velika razina heterogenosti u poslovnim sustavima koja pojedina poduzeća koriste.

Samim time, izrazito je otežana integracija tih podataka a situacija se otežava u slučaju zakonskih promjena (koje su u stanju recesija i tranzicije) izrazito česte [6]. Na taj način troše se znatni resursi na stalnu nadogradnju parcijalnih rješenja, a menadžment ostaje bez informacija važnih za ispravno donošenje odluka. Kako bi se to spriječilo, razvojni je tim dizajnirao sustav na način da se omogući:

1. ***Automatizirano prikupljanje podataka iz različitih poslovnih sustava***

Izgradnja izvještajnog sustava prikladnog za poslovno odlučivanje samo po sebi je složen proces. Samo kvalitetna i pravovremena informacija je ono što čini razliku „dobrog“ i „lošeg“ poslovnog sustava. Težnja ka izvrsnosti Končar grupacije implementirana je i u ovaj sustav a glavna misao vodilja je i naš slogan „tradicija, znanje,odgovornost“. Heterogenost poslovnih sustava bila je velik izazov pri dizajnu MIS-a. Kako bi se osiguralo da iz poslovnih sustava pravovremeno stižu informacije a da ih pri tome sustav tumači na istovjetan način omogućena je automatska razmjena podataka između poslovnog sustava društva i MIS-a bez ručnog prepisivanja na mjestima gdje je to moguće, koristeći jasno propisanu XML strukturu i web servise. Podaci se prikupljaju na nivou konta Glavne knjige. Osim same činjenice da se zahvaćaju konta i promet po kontima sustavom je definirano i za koje razdoblje se prikupljaju podaci te je dana jasna uputa računovodstvenim službama na koji način moraju provoditi knjiženja u svojim poslovnim sustavima (usklada sa kontnim planom Grupe).

Sustav prihvaća dvije XML datoteke pri čemu jedna služi za generiranje izvještaja Bilance, Računa dobiti i gubitka (podaci su salda dugovnog i salda potražnog prometa na nivou razdoblja i konta) a druga za generiranje Novčanog toka (podaci su salda dugovnog i salda potražnog prometa na nivou razdoblja i konta koja



Slika 1 – Drill down funkcionalnost

se odnose na transakcije vezane za novčana sredstva). Sustav prema zadanim kriterijima vrši logičku i formalnu provjeru integriteta podataka. Svako društvo odgovorno je za svoje podatke pa sukladno tome sve dok je aktivan period za unos/razmjenu podataka (definira se radnim kalendarom) društva mogu „slati“ podatke iz svojih poslovnih sustava.

2. ***Prikupljanje podataka na nivou konta Glavne knjige***
3. ***Skraćeno vrijeme potrebno za unos podataka***
4. ***Unapređenje procesa***
   a) Poboljšana kontrola unosa (radni kalendar)
   b) Mogućnost povijesnog pregleda transakcija
5. ***Fleksibilan izvještajni sustav***
   a) Mogućnost korisničkog definiranja predložaka
   b) Analiza podataka (statička i dinamička)

Analiza financijskih izvještaja najvažniji je dio cjelokupne analize poslovanja društva. U vrijeme donošenja godišnjih temeljnih financijskih izvještaja javlja se pojačan interes za njihovom analizom a onda i naprednim mogućnostima koje informacijski sustav donosi.

Na temelju sustavne analize može se donijeti mjerodavan sud o poslovnim performansama promatranog društva . Analiza omogućava odgovor na pitanje kako je poduzeće poslovalo u prethodnom razdoblju i kako poboljšati rezultat u narednom.

Analiza financijskih izvješća ima dva cilja: analizirati stanje imovine te analizirati stanje kapitala i obveza društva.

Za analizu stanja imovine, kapitala i obveza društva, podloga je bilanca na određeni dan, dok je za analizu prihoda i rashoda te poslovnog rezultata podloga račun dobiti i gubitka za promatrani period.

MIS omogućava korisniku analizu podataka na osnovu mnoštva unaprijed zadanih i strukturiranih izvještaja ali mu isto tako pruža mogućnost ad-hoc analize. Korisniku sustava je na taj način omogućeno da promatra podatke za određena razdoblja, kreira razlike, indekse te na osnovu toga donosi zaključke. Svaki od tako generiranih izvještaja moguće je integrirati i sa MS Office alatima za analizu podataka.

Pored ovakve analize podataka važno mjesto zauzima i analiza ekonomičnosti, proizvodnosti i rentabilnosti koja je osigurana nizom financijskih pokazatelja.

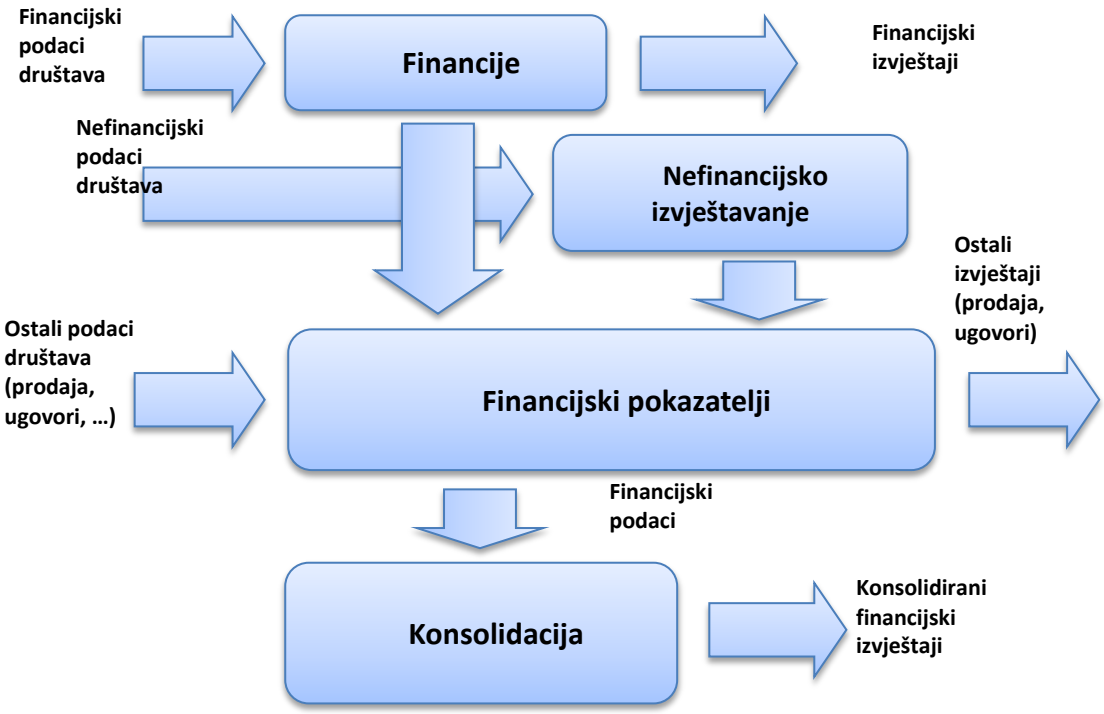
c) Pristup detaljnim informacijama (drill-down)

Izvještajni sustav obuhvaća niz izvještaja sa zadanom strukturom i prikazom na zadanim razinama detalja.

Međutim kako se podaci prikupljaju na analitičkom nivou moguće je i podatke prikazati na većoj ili manjoj razini detalja. Sustav sam rekurzivnom metodom izračunava koliko nivoa je moguće prikazati te. omogućava korisniku da se korištenjem funkcionalnosti drill- downa „šeće“ kroz podatke. Drill down funkcionalnost prikazana je na Slici 1.

Ovisno o ovlastima korisnika, korisnik može ponirati od sintetičke do analitičke stavke za svoje društvo ili pak određenu skupinu društava pa sve do cijele Grupe. Svaki izvještaj moguće je pretvoriti u izvještaje proizvoljnog formata.

d) Sposobnost stvaranja ad-hoc izvješća koja nadilaze standardna, unaprijed definirana izvješća
e) Kreiranje stavaka uvođenjem formula



Slika 2 - Struktura sustava MIS

6. ***Fleksibilnost (prema zakonodavcu, prema korisniku)***
7. ***Otvorenost aplikacije za eventualne promjene i dorade postojećih parametara i izvještaja***

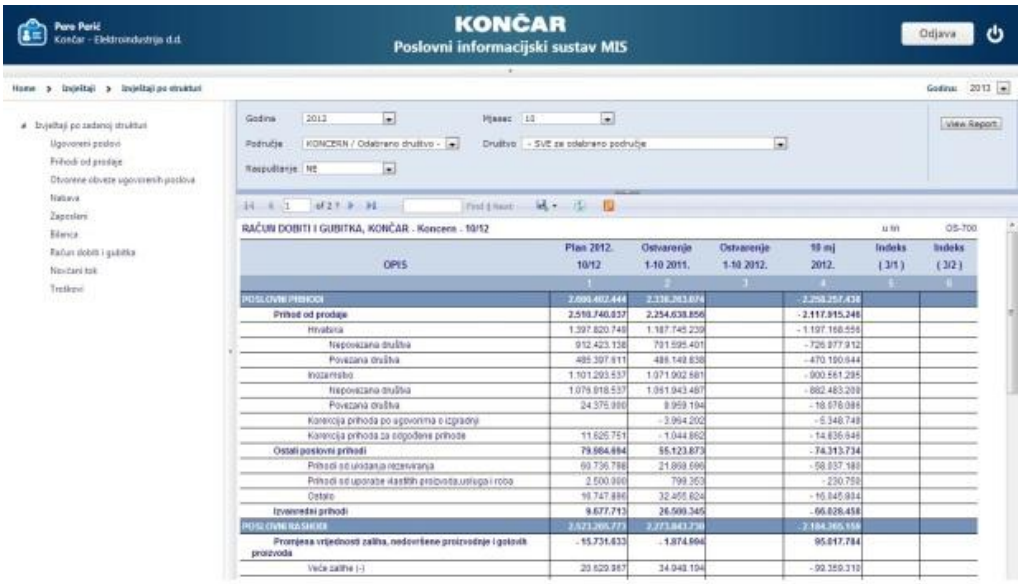
Ključna značajka aplikacije MIS je „imunost“ na zakonske promjene koje se mogu očekivati u dijelu izvještavanja za financijska područja (Bilanca, Troškovi, Račun dobiti i gubitka te Novčani tok). Jedan od mehanizama osiguravanja fleksibilnosti sustava je definiranje zajedničkog kontnog plana Grupe. Sve pozicije izvještajnog sustava ustvari su rezultat stanja na kontima Glavne knjige. Primjenom zajedničkog kontnog plana bilo je moguće projektirati sustav na način da je korisnik sam taj koji kreira izvještaje i brine o tome od kuda podaci dolaze. Srce aplikacije MIS je modul za održavanje pozicija izvještajnog sustava te modul za kreiranje predložaka izvještaja. Na taj način sustav je u cijelosti otvoren korisniku odnosno optimalno konfigurabilan. Svaka sljedeća promjena zakona ili načina knjiženja više ne zahtjeva intervenciju programskog odjela već samo jednostavnu izmjenu predloška putem sučelja aplikacije.

Informacije koje sadrži MIS podržavaju sve tri razine odlučivanja: strateške, taktičke i operativne. Izvještajni sustav organiziran je u logičke cjeline što uvelike olakšava korisniku da locira informaciju koju traži (Slika 3). Izvještaji Bruto bilanca, troškovi, račun dobiti i gubitka, novčani tok, prodaja, ugovoreni i zaključeni poslovi su osnova za sve tri razine odlučivanja, budući da aplikacija pamti dinamičke podatke iz prošlosti i budućnosti (srednjoročni planovi).

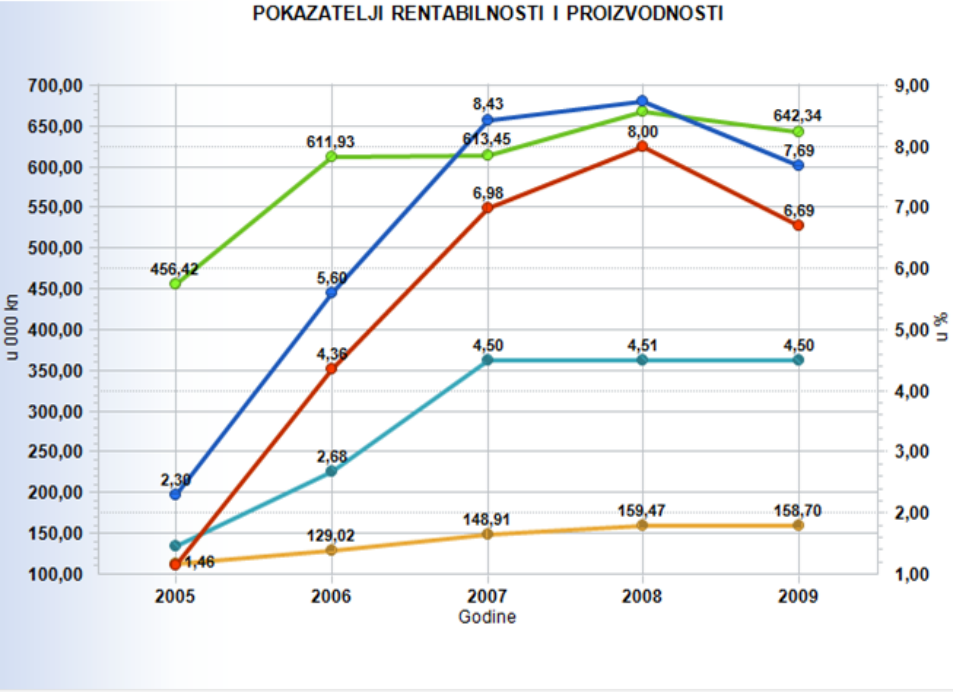
Veliki broj pokazatelja uspješnosti zasnovan na tim informacijama jesu podloga za sam proces upravljanja kako na nivou pojedinih društava tako i Grupe. Sve ove značajke imaju veliki potencijal za bolje, brže i preciznije analize financijskih i drugih podataka, što je važna podloga upravljanju i donošenju odluka. U uvjetima krize, to ponekad može napraviti dugoročni učinak i donijeti značajnu prednost poduzećima kojima se upravljanja na takav način. Na Slici 4 prikazani su Pokazatelj rentabilnosti i proizvodnosti na osnovu kojih se analiziraju i prate se povrat na kapital (ROE), povrat na prodaju (ROS), prodaja proizvoda i usluga po zaposlenom (PPUZ) te novoostvarena vrijednost (VA). Ovo su samo neki od brojnih pokazatelja implementiranih u MIS.

Informacije su u sustavu grupirane u nekoliko područja koja su pokrivena sa više od 300 predefiniranih izvještaja:

1. ***Financijsko Područje***
   a) Bilanca
   b) Račun dobiti i gubitka
   c) Novčani tok
   d) Troškovi
2. ***Nefinancijsko Područje***
   a) Ugovoreni poslovi
   b) Prihodi od prodaje
   c) Otvorene obveze ugovorenih poslova
   d) Nabava
3. ***Zaposleni***
4. ***Pokazatelji poslovanja***
   a) Ključni financijski pokazatelj
   b) Pokazatelji tijekom godine (rentabilnosti i proizvodnje, likvidnosti, aktivnosti, EBIT, EBITDA)
   c) Petogodišnji pokazatelji (rentabilnosti i proizvodnje, likvidnosti, aktivnosti, EBIT, EBITDA)
   d) Rangiranje društava



Slika 3 - Izvještajni sustav



Slika 4 - Pokazatelji rentabilnosti i proizvodnosti

5. *Konsolidacija*

a) Interni izvještaji za Nadzorne odbore
b) Eksterni izvještaji – Komisija za vrijednosne papire
c) Izvještaji o konsolidacijskim postupcima
d) Pokazatelji poslovanja

6. *Srednjeročno planiranje*

a) Prodaja
b) Bilanca
c) Račun dobiti i gubitka
d) Novčani tok

III. ZAKLJUČAK

Problem konsolidacije i menadžerskog izvještavanja za sva povezana i heterogena poduzeća problem je za koji ne postoje jednostavna rješenja, osobito kad se sustav optereti čestim zakonskim promjenama. Opisanim sustavom se predviđa unapređenje postojećih procesa te stvaranje dodane vrijednosti kako za maticu društva Končar tako i za sva Končareva društva i sve korisnike sustava. Zbog primjene zajedničkog kontnog plana Grupe, omogućena je velika kontrolabilnost ali i fleksibilnost sustava, kao i primjena realiziranog rješenja na drugim društvima srodne uređenosti.

MIS je trenutno u završnoj fazi razvoja i implementacije od strane KONČAR - Elektronika i informatika d.d. U zahtjevnim ekonomskim situacijama, mudro je ulagati u optimirane informacijske sustave. Kada su napravljeni dobro, oni zaista mogu olakšati i svako dnevno poslovanje, i donošenje teških odluka.

IV. LITERATURA

[1] Tintor, J. :"Poslovna analiza", Zagreb, 2009
[2] Gulin, D; Perčević, H; Tušek, B; Žager, L., „Poslovno planiranje, kontrola i analiza", Zagreb, 2012.
[3] SAP Thought Leadership Paper: Making Information Governance a Reality for Your Organization (Maximize the Value of Enterprise Information, SAP 2013, preuzeto sa http://goo.gl/0XJAD, veljača 2013.
[4] Herman Aguinis :"Performance Management", Second Edition, 2008.
[5] Ante-Željko Batarelo :"Informacije, informacijski sustavi i tehnike prikazivanja u funkciji poslovnog odlučivanja", Poslovna analiza i upravljanje br 3, Zagreb, 2000.
[6] Lj. Vidučić: „Financijski management", RRiF, Zagreb, 2006.
[7] Krakar, Z.: «Upravljanje informatizacijom», Fakultet organizacije i informatike, Skripta, Varaždin, 2005.

DE

International Conference on

DIGITAL ECONOMY – 10th Alpe Adria Danube Universities Initiative (ALADIN)

Steering Committee

Chairs: Edvard Tijan, University of Rijeka, Croatia
Mladen Mauher, Polytechnic of Zagreb, Croatia

Members: Andras Gabor, Corvinus University of Budapest, Hungary
Jože Gričar, University of Maribor, Kranj, Slovenia
Walter Ukovich, University of Trieste, Italy
Christian Kittl, Evolaris Research Lab, Graz, Austria
Anton Lavrin, Technical University of Košice, Slovakia
Bernhard Katzy, Universität der Bundeswehr München, Germany

PAPERS

Business Valuation in Oil&Gas Industry: New Challenges

Sonja Brlečić Valčić*, Branka Crnković-Stumpf ** and Jana Katunar**
* Saipem SpA Croatian Branch/AFC, Rijeka, Croatia
** University of Rijeka, Faculty of Economics, Rijeka, Croatia
sonja.brlecic@gmail.com, branka@efri.hr, jana@efri.hr

Abstract - **Business valuation of companies in oil and gas industry can be described as specific and difficult in many aspects within existing methodology and currently used approaches in appraisal process. Identification and evaluation of the significant factors that create and determine enterprise value in oil and gas industry is based on complex calculations involving many variables. Regardless of this reason, existing business valuation methods for such companies have to be improved with taking into account a numerous qualitative and even additional quantitative factors. Therefore, economic experts and scientists in the field of business valuation are confronted with new challenges in determination of appropriate approaches that should be able to eliminate the disadvantages of existing valuation methods. In this paper, authors have suggested some guidelines for setting up the development of valuation approach based on application and adaption of selected financial and non-financial indicators by means of artificial neural networks and fuzzy logic. Importance of properly configured commercial databases of oil and gas industry indicators with capability of optimal data processing is also emphasized.**

I. INTRODUCTION

According to the estimated population growth to 8.3 billion people until 2030, and in line with the standard of living and the industry accompanying such growth, energy consumption has been predicted to yearly increase by 1.6%, i.e. global energy consumption in 2030 will be 36% greater than in 2011 [1]. Along with the increase in energy consumption, the same source predicts an increase in production of energy resources, of which oil 116.26% (from 3995.6 million tons in 2011 to 4645.5 million tons in 2030), and natural gas 144.87% (from 2954.8 million tons in 2011 to 4280.7 million tons in 2030). This illustrates the importance of Oil&Gas (O&G) industry within the global industrial movements in the near future. The significance of business valuation within the O&G industry is best illustrated by the average of as much as 4 transaction activities of such companies a day (e.g. 1616 transactions in 2012 with the sum of 402 trillion US dollars) with the tendency of increase in the number and the value of transactions in the following years [2]. Regardless of whether they are related to entire companies, parts of companies or individual assets, mergers and acquisitions in the O&G industry take place primarily as a result of the search for additional oil and gas resources, and secondarily for reasons of strengthening the position on the global market and establishing a presence on certain markets, and acquiring an access to increasingly demanding technology, in realizing large and challenging projects, as well as to the knowledge needful for their realization (know-how).

Several essential sectors define the O&G industry and each of them is focused on specific processes within the value chain. Being that this highly active industry is marked by significant complexities, many companies focus on just one of them. Thus, in the jargon of this industry, one often comes across a division into:

- Exploration and Production (E&P) Companies (Upstream) - perform all operations associated with finding and extracting O&G;
- Transportation and Storage Companies (Midstream) - perform all operations associated with storing O&G and transporting it from fields to refineries and processing plants;
- Refining and Marketing (R&M) Companies (Downstream) - perform crude oil refining into petroleum products and selling them to end-users;
- Oil Field Services (OFS) - provide infrastructure, equipment, drilling assistance, intellectual property and all other services required by international O&G industry;
- Integrated Majors - mostly huge global companies that operate across multiple segments and with E&P sector include all or some of aforementioned O&G sectors.

Each of the mentioned sectors of the O&G industry has its specificities that affect planning of future costs and revenues, which in turn influences determining the company's value. The most important differences between these company types arise from the fact that O&G industry is very cyclical and dependent on commodity prices. One of the most frequent characteristics of E&P companies, but often present in other O&G sectors as well, relates to the inability to control revenues. Furthermore, this kind of companies are highly asset-centric, therefore the revenues of E&P companies will depend on the reserves, while the revenues of companies pertaining to other O&G groups will depend on technologies supporting this industry. The differences in accounting treatments and financial statement items, which differ from those of other industries, also classify

these companies into distinguished companies. The above mentioned specificities are manifested in production processes, production equipment and business operations [3], and result primarily in accounting [4] and financial [5] [6] characteristics. The tendency of oil and gas prices, and risks that accompany this industry greatly affect financial results of these companies [7] [8], which in addition to the above mentioned specificities makes business valuation within this industry very distinctive [9] [10].

For reasons mentioned above, valuation tools are not applicable in O&G industry without certain adjustment and further analyses necessary for modeling trends of costs and revenues in future periods. Thus, the scope of the few scientific papers related to valuation of O&G industry companies mostly includes additional expertise needful for valuating such companies, as well as implementation of appropriate valuation tools. The valuation of O&G companies is based on a combination of three conventional approaches, namely, market, income and asset approach [11]. Due to the characteristics of processes in cash flow, shortcomings of the most popular Discounted Cash Flow (DCF) method in valuating such companies make this method almost inapplicable, when used exclusively [10]. Hence, in addition to the combination of standard valuation methods adapted to this industry, most commonly used tool is the ROACE ratio (Return on Average Capital Employed) [9].

Most challenges in the existing business valuation procedures in the O&G industry are related to the prediction of O&G prices by modeling different macro and microeconomic scenarios, the lack of published specific information regarding each individual segment of business, normalization of accounting items, precise determination of focus of the E&P companies and the necessity of detailed analysis of the rate of creating revenues and costs of various activities within the O&G industry with Integrated Majors Companies.

On the other hand, new challenges of business valuation in O&G industry are focused on determining interdependencies of quantitative parameters reported in the financial statements. Furthermore, this applies to the qualitative parameters that affect the creation of company value as well, but are not quantitatively stated in those statements and are not taken into consideration by the existing valuation methods. Consequently, a need arises for a different kind of approach to business valuation that can meet these challenges. Such an approach is manifested primarily in the implementation of algorithms of computational intelligence, namely, artificial neural classification and approximation networks, and fuzzy logic.

II. BUSINESS VALUATION IN OIL&GAS INDUSTRY

The existing procedures in O&G business valuation are based on the combination of three basic valuation approaches:

- the asset approach, in which the value of an asset is estimated based on the capacity to create revenues,
- the market approach, in which the market value (the value of the already realized transaction) of similar companies within the industry is compared to the company to be valued,
- the income approach, which predicts creating of future cash flows, and by discounting them to the present value, the market value is determined.

In the E&P sector, the asset approach is mostly used in valuation of oil and gas reserves and their capacity to create future revenues with regard to the treatment of the reserves, which can be proved (reserves that are expected to be produced from existing wells), probable (there should be at least 50% chance of being technically and economically producible) and possible (estimated to have a significant, but less than 50%, chance of being technically and economically producible).

In other sectors, this approach incorporates complex valuation studies of technological assets in terms of their useful life, the assessment studies concerning replacing costs of certain assets, and the assessment of the capacity of creating revenues of certain asset sectors (individual assets or facilities).

In the market approach to business valuation using the most commonly applied method of Comparable Public Companies & Precedent Transactions [12], there are certain differences in determination of comparable parameters, especially in the E&P sector and some parts of the Integrated Majors sector that deal with exploitation and production, which can be expressed in the following manner:

- instead of the comparison by types of revenues or EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortization), a comparison is performed using a selection based on proved reserves or daily production, in addition to the usual geographical and industrial criteria,
- instead of the traditional metrics, such as revenues or EPS (Earnings Per Share), the index of metrics and multiples is based on EBITDAX (Earnings Before Interest, Taxes, Depreciation, Depletion, Amortization and Exploration Expenses), proved reserves, daily production, oil mix (revenue ratio obtained by oil and gas extraction), etc.

Common metrics for comparison of this type of companies are shown in Table 1 [12].

Considering pricing metrics mentioned in Table 1, the comparison of the already sold companies and companies to be valued is performed so that the enterprise value (EV) of an already sold company is compared with one or more mentioned metrics of another already sold company (e.g. EV / Proved Reserve quantities, EV / Daily Production, EV / EBITDAX, etc.).

The mentioned metrics are calculated for the company to be valued as well, and by comparing the metrics of the already sold company and the metrics of the company to be valued, the value of the company under valuation is determined.

In income approach to company valuation, the commonly used method is DCF method based on the theory that the value of a business is the sum of the expected future cash flows discounted to the present value by the corresponding discount rate.

In appraising the value of O&G companies by implementing the DCF method, key differences are related to the following factors [12]:

- supplemental non-cash expenses in addition to the common expenses, such as DD&A (depreciation, depletion and amortization) and stock-based compensation,
- usage of multiples of daily production, EBITDA or EBITDAX instead of the multiple based on free cash flow as final output,
- in predicting long-term growth using the Gordon Growth model, the assumed growth rate for E&P companies is 0% because O&G assets get depleted over time, thus it is assumed that with time the amount of well resources becomes equal to zero,
- for the O&G companies, instead of WACC (Weighted Average Cost Of Capital), industrial standard discount rate of 10% is being used,
- in this industry, commodity prices are commonly used instead of revenue growth variables or EBITDA margins in setting the sensitivity tables.

Considering the above, the disadvantages of the DCF method in O&G business valuation are the following:

- the obtained present value is sensitive to the assumptions and methodology,
- terminal value is almost impossible to calculate because it depends on long-term predictions of oil and gas prices,
- high CapEx (capital expenditure) requirements that reduce free cash flow may create decreasing or even negative cash flow,
- in order to obtain realistic projected future financial trends, it is necessary to observe them within at least one entire production cycle of 7-10 years for the cash flow movements to normalize.

Since the E&P sector is the most sensitive to the above disadvantages of the DCF method, the so-called Net Asset Value (NAV) model is used as an alternative. The NAV model assumes that the company never increases the existing reserves, thus there will be neither CapEx nor reinvestments in the future. Consequently, the model is (in most cases) based on the existing proved reserves, and sometimes the analysts take into consideration probable and even possible reserves. Unlike the DCF model used to value the company on the corporative level (valuing the entire company), the NAV model values the company starting at the asset level (the value of the reserves), and based on future projections determines the value of the company as a whole at the end of the valuation. In other words, the NAV method bases the value of the company on the existing oil and natural gas reserves, and the income that can be generated from such reserves until the end of their useful life.

Business valuation within the O&G industry requires a combination of all three approaches (asset, market and income approach). Heretofore, no single method has been designed that would make this possible. Each of these approaches has its shortcomings, especially in complex industries as O&G, and due to the increasing market

TABLE I. KEY OPERATING METRICS AND RATIOS FOR E&P SECTOR OF O&G INDUSTRY [12]

Metric	**Calculation method(s)**	**What does it mean?**
EBITDA	Operating Income + DD&A (Depreciation, Depletion & Amortization)	How healthy is our cash flow?
EBITDAX	EBITDA + Exploration Expense on Income Statement	How healthy is our cash flow, regardless of whether we expense or capitalize Exploration?
Production Costs Per Mcfe (1000 cubic feet equivalent) or BOE (Barrels of Oil Equivalent)	Total Production Expense / Total Production Volume	How much does it cost us to produce 1 unit of energy from the ground?
Finding & Development (F&D) Costs - All Sources (Mcfe or BOE)	(Development + Exploration + Acquisition Expenses) / Net Reserve Additions	How much does it cost us to add 1 unit of energy to our reserves?
F&D Costs - Excluding Purchases and Sales (Per Mcfe or BOE)	(Development + Exploration Expenses) / (Reserve Revisions, Extensions, Additions, and Discoveries)	How much does it cost us to add 1 unit of energy to our reserves without buying assets or other companies?
Production Replacement Ratio - All Sources	Net Reserve Additions / Annual Production Volume	Are we replacing all the reserves we're depleting each year?
Production Replacement Ratio - Excluding Purchases and Sales	(Reserve Revisions, Extensions, Additions, and Discoveries) / Annual Production Volume	Are we replacing all the reserves we're depleting each year, even without acquiring other assets or companies?
Proved Reserves	Search in the annual report	How much more energy can we produce if we never find or purchase additional reserves?
Proved Developed Reserves / Proved Reserves	Search in the annual report	What % of the Proved Reserves are we 100% certain to produce?
Reserve Life Ratio or R/P Ratio	Proved Reserves / Annual Production Volume	How many years until we run out of energy completely?
Daily Production	Annual Production Volume / Days in Year	How much energy do we produce each day?
Oil Mix (%)	Proved Oil Reserves / Total Proved Reserves or Oil Production / Total Production	Are we mostly an oil company, or mostly a natural gas company?

uncertainty, it is very difficult to predict future movements (especially of resources and resource prices). Consequently, it is often necessary to perform additional financial analyses that include a certain number of significant metrics. The selection of those metrics is very important due to the fact that almost 120 such parameters have been identified in valuation procedures of various industries [13]. For this very reason, it is necessary to explore interdependency of those financial parameters (individually or in certain groups) and the value of the company reached by the sale on the market in order to examine which of those factors, to what extent, and by which criteria influence the creation and accumulation of the company value.

Furthermore, attention should be directed to the need for an analysis and inclusion of relevant qualitative metrics that affect movement and business sustainability, as well as positive business results. Some of them are: intellectual capital, know-how, development of human resources, business development, base of clients and vendors, relation with clients and suppliers, quality of products and/or services, organization structure and management quality, marketing, etc.

Studying and analyzing qualitative factors that create value, the international sustainability agency Two Tomorrows places them among five key categories, or domains [14]:

- Strategy: Alignment between sustainability efforts and core business strategy, management of major sustainability impacts, opportunities and risks;
- Governance: Quality of top-level governance of sustainability issues;
- Engagement: Extent to which stakeholder concerns are understood and acted on;
- Value Chain: Management of impacts through the value chain from suppliers to distributors, including the lifecycle of products;
- Innovation: Effectiveness of efforts to develop products and services that address social and environmental challenges in a profitable and scalable way.

The fact that the existing valuation methods do not have the capacity to take the mentioned qualitative metrics into account creates a need for the development of a methodology that would be able to quantify these metrics that could be used with the usual financial quantitative metrics. It is obvious that an alternative or hybrid approach for O&G business valuation, which will have the capacity to use all the previously identified quantitative and qualitative factors that affect the value of the company, should be developed. One possible approach that meets these requirements is presented below.

III. A New Approach in Business Valuation of Oil&Gas Companies

The improvement of information and computer technology made way for the progressive growth of the methodological spectrum available in various areas for solving numerous problems. In addition to the statistical and operational research methods, models of computational intelligence based on the learning theory have contributed to the possibility of using prior knowledge for effective decision making. Particularly, computational intelligence is proven to help in solving the problems of identification of parameters in a system, as well as in determining their interdependencies (artificial neural networks). This even includes fusion of completely different and usually incompatible information with the aim to assess their influence to the final, beforehand unknown, information (fuzzy logic).

Artificial neural networks (ANN), as one of the computational intelligence methods structured after the human brain, started to be developed in 1943 in paper of McCulloch and Pits. Since then, they are constantly being developed, so the computational intelligence based on the learning theory enhanced the possibility of using prior knowledge (via expert systems and fuzzy logic) and data (via ANN) for complex information processing. These models are often used for predicting bankruptcy [15] [16] and stock market trends [17], but the field of implementation is spreading to the financial analyses, research of interdependencies of certain metrics, business optimization, and company valuation as well [18]. Next to production and operations, finances and investments are business areas that most frequently use computational intelligence in economics [19].

These very tools enabled a contemporary approach to business valuation [20] and especially integration of financial [21], strategic [22], managerial [23] and other information of quantitative and qualitative nature [24] as the value's moving force, and using fuzzy logic [25] [26] [27], which further exposed the shortcomings of the standard valuation process and the necessity of implementing different methods and analytical techniques [28].

Researches concerning the implementation of these new tools and the modern approach to an integral valuation of companies in the Oil & Gas industry have not been published yet. Occasional studies deal with using the ANN in estimating oil and gas reserves [29], which also indicates the topicality of this research area.

A modern approach to O&G business valuation, proposed in this paper, can be expressed in three basic phases:

- search, gathering and classification of data (initial phase),
- determination of parameter interdependencies and their impact on business value by means of ANN (neuro-determination phase),
- development of fuzzy inference system (FIS) for business valuation of O&G companies (fuzzy-valuation phase).

The entire concept of this approach is also presented in Fig. 1.

The initial phase comprises of three stages. In the first stage, it is necessary to make a selection of all relevant

useful information sources and business databases that will be used to form an appropriate knowledge base. Given the complexity of the business within the O&G industry, the existing business databases can hardly meet all the appraisals' requirements independently. Thus, it is often necessary to use several business data bases simultaneously along with additional information sources, which makes the valuation process even more complicated and difficult. Regardless of this problem, in the second stage of this phase, it is necessary to search for and collect all information and data relevant to the process of valuation of the O&G Company. After the collection of those data is performed, in the third phase they are divided into two basic groups, namely, quantitative and qualitative.

It is also important to emphasize that it is necessary to implement and enable additional queries for the end user, tightly linked to business valuation, within the existing business databases. Specifically, this means that the user should have the option to choose the industry (e.g. O&G industry), the sector of the O&G industry (e.g. Integrated Majors), the company within the sector (e.g. ExxonMobil) and the procedure (e.g. business valuation). This would make searching and collecting the required data significantly faster and the procedure itself easier. The reason for this is obvious: data relevant to contemporary valuation are mainly located in several databases, and even if they are all placed within a single database, they are almost impossible to search using just one query due to the existing approach in the organization of business databases. Furthermore, if the business databases specialized for this area would be additionally expanded with appropriate non-financial data relevant for a certain industry, their efficiency would increase considerably. Such quantitative metrics are often "hidden" in annual reports; research studies conducted by audit and business advisory companies, institutes and classification societies; various relevant Internet sources etc., but are not taken into consideration in business databases.

The second phase of the modern approach to O&G business valuation, i.e. neuro-determination phase, also consists of three stages. In the first, quantification of qualitative parameters is to be performed in order to enable the analysis of the interdependencies of all the selected parameters, including the quantitative ones, as well as the analysis of their affect to the value of the company. According to this very affect, as one of the possible criteria, the selection of all relevant factors will be enabled in the third stage. This type of an analysis will not only determine the list of all significant quantitative and qualitative metrics, but will also quantify their significance in a view of the estimated value of the company. Moreover, thus determined list of all significant metrics can be used for defining user queries in the initial phase of searching business databases, as well.

For classifying parameters in this stage, classification neural networks (e.g. Kohonen's self-organizing map, probabilistic ANN, etc.) are recommended, while for determining interdependencies between parameters and the value of the company, approximation neural networks (e.g. regression ANN based on radial basis functions) are advised.

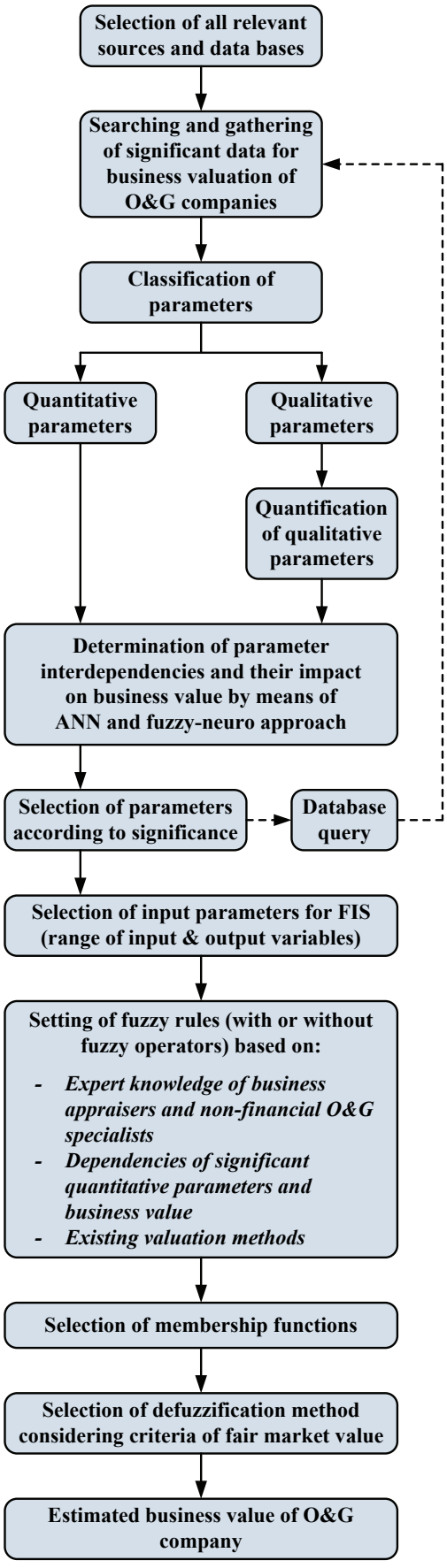


Figure 1. Flowchart of a modern approach to O&G business valuation

Fuzzy-valuation phase begins with the selection of input parameters for FIS and the fuzzification of their impact (range of input variables) to the value of the company (range of output variable). The greatest challenge is creating fuzzy rules (with or without fuzzy operators) that must be based upon:

- classical methodology of business valuation (all 3 approaches),
- determined interdependencies of business value and significant quantitative and qualitative parameters,
- expert knowledge of business appraisers and non-financial specialists from O&G industry.

Apart from selecting membership functions, it is of great importance to emphasize the significance of the selected defuzzification method that directly influences the final "fuzzy-value" of the company. The estimated value of the company determined by this contemporary neuro-fuzzy approach must tend to the fair market value, thus the defuzzification method must be selected accordingly.

IV. CONCLUSION

This paper presents the modern neuro-fuzzy approach to business valuation within the O&G industry. The authors examine the shortcomings of the existing methods of valuation within the industry of such complexity and suggest a contemporary model based on computer intelligence algorithms. In doing so, the emphasis is put on the application of the ANN in determining the interdependencies of significant quantitative and qualitative metrics, that create the value of the company, on the one hand, and on the development of FIS, by which the value of the company, which can objectively and reliably be considered fair market value, can be determined, on the other. The value of the O&G company estimated in this manner could be of great significance for the potential sellers and purchasers, business appraisals, investment bankers, as well as all other entities within the O&G sector.

REFERENCES

[1] Energy Outlook 2030. BP p.l.c., January 2013. [Online]. Available: www.bp.com/energyoutlook

[2] Global Oil and Gas Transactions Review 2012. Ernst&Young, EYGM Limited, 2013. [Online]. Available: http://www.ey.com/

[3] M.S. Raymond and W.L. Leffer, *Oil and Gas Production in Nontehnical Language*. Tulsa, OK: PennWell Corporation, 2006.

[4] C.J. Wright and R.A. Gallun, *Fundamentals of Oil & Gas Accounting*. 5th ed., Tulsa, OK: PennWell Corporation, 2008.

[5] D. Johnston and D. Johnston, *Oil Company Financial Analysis*, Tulsa, OK: PennWell Corporation, 2006.

[6] D. Cormier and M. Magnan, "Performance reporting by oil and gas firms: contractual and value implications," J. Int. Account. Auditing. Taxation., vol. 11, no. 2, pp. 131-153, 2002.

[7] S.A. Basher, A.A. Haug and P. Sadorsky, "Oil prices, exchange rates and emerging stock markets," Energ. Econ., vol. 34, no. 1, pp. 227-240, 2012.

[8] M.A. Adelman and G.C. Watkins, "U.S. oil and gas reserve prices, 1982-2003," Energ. Econ., vol. 27, no. 4, pp. 553-571, 2005.

[9] P. Osmundsen, F. Asche, B. Misund and K. Mohn, "Valuation of International Oil Companies," Energ. J., vol. 27., no. 3, pp. 49-64, 2006.

[10] B. Misund, F. Asche and P. Osmundsen, "Industry upheaval and valuation: Empirical evidence from the international oil and gas industry," Int. J. Account., vol. 43, no. 4, pp. 398-424, 2008.

[11] A.W. Howard and A.B. Harp, "Oil and Gas Company Valuations," Bus. Valuation Rev., vol. 28, no. 1, pp. 30-36, 2009.

[12] Oil&Gas Modeling - Enterprise Value and Key Metrics. Online Video Course. Breaking Into Wall Street, 2013. [Online]. Available: http://breakingintowallstreet.com/biws/oil-gas-modeling/

[13] Financial ratios and measures. Glossary and Terms. Damodaran Online, 2013. [Online]. Available: http://pages.stern.nyu.edu/~adamodar/New_Home_Page/definitions.html

[14] Tomorrow's Value Research 2012 - The annual assessment of corporate sustainability practices worldwide. Two Tomorrows Group Ltd, 2012. [Online]. Available: http://www.twotomorrows.com/ media/pdf/TVR_2012_PDF_final2.pdf

[15] G. Zhang, M.Y. Hu, B.E. Patuwo and D.C. Indro, "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis," Eur. J. Oper. Res., vol. 116, no. 1, pp. 16-32, 1999.

[16] Z.R. Yang, M.B. Platt and H.D. Platt, "Probabilistic Neural Networks in Bankruptcy Prediction - A New Tool for Predicting Thrift Failures," J. Bus. Res., vol. 44, no. 2, pp. 67-74, 1999.

[17] M. Zekić-Sušac, "Neural Networks in Stock Market Predictions," University of Zagreb, Faculty of organization and informatics, Varaždin, Croatia, 1999. (in Croatian)

[18] K.A. Smith and J.N.D. Gupta, *Neural Networks in Business: Techniques and Applications*, London, UK: Idea Group Publishing, 2002.

[19] B.K. Wong and V.S. Lai, "A survey of the application of fuzzy set theory in production and operations management: 1998-2009," Int. J. Prod. Econ., vol. 129, no. 1, pp. 157-168, 2011.

[20] J. Seng and J.T. Lai, "An Inteligent information segmentation approach to extract financial data for business valuation," Expert Syst. Appl., vol. 37, no. 9, pp. 6515-6530, 2010.

[21] P. Sevastjanov, L. Dimova and D. Sevastianov, "Fuzzy Capital Budgeting: Investment Project Evaluation and Optimization," in *Fuzzy Applications in Industrial Engineering*, vol. 201, C. Kahraman, Ed., London, UK: Springer, 2006, pp. 205-228.

[22] R.T. McIvor, A.G. McCloskey, P.K. Humphreys and L.P. Maguire, "Using a fuzzy approach to support financial analysis in the corporate aquisition process," Expert Syst. Appl., vol. 27, no. 4, pp. 533-547, 2004.

[23] J. Chen and S. Lin, "An interactive neural network-based approach for solving multiple criteria decision-making problems," Decis. Support Syst., vol. 36, no. 2, pp. 137-146, 2003.

[24] E.E. Karsak and E. Tolga, "Fuzzy multi-criteria decision-making procedure for evaluating advanced manufacturing system investments," Int. J. Prod. Econ., vol. 69, no. 1, pp. 49-64, 2001.

[25] S. Malagoli, G. Mastroleo and C.A. Magni, "The use of fuzzy logic and expert systems for rating and pricing firms: a new perspective on valuation," Managerial Financ., vol. 33, no. 11, pp. 836-852, 2007.

[26] J.S. Yao, M.S. Chen and H.W. Lin, "Valuation by using a fuzzy discounted cash flow model," Expert Syst. Appl., vol. 28, no. 2, pp. 209-222, 2005.

[27] A.M. Gil-Lafuente, C. Castillo-Lopez and F.R. Blanco-Mesa, "A Paradigm Shift in Business Valuation Process Using Fuzzy Logic," in *Soft Computing in Management and Business Economics*, vol. 2, A.M. Gil-Lafuente, J. Gil-Lafuente and J.M. Merigo-Lindahl, Ed., London, UK: Springer, 2012, pp. 177-189.

[28] D. Tillman Cassone, "A process to estimate the value of a company based on operational performance metrics," Ph.D. dissertation, Dept. of Ind. and Manuf. Syst. Eng., Kansan State Univ., Manhattan, KS, 2005.

[29] Y.H. Pan and H. He, "Evaluation Techniques for Oil Gas Reservoir Based on Artificial Neural Networks Techniques," in Proc. of the 3rd Int. Conf. on Bus. Intell. and Financ. Eng., BIFE 2010, Hong Kong, SAR of China, August 13-15, 2010, pp. 28-31.

Komparativna analiza hrvatskih online novina koje se plaćaju

mr.sc. Olivia-Silvana Prlić*, Andriana Lacković, bacc.ing.tech.inf. ** i Filip Lončar, bacc.ing.tech.inf. ***
* Tehničko veleučilište u Zagrebu/Informatika, Zagreb, Hrvatska
** Tehničko veleučilište u Zagrebu/Informatika, Zagreb, Hrvatska - studentica
*** Nacionalna i sveučilišna knjižnica u Zagrebu / IT odjel, Zagreb. Hrvatska

Sažetak - Donedavno su papirnate novine bile jedini oblik novina, ali razvoj tehnologija i dostupnost Interneta omogućio je nastanak Internet portala s novostima i online novina. Danas su, zbog svojih mnogobrojnih prednosti pred papirnatim novinama, Internet portali s novostima i online novine toliko dobro prihvaćeni da su počeli preuzimati tržište. Kao što je pokazalo naše prethodno istraživanje o čitanosti različitih vrsta novina [1], u Hrvatskoj je najčitanija vrsta novina besplatni Internet portali s novostima koji se financiraju samo od reklama, a najslabije čitana vrsta novina su online novine koje se plaćaju. Upravo zbog toga je odlučeno više se posvetiti ovoj temi i analizirati online novine koje se plaćaju da se sazna koje su njihove najveće prednosti i mane te pronađe što bi kod njih trebalo poboljšati kako bi se povećala njihova čitanost.

I. UVOD

Općeniti trend papirnatih novina u svijetu je izrazito negativan. Broj čitatelja papirnatih novina u stalnom je padu te je posljedica sveopće nagađanje koliko će papirnate novine kao novine još uopće postojati. Svoje čitatelje papirnate novine danas većinom broje među starijim generacijama, dok se one mlađe gotovo isključivo orijentiraju na online novine kao izvore za zadovoljavanje svojih interesa. Najvjerojatniji razlog tome je što starije generacije čitatelja nisu u dovoljnoj mjeri upoznate i/ili zainteresirane za mogućnosti koje nude moderne tehnologije. Također, u obzir se uzima i činjenica da postoji gotovo cjeloživotna navika starijih generacija čitatelja koja se sastoji od listanja papirnatih (opipljivih) novina, a za tako nešto se ne očekuje da će se olako iskorijeniti. Upravo je to razlog zašto će papirnate novine i dalje postojati, ali će s vremenom imati sve manji i manji broj čitatelja. Realno je i očekivanje da će se sve novine koje u budućnosti žele opstati prebaciti barem jednim dijelom online, a s vremenom i potpuno. Jedan od najvećih razloga zašto će se to dogoditi jest da papirnate novine sve teže prate konkurentne online novine i prednosti kojima se one odlikuju. Dinamičnost i brzina kojom online novine uređuju i objavljuju sadržaj je prednost kakvu papirnate novine ne mogu nadoknaditi ni na koji način. Uz to, online novine nude nevjerojatnu mogućnost interaktivnosti putem poveznica na izvore ili srodne članke, komentiranje vijesti ili njihovo jednostavno prosljeđivanje drugim čitateljima. Online novine u svom članku mogu imati multimedijalni sadržaj poput slika, snimke zvuka ili videa koji povećavaju doživljaj vijesti i nude bolju pozadinu priče. Sve ove prednosti rezultirat će time da će online novine imati sve veći i veći broj čitatelja i to gotovo svakim idućim danom. Sve je češća praksa da se u papirnatim novinama čitatelje navodi da više informacija o temi nekog članka potraži upravo na online izdanju istih novina.

Glavni dijelovi ovog rada su:

a) analiza online novina koje se plaćaju u Hrvatskoj,

b) koliko su online novina koje se plaćaju interaktivne i povezane s drugim sadržajima,

c) po čemu i koliko su online novina koje se plaćaju kvalitetnije od besplatnih Internet portala s novostima i

d) komparativna analiza hrvatskih online novina koje se plaćaju i najčitanijih online novina u svijetu.

II. RAZLIKA IZMEĐU PAPIRNATIH I ONLINE NOVINA KOJE SE PLAĆAJU TE BESPLATNIH INTERNET PORTALA S NOVOSTIMA

Gotovo sve papirnate novine koje se plaćaju imaju i svoja online izdanja. S obzirom na spomenuti negativni trend i sve posljedice koje se njime stvaraju za papirnate novine, to je sasvim očekivana pojava. Međutim, veliki je paradoks da je većina online izdanja papirnatih novina besplatno, iako se papirnata izdanja novina plaćaju. U obzir treba uzeti činjenicu da se većina online novina financira gotovo isključivo preko reklamnog sadržaja ukomponiranog u njihov sastav. U razvijanju je novi trend koji nudi alternativnu mogućnost financiranja za online izdanja novina – plaćanje i pretplata na premium sadržaj. Ovaj princip novina osmišljen je tako da korisnicima nudi mogućnost plaćanja za pristup opsežnijem, ekskluzivnijem i kvalitetnijem sadržaju koji nije dostupan korisnicima besplatnog dijela online novina. Navedeni sadržaj se ističe reportažama s mjesta događaja, opširnijim analizama i intervjuima koji omogućuju premium korisnicima "stepenicu više" u kvaliteti i doživljaju prilikom informiranja putem online novina[1]. Večernji list su prve hrvatske online novine koje su spomenuti trend provele u djelo te nude svim korisnicima pristup premium sadržaju uz plaćanje, dok svim pretplatnicima na tiskano izdanje novina nude premium sadržaj besplatno (odnosno sve do kraja trajanja njihove pretplate). Na svjetskoj razini je ova praksa mnogo razvijenija te novine svojim čitateljima konstantno nude bogat premium sadržaj, dok se kod nas ta praksa još uvijek uhodava i razvija.

Unatoč tome, anketom provedenom u sklopu našeg prijašnjeg rada dođeno je do rezultata koji pokazuju kako su online novine koje se plaćaju najnepopularniji oblik novina s jedva 2% čitatelja koji ih preferiraju. S druge strane, papirnate novine koje se plaćaju imaju gotovo četiri puta više poklonika koji sveukupno čine 8% anketiranih. Besplatne papirnate novine čita 25% anketiranih, dok čak 73% anketiranih najradije čita novosti na Internet portalima i druge oblike online novina koje se ne plaćaju [1].

Iz toga se vidi da se većina čitatelja u Hrvatskoj još uvijek nerado odlučuje plaćati za online novine (unatoč svim prednostima koje nude naspram papirnatih). Najvjerojatniji razlog tome je taj što je većini čitatelja pojam Interneta i njegovog sadržaja povezan s njegovim slobodnim dijeljenjem i potpunoj dostupnosti. Također, s obzirom na jednostavnost i dinamiku kojom se vijesti preko Interneta šire, većina korisnika ne osjeća potrebu plaćati online novine, kada znaju da do istih informacija mogu doći besplatno surfajući Internetom i istraživanjem brojnih alternativnih izvora. Upravo je to ključni razlog zašto se papirnate novine još uvijek bolje "prodaju" od online novina. Dodatni subjektivni razlog koji se također ne smije izostaviti je da ljudi preferiraju plaćati opipljiv proizvod od onog virtualnog jer to neupitno daje osjećaj da su zaista dobili ono što su platili. Online izdanja novina su kod nas još uvijek u fazi ranog razvoja te najčešće zapravo predstavljaju online izdanje gotovo identičnog sadržaja kao i papirnata izdanja istih novina. To je još jedan razlog zašto plaćeni tip online novina kod nas trenutno nije posebno popularan, iako se očekuje da će svojim razvitkom online novine to u budućnosti promijeniti.

III. KARAKTERISTIKE IDEALNIH ONLINE NOVINA KOJE SE PLAĆAJU

Prema McIntyreovom članku „Rating The Top 25 Newspaper Websites" i prethodnom iskustvu autora ovog članka najbitnije karakteristike online novina koje se plaćaju su lakoća pronalaženja pomoću Internet tražilica, lakoća navigacije, kvaliteta sadržaja, redovitost ažuriranja, količina reklama, povezanost s drugim sadržajima, interaktivnost i cijena [2].

A. Lakoća pronalaženja online novina pomoću Internet tražilica

Jedan od vrlo važnih faktora koji bi trebale imati dobro odrađen sve online novine koje se plaćaju je lakoća pronalaženja pomoću Internet tražilica, odnosno SEO (engl. search engine optimisation). Da bi neka Internet stranica uopće imala posjetitelje, oni je trebaju moći što lakše pronaći pomoću Internet tražilica [3].

B. Lakoća navigacije

Posjetitelji svih Internet stranica pa tako i online novina žele što prije naći tražene informacije zbog čega je vrlo važno dobro razraditi način navigacije. Ako neka stranica ima kompliciranu navigaciju i posjetitelju je teško doći do određene informacije, vrlo brzo će odustati, otići sa stranice i početi tražiti neku drugu stranicu gdje će pronaći željenu informaciju. Zbog toga je vrlo važno da online novine koje sadrže mnogo različitih informacija imaju jako dobro organiziranu navigaciju [4].

C. Kvaliteta sadržaja

Kvalitetne online novine koje se plaćaju bi trebale svojim sadržajem privlačiti pažnju čitatelja i što više koristiti dodatne mogućnosti koje nudi Internet [2].

Kvalitetni članci u online novinama su oni koji nisu prerađene verzije novosti preuzetih s nekog drugog mjesta, koji nude više informacija nego besplatni Internet portali s novostima i koji nude vijesti prije konkurencije ili na drugi čitateljima primamljiv način. Čitateljima je također vrlo važno i da su novine vjerodostojne te da u njima nema pogrešaka [5].

D. Redovitost ažuriranja

Ovo je jedna od najvećih prednosti online novina pred papirnatim koje svoj sadržaj ne mogu mijenjati i nadopunjavati u realnom vremenu. Vrlo je važno da i online novine koje se plaćaju, poput Internet portala s novostima objavljuju vijesti tijekom cijelog dana te ih objavljuju trenutno. Tako su mnogo aktualnije od papirnatih novina koje se najčešće na dan distribucije zapravo sastoje od jučerašnjih novosti. Unatoč tome, i dalje ima online novina koje se plaćaju koje ne iskorištavaju tu mogućnost i samo su elektroničko izdanje papirnatih novina [1].

E. Količina reklama

Besplatni Internet portali s novostima imaju mnogo reklama jer upravo to je njihov jedini izvor financiranja te bi jedna od razlika između njih i online novina koje se plaćaju trebala biti upravo količina reklama. Danas se prenatrpanost oglasima navodi kao glavna mana novina [6].

F. Povezanost s drugim sadržajima

Online novine na jednostavan način mogu referencirati sadržaj svog članka na njegov izvor ili slične članke. Tako čitatelji mogu nakon čitanja članka nastaviti čitati o istoj ili sličnoj temi bez poteškoća koje imaju čitatelji papirnatih novina koji se jedino mogu osloniti na to da će autor članka u sadržaju spomenuti neki dodatni izvor ili ga na neki način priložiti [1].

G. Interaktivnost

Pod interaktivnost kod online novina podrazumijeva se mogućnost komentiranja, dobivanja povratnih informacija od samog autora, dobre povezanosti s drugim sličnim člancima i izvorima, mogućnost dijeljenja sadržaja putem socijalnih mreža, prisutnost videa i mogućnost glasanja u anketama.

Kada korisnici imaju veću kontrolu nad nekim iskustvom mogu se više uživjeti u njega. Zbog toga je jako važno što sve korisnici mogu raditi sa sadržajem. Korisnici radije aktivno traže otkriće nego da budu samo pasivno informirani [7].

H. Cijena

Troškovi pokretanja, izdavanja i održavanja papirnatih novina su osjetno veći nego za online izdanja. Papirnate novine se moraju bazirati na određenoj cijeni

kako bi mogle pokriti navedene troškove. Automatska posljedica toga je da su papirnate novine skuplje.

Dok se Internet portali s novostima najčešće financiraju prihodima od reklama te su besplatni za čitatelje, online novine koje se plaćaju imaju svoju cijenu zbog čega moraju biti jače u drugim navedenim faktorima, posebno zato što ima osoba kojima mnogo znači cijena novina [1].

IV. DRUGI IZVORI

Najčešći razlog kojim urednici opravdavaju odluku o naplaćivanju pristupa sadržaju online novina jest da se navedeni sadržaj smatra predragocjenim da bi ga se dalo besplatno. Većina svjetskih online novina je prije desetak godina krenula s potpuno besplatnim sadržajima na svojim web stranicama. S vremenom su neke od online novina krenule naplaćivati određene usluge i dijelove svog sadržaja poput objavljivanja osmrtnica i pristupa arhivskim materijalima. Tako se postepeno razvio model pretplate na sadržaj kojeg mnoge svjetske online novine primjenjuju u svakodnevnom poslovanju.

Prema članku L. Indvik, kako bi takav način naplaćivanja mogao funkcionirati, osmišljen je sistem kojim online novine sprječavaju dostupnost svom sadržaju svima osim plaćenim korisnicima. Navedeni sistem zove se "paywall" i dijeli se na 3 kategorije, ovisno o razini striktnosti kojom se primjenjuje. "Hard paywall" je najrigoroznija kategorija koja omogućuje ili minimalan ili nikakav pristup sadržaju onima koji nisu pretplatnici. Druga kategorija, "soft paywall" primjenjuje fleksibilniju metodu kojom omogućuje djelomičan pristup sadržaju ili određenoj količini članaka prije nego počne inzistirati na pretplati. Treća kategorija je kombinacija dvije navedene kategorije te ona kao takva dozvoljava besplatan pristup određenom sadržaju dok premium sadržaj čuva samo za pretplatnike. Prošle godine u SAD-u je gotovo 20% online novina nekom od metoda "paywalla" zahtijevala plaćanje za pristup svom sadržaju što je dvostruko više nego prije dvije godine [8]. Tako online novine povećavaju svoje prihode, ali i na određen način i dalje potiču određen postotak populacije na kupnju papirnatih izdanja novina (pogotovo onaj dio populacije koji smatra da već dovoljno plaća za pristup internetu, a kamoli za sadržaj novina na njemu).

Međutim, prema medijskom stručnjaku B. Mitchellu sa Poynter instituta, jedini način da se "paywallom" dođe do stalnog prihoda jest da online novine stvore neku novu vrijednost u vidu više kvalitete ili inovativnosti u online sadržaju kojeg su odlučile naplaćivati [9].

Uvođenjem plaćanja pristupa sadržaju, online novine su u početku gubile velik broj svojih čitatelja. Primjerice, novine Daily Gazette su u kolovozu 2009. godine zabilježile gubitak od gotovo 40% čitatelja koji su nisu bili spremni odjednom početi plaćati za sadržaj koji je do tada bio besplatan [10]. Ipak, s vremenom je broj pretplatnika počeo stabilno rasti, ponajviše zbog činjenice da su online novine mnogo poradile na kvaliteti i prezentaciji ponuđenog sadržaja. Kao rezultat toga, danas postoje novine koje ubiru značajan prihod od svojih pretplatnika. New York Times je prošle godine brojao oko 325 tisuća pretplatnika što je tim novinama donosilo prihod od gotovo 80 milijuna dolara[11].

Online izdanje New York Timesa dnevno broji preko 12 milijuna posjetitelja (odnosno čitatelja) te prema provedenim istraživanjima [2] slove kao najbolje online novine u SAD-u. Navedeno istraživanje kao ključne razloge tog uspjeha navodi maksimalno iskorištavanje svih prednosti modernih tehnologija koje stoje na raspolaganju online novinama poput visoko kvalitetnih "slideshowova" koji reproduciraju slike i video sadržaje nedostupne čitateljima papirnatih izdanja novina ili pak mogućnost komentiranja svakog članka te sudjelovanja u online anketama. Velik broj novinara zadužen je za pisanje, objavljivanje i uređivanje stotina vijesti i blogova dnevno što New York Timesu osigurava tako velik broj čitatelja kao i garanciju da će ti isti čitatelji navraćati na stranicu nekoliko puta dnevno. Iako određen dio sadržaja također postoji i u papirnatom izdanju novina, značajan dio sadržaja ipak je namijenjen isključivo online izdanju New York Timesa. Ta prednost, kao i sve ostale navedene, čini online izdanje ovih novina neupitno boljim i kvalitetnijim od svog papirnatog izdanja. Svojim poslovanjem i dobrim praksama ove novine mogu poslužiti kao primjer gotovo svim online novinama, a pogotovo onima koje su tek u ranoj fazi implementacije [2].

V. KOMPARATIVNA ANALIZA ONLINE NOVINA KOJE SE PLAĆAJU

U ovom poglavlju analiziraju se, prema navedenim kriterijima, tri različite hrvatske online novine koje se plaćaju, a to su 24 sata, Jutarnji list i Večernji list. U zadnjem dijelu ovog poglavlja karakteristike hrvatskih novina ukratko su uspoređene s najvažnijim karakteristikama najposjećenijih svjetski poznatih online novina Daily Mail i New York Times [12]. Odabrane su upravo ove novine zbog toga što su u doba kada su papirnate novine bile najčitaniji oblik novina bili najprodavanije novine u Hrvatskoj [13].

A. Lakoća pronalaženja online novina pomoću Internet tražilica

Niti jedne od uspoređivanih hrvatskih online novina koje se plaćaju nisu imale dobar rang na Internet tražilici Google (Google je uzet za relevantnu tražilicu preko koje ćemo uspoređivati online novine zbog toga što je to najkorištenija Internet tražilica [14]) i jedino kako ih se moglo naći je preko besplatnih Internet portala s novostima istih novina.

Kada se dođe na Internet portal 24 sata, poveznica na online novine koje se plaćaju se može pronaći skroz pri dnu stranice što nije dobro za lakoću pronalaženja.

Kod Jutarnjeg lista ih je bilo još teže pronaći, na Internet portalu ne postoji vidljiva poveznica na online novine koje se plaćaju, pronađene su tek nakon traženja preko Google-a riječi „Jutarnji list pretplata".

Na Internet portalu Večernjeg lista pri samom vrhu stranice postoji poveznica pod nazivom „E-izdanje", ali nije istaknuto pa je također otežano pronalaženje. Unatoč tom nedostatku, u ovom području je vodstvu pred 24 sata i Jutarnjim listom. Večernji list također ima i na Internet

portalu neke sadržaje koji su označeni s premium oznakom koji se u potpunosti mogu pročitati nakon plaćanja sadržaja. Ove sadržaje je vrlo lako naći i neki od njih su istaknuti već na prvoj stranici te je po tome preduhitrio konkurenciju.

Zbog navedenog u ovom poglavlju se uspoređuje lakoća pronalaženja preko tražilica Internet portala uspoređivanih novina.

Prema alatu Google keywords, daleko najčešće tražene ključne riječi povezane uz novine su „novine“ (450.000 upita mjesečno) i „vijesti“ (823.000 upita mjesečno). Prilikom traženja riječi „novine“ najbolje stoji Jutarnji list koji je drugi po redu na prvoj stranici rezultata dok 24 sata i Vjesnik nisu na prve tri stranice. Za ključnu riječ „Vijesti“ najbolje stoji Večernji list koji je prvi po redu na prvoj stranici rezultata, ali to nije zbog dobre optimizacije za tražilice nego zbog plaćenog oglašavanja na Google-u. Jutarnji list je za ovu ključnu riječ četvrti, a 24 sata peti po redu na prvoj stranici rezultata.

S obzirom na to da su sve istraživane novine dobro razvijeni proizvodi, treba uzeti u obzir i rezultate pretrage točno ciljanih novina. Najčešće tražene novine od istraživanih su 24 sata sa 4.090.000 traženja mjesečno, nakon njih slijedi Jutarnji list sa 3.413.000 traženja i na kraju Večernji list sa 2.280.000 traženja.

New York Times i Daily Mail vrlo poznate novine pa je lakoća njihova pronalaženja na Googleu promatrana s obzirom na njihova imena kao ključne riječi. New York Times ima 9.140.000 traženja mjesečno, a Daily Mail 20.400.000.

Unatoč tome, ako se gleda ranking svih web stranica na Internetu, na malo boljoj poziciji je New York Times s Alexa Traffic Rankom 125, dok Daily Mail ima Alexa Traffic Rank 129. Alexa Traffic Rank broj pokazuje koja po redu je je određena web stranica prema posjećenosti [15].

Prema spomenutoj stranici za online ranking Alexa, od hrvatskih online novina na najvišoj poziciji prema posjećenosti je Jutarnji list (2.992), zatim slijede 24 sata (3.667) i Večernji list (4.756) [15].

B. Lakoća navigacije

Sve hrvatske online novine koje se plaćaju su samo papirnate novine koje se mogu kupiti na kioscima prenesene u elektronički oblik s dodanim efektom da izgledaju kao da se listaju. Zbog toga je kod svih isti način navigacije, listanje kao i kod papirnatog oblika pa je mnogo teže snaći se u plaćenim online novinama nego na besplatnim portalima s novostima.

Večernji list i 24 sata koriste posve istu tehnologiju za online novine koje se plaćaju, koriste online aplikaciju Newspaper Direct i imaju malo više mogućnosti nego Jutarnji list. Na slici 2. se vidi da se u aplikaciji Newspaper Direct kod klika na određeni članak otvara uvećani tekst članka (nije samo slika papirnatih novina koja se može uvećati kao kod Jutarnjeg lista) što olakšava čitanje i snalaženje.

Jedino se Večernji list odmaknuo malo više od ostalih sa svojim premium sadržajem (uz navedeni oblik papirnatih novina prenesenih na Internet). Navigacija

Slika 1: članak objavljen online preko aplikacije Newspapers Direct

premium sadržaja je kao na besplatnim portalima s novostima pa je i u ovoj karakteristici bolji od 24 sata i Jutarnjeg lista.

Za razliku od hrvatskih online novina koje se plaćaju, Daily Mail i New York Times imaju mnogo jednostavniju navigaciju i poput Internet portala s novostima imaju sadržaj podijeljen na glavne teme, koriste poveznice na druge slične sadržaje i u samo nekoliko klikova mišem može se doći do tražene teme bez potrebe za virtualnim „listanjem“.

C. Kvaliteta sadržaja

U online novinama 24 sata koje se plaćaju postoje neke vijesti koje su opširnije u verziji koja se plaća ili ih uopće nema na Internet portalu, ali postoje i obrnuti slučajevi. Vijesti u novinama koje se plaćaju često su zapravo vijesti koje su dan prije objavljene na besplatnom Internet portalu s novostima, ponekad malo nadopunjene. Internet portal je bolji i po tome što kod njega postoji mogućnost komentiranja. Zbog svega navedenog nema puno razloga za kupovanje online novina koje se plaćaju.

Kod Jutarnjeg lista u online novinama koje se plaćaju ima članaka koji nisu objavljeni na Internet portalu s novostima, a ako i jesu, objavljuju ih kasnije nego što se mogu pročitati u verziji koja se plaća. Zbog toga se može zaključiti da funkcioniraju bolje od 24 sata.

Večernji list je po karakteristikama sadržaja koji se plaća i besplatnog sadržaja poput 24 sata, nema nekog pravila, ponekad su ranije objavljeni ili opširniji članci na besplatnom Internet portalu, a ponekad u online novinama koje se plaćaju.

Prema ranije navedenim kriterijima, Daily Mail nije usporediv s hrvatskim online novinama koje se plaćaju jer ima samo besplatno online izdanje.

Za razliku od njega, online izdanje New York Timesa ima besplatne sadržaje kao i one koji se plaćaju. Online sadržaji New York Timesa u potpunosti iskorištavaju sve mogućnosti online novina. Stalno objavljuju nove članke koje ažuriraju i po nekoliko puta dnevno. Veći dio sadržaja objavljuje se samo u online izdanju koje se plaća. Prema promatranim kriterijima, ove novine su najbolje kvalitete.

D. Redovitost ažuriranja

Kao što je već napisano u poglavlju o lakoći snalaženja, budući da su sve hrvatske online novine samo

verzije papirnatih, ažuriraju se samo jednom dnevno i ne iskorištavaju prednosti koje im pruža Internet.

Ponovno s izuzetkom premium sadržaja Večernjeg lista kao i New York Timesa te Daily Maila koje se ažuriraju više puta dnevno, kako dolaze nove vijesti.

E. Količina reklama

Količina reklama kod hrvatskih online novina računata je kao udio prostora u novinama koji zauzimaju reklame. Na primjer, ako reklama zauzima cijelu stranicu, zbroju stranica s reklamama pribrojen je broj jedan, ako iduća zauzima pola stranice zbroju je pribrojeno 0,5 i tako dalje. Dobiven broj je podijeljen s ukupnim brojem stranica u novinama i izračunat je postotak prostora koji zauzimaju reklame.

Najviše reklama imale su online novine 24 sata, otprilike oko 25% sadržaja s time da su kod njih sve reklame bile puno veće i 12% novina su bile reklame preko cijele stranice.

Večernji list i Jutarnji list su imali podjednaku količinu reklama, manju od 20% sadržaja, a velik broj reklama bile su najave sadržaja iz budućih izdanja.

Kod premium sadržaja Večernjeg lista reklame su se nalazile u većem dijelu užeg, desnog stupca. Suprotno tome Daily Mail i New York Times su imali samo po jednu reklamu uz svaki članak. Reklame su također bile smještene u desnom stupcu web stranice.

F. Povezanost s drugim sadržajima i interaktivnost

S obzirom na to da su Jutarnji list online novine koje se plaćaju samo papirnate novine u elektroničkom obliku, nisu povezane s drugim sadržajima pomoću poveznica niti interaktivne.

24 sata i Večernji list za online novine koje se plaćaju koriste Newspaper Direct. Interaktivnije su i imaju mogućnosti dijeljenja sadržaja preko društvenih mreža, komentiranja, mijenjanja veličine slova, prijevoda na engleski, a Večernji list ima i poveznice koje vode na druge članke slične tematike ranije objavljene u Večernjem listu (nema vanjskih poveznica).

Korak naprijed od svega prethodno spomenutog napravio je Večernji list, ali ne s online novinama koje se plaćaju nego premium sadržajima na Internet portalu. Premium sadržaji sadrže poveznice na druge slične članke, povezani su s Facebook-om, Twitter-om i Google plus-om, imaju mogućnost komentiranja, a neki od njih imaju i ankete vezane uz teme u kojima čitatelji mogu glasati kakvo je njihovo mišljenje o temi. Dodatna pogodnost je i biranje veličine slova tako da osobe koje slabije vide mogu koristiti veće povećanje. Premium članci su i personalizacija novina jer čitatelji ne moraju kupiti cijele novine nego biraju i kupuju samo teme koje ih zanimaju.

Daily Mail i New York Times imaju dobro povezane sadržaje s drugim sadržajima kao i interaktivnost. Imaju iste mogućnosti kao i upravo opisani premium sadržaju Večernjeg lista (izuzev biranja veličine slova).

G. Cijena

Sve hrvatske online novine nude opciju besplatne probe prvih tjedan dana, a Večernji premium nudi besplatno dva premium članka. New York Times ima ponudu prvih mjesec dana za 0,99$ (5,82kn). Daily Mail online sadržaje nudi besplatno pa ga nema u priloženoj tablici 1.

U tablici 1 prikazan je odnos cijena različitih vrsta online novina (cijene od ožujka 2013.).

Iz tablice 1 je vidljivo da je Večernji list premium najpovoljniji, ali trenutno ne nudi mnogo više članaka od besplatnog dijela Internet portala pa zapravo nije toliko mnogo povoljniji.

H. Posebne ponude

24 sata pretplatnicima svojih online izdanja nude besplatne priloge (izlaze povremeno, rjeđe od jednom tjedno) poput Cafea 24, Super Mile i Budi.in.

Jutarnji list nema dodatnih posebnih ponuda.

Večernji list pretplatnicima na online izdanja nudi besplatne priloge (ne izlaze svaki dan) Ekran, V magazin, City, Obzor, Auto +, Kvadrat po mjeri i Biznis +. Pretplatnici na tiskano izdanje imaju besplatan pristup svim premium sadržajima na Internet portalu.

New York Times nudi 50% popusta za pristup online sadržajima svima koji su pretplaćeni na tiskana izdanja New York Timesa ili International Herald Tribunea, studentima i zaposlenicima obrazovnih ustanova. Na upit su mogući grupni popusti za tvrtke.

VI. BUDUĆNOST ONLINE NOVINA KOJE SE PLAĆAJU

Online izdanje novina na jedan način predstavlja spas od propasti, ali nema garancije da će proizvod koji gubi na papiru odjednom uspjeti u online izdanju. Nema konsenzusa o tome kako postići balans između tiska i online izdanja. U zemljama u kojima online izdanja i tiskovna izdanja čitavo desetljeće postoje jedno pored drugog počinje vrijeme promjena. Broj izdanja dnevnih novina sve više opada, dnevni tisak nestaje i analitičari sugeriraju izdavačima tiskanih izdanja da se okreću analitičkim ili istraživačkim tekstovima. Izdavačima hrvatskih online izdanja je jasno da ne mogu preživljavati samo na temelju prihoda od oglašavanja već moraju pronaći svoj vlastiti model ponude i kupnje informacija. Oni moraju preispitati srž svojeg poslovanja i mijenjati sve elemente svojeg marketinga od proizvoda kojeg nude, metoda određivanja cijena, distribucija putem novih platformi, marketinške komunikacije, sve do upravljanja odnosima s korisnicima u cilju izgradnje lojalnosti. Online vijesti je teže prodati, ali korisnicima nudi mnoge pogodnosti ne bi li ih se pridobilo. Može se pretpostaviti

TABLICA I. ODNOS CIJENA RAZLIČITIH VRSTA ONLINE NOVINA U KUNAMA

	24 sata	**Jutarnji list**	**Večernji list**	**Večernji premium**	**New York Times**
proba	0 kn - 1 tjedan	0 kn - 1 tjedan	0 kn - 1 tjedan	0 kn - 2 članka	5,82 kn - 1 mjesec
1 dan	2,42	-	4,66	-	-
1 tj.	17,22	-	25,86	-	-
1 mj.	69,07	100	93,08	45,00	88,20
3 mj.	172,76	230	227,47	65,00	-
6 mj.	-	360	-	130,00	-
12 mj.	633,60	470	806,53	260,00	-

da će biti održiv onaj model koji čitateljima ponudi dvije stvari: jednostavnost korištenja i percepciju dobre vrijednosti za novac. U ovom trenutku Večernji list bilježi značajan rast prihoda od online poslovanja. Ono što je ključno svakako je da se razvije zakonska regulativa koja će nakladnike i same autore zaštiti od krađe informacija i neovlaštenog objavljivanja.

VII. ZAKLJUČAK

Provedena komparativna analiza online novina koje se plaćaju pokazala je da je ovaj oblik novina u Hrvatskoj još uvijek u fazi razvoja i ima dosta prostora za poboljšanja.

Prema kriterijima ove analize, premium sadržaji na Internet portalu Večernjeg lista su najbliži onome kako u svijetu izgleda ovaj oblik novina te imaju mnogo prednosti pred konkurencijom (tablica 2).

Najveće mane većine hrvatskih online novina koje se plaćaju su statičnost (većina ih je samo preslika papirnatih novina s dodanim efektom listanja), neiskorištavanje mogućnosti interaktivnosti koju pruža objavljivanje na Internetu, teško pronalaženje pomoću Internet tražilica kao i na stranicama pojedinih novina, nedovoljno redovito ažuriranje sadržaja te prevelika količina reklama.

Zbog toga bi, s ciljem poboljšanja prodaje online novina koje se plaćaju trebalo povećati njihovu interaktivnost, poraditi na marketingu i upoznavanju publike s online novinama i njihovim prednostima, poboljšati njihovu optimizaciju za Internet tražilice (SEO), ponuditi čitateljima sadržaje koje ne mogu dobiti na besplatnim Internet portalima, personalizirati novine te smanjiti količinu reklama.

Hrvatske online novine koje se plaćaju su se pojavile na tržištu relativno nedavno i imaju mnogo prostora za razvoj i napredak. Iako publika trenutno nije jako zainteresirana za njih, u budućnosti će se to sigurno promijeniti jer će razvojem tehnologija i primjenom najboljih praksi iz svijeta hrvatske online novine postati neusporedivo bolje od svojih papirnatih verzija.

LITERATURA

[1] O.S. Prlić, A. Lacković, F. Lončar, "Papirnate novine i e-novine u 21. stoljeću", Tiskarstvo i dizajn, 2013

[2] D.A. McIntyre, "The Newspapers: Rating The Top 25 Newspaper Websites," 24/7 Wall St, http://247wallst.com/2008/06/03/the-newspapers/, 03.06.2008.

[3] S. Lazarić, "Optimizacija web stranica za tražilice," neobjavljen

[4] J. Clark, „Constructing Accessible Web Sites," New Riders, 2002

[5] T. Jefferson, E. Wager, F. Davidoff, "Measuring the Quality of Editorial Peer Review," in JAMA, vol. 287(21), 2002, pp. 2786-2790.

[6] B. Krce Miočić, Lj. Zekanović-Korona, M. Mirković, „Oglašavanje u tiskovinama jučer, danas, sutra," Kraljski Dalmatin – 200 godina zadarskog i hrvatskog novinarstva u europskom kontekstu, 2007

[7] K. Kenney, A. Gorelik, S. Mwangi, „Interactive features of online newspapers," First Monday, vol. 5(1), 2000

[8] L. Indvik, „20% of Newspapers Now Have Online Paywalls," Mashable, http://mashable.com/2012/08/28/us-newspapers-paywalls/, 28.08.2012.

[9] B. Mitchell, „It's time: 5 reasons to put up a metered paywall," Poynter, http://www.poynter.org/latest-news/business-news/newspay/149953/its-time-5-reasons-for-taking-the-plunge-into-a-metered-paywall/, 20.10.2011.

[10] J. Tartakoff, „Taking The Plunge: How Newspaper Sites That Charge Are Faring," Paid Content, http://paidcontent.org/2009/09/02/419-taking-the-plunge-how-newspaper-sites-that-charge-are-faring/, 02.09.2009.

[11] M. Ingram, „Debunking the "original sin" of online newspapers," Gigaom, http://gigaom.com/2012/02/06/debunking-the-original-sin-of-online-newspapers/, 06.02.2012.

[12] "Most Read Online Newspapers in the World: Mail Online, New York Times and The Guardian," comScore Data Mine, http://www.comscoredatamine.com/2012/12/most-read-online-newspapers-in-the-world-mail-online-new-york-times-and-the-guardian/, 13.12.2012.

[13] T. Krasnec, "Jutarnji list najčitaniji, slijedi Večernji pa 24 sata," Jutarnji list, http://www.jutarnji.hr/jutarnji-list-najcitaniji--slijedi-vecernji-pa-24-sata/150007/, 18.05.2006.

[14] D. Sullivan, „Google Still World's Most Popular Search Engine By Far, But Share Of Unique Searchers Dips Slightly," Search Engine Land, http://searchengineland.com/google-worlds-most-popular-search-engine-148089, 12.02.2013.

[15] http://www.alexa.com/, 08.04.2013

TABLICA II. PREGLED RAZLIKA IZMEĐU RAZLIČITIH ONLINE NOVINA KOJE SE PLAĆAJU, PREDNOSTI SU OBOJANE U SIVO

	24 sata	**Jutarnji list**	**Večernji list**	**Večernji premium**
lakoća pronalaženja	poveznica pri dnu stranice	Google: „Jutarnji list pretplata"	poveznica pri vrhu stranice	jasno naznačeni svi sadržaji
posjećenost (Alexa Traffic Rank)	3.667	2.992	4.756	4.756
lakoća navigacije	Newspaper Direct	kao papirnate novine	Newspaper Direct	Internet portal
kvaliteta sadržaja	nema većih predosti pred besplatnim sadržajima	opširniji članci, ranije objavljeni	nema većih predosti pred besplatnim sadržajima	opširniji članci, ranije objavljeni
redovitost ažuriranja	jednom dnevno	jednom dnevno	jednom dnevno	više puta dnevno
količina reklama	≈ 25% površine	manje od 20% površine	manje od 20% površine	desni stupac
povezanost s drugim sadržajima i interaktivnost	mogućnost dijeljenja sadržaja preko društvenih mreža, komentiranja	kao papirnate novine	mogućnost dijeljenja sadržaja preko društvenih mreža, komentiranja, poveznice na slične članke (nema vanjskih poveznica)	mogućnost dijeljenja sadržaja preko društvenih mreža, komentiranja, ankete, poveznice na slične članke, personalizacija
cijena za 1 mj.	69,07 kn	100,00 kn	93,08 kn	45,00 kn
cijena za 12 mj.	633,60 kn	470,00 kn	806,53 kn	260,00 kn

ePrivacy rules and data processing in users' terminal equipment: a Croatian experience

Nina Gumzej and Snježana Grgić

University of Zagreb Faculty of Law, Zagreb, Croatia
E-mail: nina.gumzej@pravo.hr

Personal Data Protection Agency, Zagreb, Croatia
E-mail: snjgrgic@gmail.com

Abstract – Implementation of revised ePrivacy Directive rules on accessing, and using information stored in users' terminal equipment was due in 2011 throughout the EU. Corresponding rules were already in 2011 transposed into Croatian law. In this paper authors analyze new ePrivacy Directive rules while focusing on cookies and similar technologies. They examine interpretations of compliance requirements as to selected types and purposes of cookies, and observed ambiguities towards reaching effective and user-friendly implementation in practice that complies with ePrivacy Directive. General conclusions are presented on research conducted into transposition and enforcement in the EU. This is complemented with research into legislative solutions of one EU State, not as representative or suggested best-practice but as example of most recently adopted rules in EU, while taking into account especially earlier observed ambiguities in practical implementation and nuances in adopted legal solutions EU-wide. Analysis of the setting in Croatia begins with a quick-test on cookie practices as regards domestic websites, followed by analysis of relevant legislation and related developments in Croatia. Research conducted supports authors' conclusions on needed guidance and awareness-raising activities together with monitoring and careful selection of best-practice approaches towards meeting compliance including, in particular, effective technological solutions supporting ePrivacy Directive requirements.

I. INTRODUCTION

In 2009 a new EU regulatory framework in electronic communications was adopted, which also included amendments to Directive 2002/58/EC (further: ePrivacy Directive). In this paper we are interested specifically in relevant amendments of rules that are intended to safeguard privacy of users of terminal equipment in cases of storing information, or accessing information already stored in such terminal equipment (personal computers, smart-phones and any other terminal equipment capable of connecting to the internet). [1] The need to have special rules in force at EU level with a focus on protection of privacy in electronic communications is explained by the fact that described activities represent forms of intrusion into private life of users. Terminal equipment of the user and any information stored in it is considered to form part of his or her private sphere requiring protection under the Council of Europe Convention for the protection of human rights and fundamental freedoms. [2] Data in terminal equipment relate to the user and they can be of sensitive nature, which especially needs to be taken into account where more users of same terminal equipment may access them. Harmful consequences of unlawful activities by third parties over a user's terminal equipment and his/her data are particularly present in viruses and software enabling secret monitoring of users and control over performance of their terminal, such as spyware. In addition to this increased attention over the past years has been directed to the practice of installing "cookies" and similar devices into users' terminal equipment and the resulting processing of data relating to users. Cookies can generally be defined as information generated by a website's server that is placed into the terminal (e.g. computer hard drive), while keeping a copy for itself. Unlike session cookies, permanent (persistent) cookies remain in the terminal and enable recognition of the user next time he or she visits a relevant website. As reported by the European Network and Information Security Agency - ENISA, activities relating to storing and sending cookies are in essence invisible to users by default. Moreover, information stored can be extensive so as to include any data provided earlier by a user to a website and personal data, such as a user name, address, e-mail address, as previously revealed by user to a website; attributes such as a name, domain and expiration date and credentials (e.g. user name and password); preferences and interface customizations; session data and data from sites, i.e. cached data, and tracking information on users. [3] While in many cases use of cookies is legitimate and necessary e.g. to be able to transmit a communication over the network or provide an online service the user explicitly asked for, there are circumstances where this is not the case and where users must be informed of cookies and approve their use. This is especially required for cases when third-party cookies are used for tracking users' behavior, habits and interests across various internet websites for profiling and targeting purposes (online behavioral advertising).

In this paper we are focusing on ePrivacy legislation and implementation thereof with respect to processing of data in users' terminal equipment by way of cookies and similar devices. We were in particular motivated for our research by observed ambiguities in implementation of relevant new ePrivacy Directive requirements throughout the EU, while also taking into account that the

corresponding legislative changes were already in 2011 adopted in the Republic of Croatia (EU candidate country at the time). We will show in this paper how noticed ambiguities in implementation relate to questions, such as types of cookies used and especially the purpose for which they are used. As to detailed interpretations of new rules at EU level we shall focus on selected interpretations and guidance that was until today issued by independent advisory body for personal data and privacy protection - The Article 29 Working Party on the Protection of Individuals with regard to the Processing of Personal Data (further: Art. 29 WP). [4] This body has had a very active role in providing guidance on meeting compliance with the new EU cookie rules and in this paper we shall primarily focus on interpretations of cookie consent exemptions. [5] Role and tasks of Art. 29 WP are not limited to general personal data and privacy protection as they extend also to privacy and data protection in electronic communications [6] and they are as such affirmed in the new EU regulatory regime for electronic communications. [7] All this is important to have in mind also because opinions it adopts according to rules of procedure [8] are of influence on relevant positions and practice of national data protection authorities (further: DPA) in Member States (whose representatives are its members), even though they are legally not binding. [9] We also note the proposed reinforced role of Art. 29 WP and especially its stronger coordinating role with respect to national DPAs in the new EU data protection framework. [10] As to its interpretations on new cookie rules, it appears on the basis of available data on national DPAs' websites that DPAs mostly do take note of them whether by simply referring to them or by making note of them in their own guidance on implementation of new cookie rules according to relevant national legislation. [11] Moreover, according to available data at least one telecom authority in an EU State having competence for enforcement of new cookie rules at national level would also be taking Art. 29 WP opinion(s) into account. [12]

Deadline to transpose the revised ePrivacy Directive in the EU was end of May 2011. New EU cookie rules were also in 2011 implemented in Croatia, with no transposition period envisaged in the relevant Electronic Communications Act (further: ECA). [13] We are, however, not aware of guidance issued on relevant ECA rules by the enforcing authority. We also noticed lacking research on this topic in national scientific literature. Consequently, the purpose of our research in this paper is also to provide a contribution towards filling of this gap.

II. NEW EPRIVACY COOKIE RULE - TOWARDS IMPLEMENTATION

According to the new rules EU Member States must ensure that storing of information, or gaining of access to information already stored, in terminal equipment of a subscriber or user is only allowed on condition that the subscriber or user concerned has given his or her consent, having been provided with clear and comprehensive information, in accordance with Data Protection Directive 95/46/EC, *inter alia*, about the purposes of the processing. This shall not prevent any technical storage or access for the sole purpose of carrying out the transmission of a communication over an electronic communications network, or as strictly necessary in order for the provider of am information society service explicitly requested by the subscriber or user to provide the service.

Prior to analysis of these requirements it is important to make two preliminary observations in relation to scope of the new rules and their relationship with general data protection rules, in cases where cookies and similar devices are used. Namely, legal protection is hereby ensured as regards any information stored in users' terminal equipment, due to the fact these belong to his or her private sphere [14], as we explained in the introduction. It is also important to point to the EU Data Protection Directive 95/46/EC (further: DP Directive), to which reference is made in the ePrivacy rules. [15] Namely, DP Directive has a close relationship with the ePrivacy Directive as it applies (*lex generalis*) to all matters with respect to personal data processing that are not specifically regulated by ePrivacy Directive (*lex specialis*). [16] In terms of ePrivacy cookie rules this means that DP Directive (national legislation adopted on the basis of it) needs to be taken into account should use of cookies or similar devices entail processing of information that is personal data. Since further analysis of the topic on identification in digital environment goes outside the scope of this paper we will here only touch on it to explain the above statement, by pointing to relevant interpretations of Art. 29 WP and in particular on personal data processing where persistent cookies with unique user IDs are used, which enables tracking of user's terminal equipment and especially in the context of profiling and targeting of a (thus identified, i.e. singled out) user. [17] To this we are adding references to identification concepts including digital identifiers in the proposed new EU data protection framework [18] and in recent case law of the Court of Justice of the European Union. [19] As to our second observation on scope, we find it clear in text of the cookie rule that it applies not only to providers of public communications networks and services but also to any party using described techniques [20], and especially information society service providers. [21] Moreover, cookie rule falls under the general rule guaranteeing secrecy of electronic communications (and related traffic data) that prohibits *any third party* from listening, tapping, storing, intercepting or any other monitoring of such data.

There are three key changes that the new ePrivacy cookie rules brought in relation to the old ones. Firstly, with the intention to ensure protection of terminal equipment and thus their users' private sphere from virus and spyware regardless of their delivery method, new rules apply also where electronic communications network was not used (e.g. delivery via external data storage media such as a CD, CD-Rom or USB key). Secondly, the old rules prescribed that users were to be *offered the right to refuse* such access/storage, while the new rules (operative provisions) require users to consent to this after having been informed. Such "informed consent" requirements in the ePrivacy Directive, pointing to DP Directive requirements [22], mean that cookie use (storing data, or accessing data already stored in terminal) is allowed when the user provided consent as a freely given specific and informed indication of his or her wishes

and by which this user signified his/her agreement with this (information having been provided to the user). Information to be provided includes in particular the processing purpose (purpose of storing, or accessing data, i.e. cookie purpose) and identity of entity doing the processing (normally, entity placing the cookies). [23]

The third change is a narrowed exemption from described legal requirements so that under the new rules only cases where technical storage or access *has the sole purpose of carrying out a communication* over the network are covered. According to Art. 29 WP transposing this requirement to cookies means there are no informed consent requirements to deliver cookies if without them it would be impossible to transmit a communication over the network, with the example of "load balancing session cookies". This also means that under the new rules use of cookies with the purpose to, for example, aid, speed up or regulate such transmission requires informed consent, as this function (where storage or access have the purpose of *facilitating* transmission of a communication) is no longer an exception.

The old rule providing for another exemption remained the same and this means that consent need not be obtained if storage or access is *strictly necessary for providing the information society service to a user who explicitly requested it*. The test proposed by Art. 29 WP to apply in cases of information society services consisting of more functionalities, such as a free online newspaper with additional functionalities for logged-in users (e.g. commenting), which function with their own cookies. When the user asks for it a cookie must be necessary to provide functionality to the user (otherwise if no cookie was used, functionality is unavailable) and the user asked for such functionality as part of requested information society service. Review of the element of "strictly necessary" was proposed to be looked at from user's point of view and not that of a service provider.

Following a common classification of cookies to session and permanent (persistent) as well to first and third party cookies, Art. 29 WP examined various types of cookies in relation to their main purpose, holding that a review of effectively applicable exemption conditions from informed consent requirement should be based on *purpose of cookie rather than its technical characteristics*. It concluded (without prejudice to DP Directive requirements, as we explained earlier) that use of particular types of cookies may under certain conditions be exempt from informed consent rule, and only if not used for additional purposes. Thus apart from load balancing session cookies that we discussed earlier (as necessary to carry out a communication over the network), eligible cookies would also be: a) "user input cookies", i.e. session-id, functioning, e.g., to keep track of input of the user filling online forms over more pages or of actively selected items in online shopping, such as by clicking on the "add to my shopping cart" button, for duration of session, or persistent cookies limited to a few hours in certain cases; b) "authentication cookies", for duration of session; c) "user security cookies", aiming to detect authentication abuses, e.g. repeated failed log-in attempts, for a limited persistent duration; d) "user interface customization persistent cookies, e.g. language and result display preference cookies, generally for duration of session; e) "multimedia content player session cookies", e.g., flash player cookies functioning to playback media content - user requested service), for duration of session; f) "third party social plug-in content sharing cookies", only for logged in social network members (typically operating via plug-in modules integrated in platforms of website operators in order to enable sharing content between social network users). Social plug-in cookies used for other purposes, such as tracking (social network members and non-members) for behavioral advertising, would not be exempted from the consent requirement.

Interestingly, upon examining first party analytics cookies and establishing these are not strictly necessary to provide users with requested service (online services with all functionalities are as a rule possible without them), Art. 29 WP did find such cookies to pose a lower risk to privacy and potentially subject to less strict exemption conditions than other types of cookies. In other words, when these cookies are only used for aggregated statistical purpose and with certain other conditions fulfilled, user-friendly opt-out mechanism would be accepted, with other requirements implemented (e.g. provided information on cookies to users in privacy policies, anonymization mechanisms/procedures applied to identifiers such as IP addresses). With that in mind Art. 29 WP suggested that any future amendment of relevant ePrivacy rules might include these cookies as another consent exemption. [24]

Third party advertising cookies that are used for behavioral advertising pose a greater risk to privacy. Art. 29 WP issued other opinions with guidance on the particular topic of online behavioral advertising and third-party cookies. [25] This is, for example, where advertising network providers place tracking cookies into user's terminal at first access to a website serving an ad of its network. For this reason such tracking cookies are referred to as "third party cookies" as opposed to "first party cookies" that are placed by website operators. Such cookie enables the ad network provider to recognize users when they revisit this website, or when they access any other partnering website (ad network). Thus tracking operates on the principle the more detailed information collected on user's habits and interests, i.e. behavior across various websites, the more detailed profiling is enabled and thus better targeting of the user. With use of these cookies collected information relating to user' online behavior is shared with third parties, such as advertisers. Situations where same terminal equipment is used by more users needs to also be taken into account as this could result in users receiving ads targeted on the basis of surfing habits of other users. With all this in mind and especially the high levels of privacy intrusions with the use of third party advertising cookies, it is understandable that their use cannot be exempt from the informed consent requirement.

With respect to new consent requirements, i.e. a switch from "right to refuse" in earlier rules into ensuring informed consent, it should be noted that was a rather debated issue in the relevant EU legislative process [26], and it seems to us that up until today a clear uniform practice to ensure effective, practical, user-friendly and altogether ePrivacy Directive compliant informed consent still appears a myth. A detailed analysis of possible ways

to satisfy informed consent requirements under the ePrivacy Directive goes outside the scope of this paper. [27] However, we would here like to briefly point to the often asked question with respect to relevant use of technology, if informed consent can validly be expressed via settings of a browser or other application. [28] Here, for example, along with many others the Art. 29 WP has been calling for effective technical solutions towards enabling affirmative actions of users to signify their valid informed consent. Its reasons for opining that informed consent requirements are not met (browser settings) are mainly motivated by attempts to construe consent validly expressed just on account of a presumption of a user using his or her browser with a default setting to accept all cookies. [29] In this respect in particular work on a global "Do Not Track" (DNT) standard needs to be mentioned, which is highly prioritized and strongly supported also by the European Commission, however, subject to effectiveness and viability of any such future standard. [30] This work is on the agenda at the World Wide Web Consortium. [31] Clearly, innovative technical solutions towards valid informed consent via browser settings or other applications continue to be explored.

III. NATIONAL IMPLEMENTATION AND ENFORCEMENT

A. *Situation in the EU – overview*

Most Member States have by beginning of February 2012 enacted relevant legislation, which was also influenced by actions of the European Commission against Member States [32]; however, a comparative overview of implementing national legislation shows nuanced approaches. [33] With this in mind, and taking into account earlier described developments in the area of cookie law compliance at EU level we decided to analyze the more specific legislative solutions implemented by a Member State with *most recently adopted* legislation, as it may be argued to have had a "comparative advantage" in terms of time to review possible implementation solutions as well as expected extent of non-compliance with new requirements. According to available information this is Slovenia, which amended its Electronic Communications Act (further: SECA) at the end of 2012 with adopted new rules on cookies, as follows. [34] The first more specific rule adopted to point to is the prescribed possibility to have consent expressed via browser settings (in accordance with the relevant recital we pointed to earlier in this paper). Next, where a breach of SECA cookie rule is at the same time a violation of Slovenian general data protection act it is prescribed that only sanctions and enforcement mechanism provided for in SECA apply (enforcing authority for cookie rules is the DPA). We may also point to the adopted solution on sanctions for violations of relevant cookie rules. These are relatively low, falling in the second lowest by category in the range of five categories. According to relevant submissions of the Slovenian DPA in the SECA legislative procedure a light sanction was considered most appropriate due to the fact that mostly small and mid-size organizations would be caught up by new cookie rules. [35] It should also be pointed out that a transition period of five months was prescribed for the new cookie rule. These solutions are not representative or suggested best-practice; however, they show us a rather careful and pragmatic approach towards enforcement of new cookie requirements.

As regards enforcement in the EU competent (enforcing) authorities have according to our research already issued helpful guidance on cookies and on compliance with the new cookie rules in some EU Member States, for both the users and relevant service providers, and we expect many more will follow during this year. We also observed that early steps towards enforcement include instructions to responsible entities, such as website operators, to carry out audits into cookies they use, their purpose and the need for them, which is followed by evaluation of exemption requirements. [36] In some Member States supervisory authorities already started to check compliance, using methods such as sending out letters to website operators asking for information on steps taken to ensure compliance. [37] Direct online reporting of concerns on use of cookies or similar technologies by websites is also being spread out and made available for the users. [38]

B. *Indication of cookie practices in Croatia: quick-test*

Before analyzing the situation in Croatia with respect to legislation transposing ePrivacy Directive cookie rules and related developments we felt it appropriate, in light of topic of this paper, to make a quick-test in order to gain at least some indication of cookie practices by domestic website operators. However, for relevant results due to limited test scope and methodology further more extensive research should be conducted. We conducted the test on thirty websites [39]. Methodology used included tracking of the number of cookies left of each page in a single session, distinguishing cookies stored on user's terminal equipment - computer from cookies that are stored locally at the side of service provider or a third party as well as cookies that come from the service provider and cookies coming from third companies. Flash cookies were not considered. We also reviewed the method of informing users on collection of information through cookies.

In this test we found a total 884 cookies in a single session, 795 on user's computer and 89 stored locally at the service provider or a third party, or average website has almost 30 cookies, 24 percent of these come from service providers and 76 percent come from third parties. As to number of cookies used, we found it interesting to compare quick-test results with data from a UK study by TRUSTe, according to which the average website has 14 cookies per page. [40] 32 percent of these come from service providers and 68 percent come from third party companies, which could be analytics companies or companies delivering advertising. In our test we observed that analyzed websites have doubled averaged cookies per page and 8 percent cookies more than in the UK study coming from third parties. Quick-test results have also shown that only a few websites clearly informed users on placement of cookies and a vast majority of others did not provide any information. Namely, three of the thirty sites informed users on placement of cookies via Terms of Use, while none asked for their consent or ensured to have it. We interpreted this finding as indicating that only the more privacy-aware users with some understanding of cookies would be able to protect themselves from

receiving unwanted cookies, by relying independently on their browser settings, such as the "Enhanced browser settings" option. This led us towards further research into relevant behavioral patterns and awareness of users according to available studies in other countries. We noted, according to a UK study from 2011 [41] that only 13 percent of people fully understand how web cookies work and 32 percent changed default settings on their browser to ensure more privacy. Taking these results into account and applying them to the Croatian environment, question is how effective could described protection really be for average internet user to identify what cookies he/she has, assess their intrusiveness, and decide on his/her strategy for accepting or rejecting them?

In our opinion results of the quick-test indicate the need for organizational culture in Croatia ensuring respect for legal privacy rules which go beyond transparency and for a role of enforcing authorities who in the first place need to provide guidance on relevant cookie rules, both to service providers and users. This is especially relevant given that legislation does not have clear answers on how cookie-relevant information to users is to be provided or their consent obtained, while taking into account that this should be user-friendly. Guidance that is already today available from competent authorities in various EU Member States, and to which we pointed in the paper, could be of great help towards setting first steps in this respect (in compliance with ECA cookie rule, of course).

C. *Legislation and related developments in Croatia*

In 2011 the Croatian legislator transposed relevant new ePrivacy Directive rule into ECA, mostly in line with it, with the exception that protection in ECA was not widened to all cases of access and/or storage, regardless of means. We explained in section II. of this paper that widened scope of protection was prescribed in new ePrivacy rules to ensure stronger and uniform protection of users from intrusions into their private sphere. Conversely, scope of protection under ECA remains applicable, as in the old rules, only to cases where access and/or storage (e.g. spyware delivery) is made over the electronic communications network. [42] With respect to possible ways to express informed consent, we noted that ECA does not expressly prescribe the browser settings option. This is in line with operative rules, i.e. articles of the ePrivacy Directive cookie rule, which do not regulate this (as explained earlier, relevant recital mentioning browser settings as an option is a non-operative rule). However, we do find room for improvement of current solution on user's consent in ECA. Namely, definition of consent in ECA is autonomous and not referenced to the Personal Data Protection Act, which transposes DP Directive into Croatian legislation. We are noting this as a potential issue since it may give room for different interpretations of consent requirements under the two laws and especially where informed consent for the use of cookies and similar devices is interpreted in practice. [43]

Violations of relevant new rules are considered serious infringements of ECA and sanctions prescribed are severe. [44] The Croatian Post and Electronic Communications Agency is the national regulatory and enforcing authority for ECA. Prescribed sanctions apply to any form of infringement, which means that any use of electronic communications network for storing data, or for access to stored data in contravention of ECA is sanctioned. A monetary fine prescribed goes from a minimum of HRK 100.000,00 (cca. EUR 13.200,00) to a maximum of HRK 1.000.000,00 (cca. EUR 132.000,00) for a legal person (responsible persons in legal persons may be fined as well, up to HRK 100.000,00), and for violations committed by natural person the fine can go from HRK 10.000,00 to 50.000,00. The fine range (for responsible person in legal person and for natural person) is doubled in violations committed with a view to obtain pecuniary gain. This could in our opinion be a factor to consider where, e.g., tracking cookies are used (in contravention of ECA) to target users with ads.

Our analysis shows that rules on cookies in ECA are in overall very strict. We are, however, unaware of guidance on implementation that may have so far been issued by the enforcing authority, both for the users and relevant service providers. This topic also does not seem to be covered or at minimum not adequately covered, in national scientific literature. It is not unlikely that urgent adoption of revised ECA cookie rule without public consultation on proposed amendment to the old rule (and thus without a proper dialogue between all stakeholders) contributes the observed situation. [45] In any case, we look forward to all future awareness-raising activities in the area.

IV. CONCLUSION

Main purpose of new ePrivacy Directive rules on cookies is to better inform users about what cookies do and to put them in control of cookie data collection. As we showed in this paper, legislation on the basis of new cookie legislation EU-wide and in Croatia is mostly transposed, though many important practical questions remain open as regards implementation of new requirements. Implementation of national rules adopted on the basis of ePrivacy Directive in our opinion needs to be subject to careful consideration and meaningful guidance provided by competent (enforcing) national authorities, especially taking into account results of our research in section II. of this paper. Cookie guidance needs to be directed toward all obliged parties such as service providers, as well as to internet users. Even though we noted that publicly available guidance issued by national competent authorities does not in absolutely all cases expressly refer to, or exactly conform to relevant guidance issued by specialized bodies at EU level, such as the Art. 29 WP to which we referred in our paper, it is in our opinion best practice for obliged parties to know of them and take them into account. Consultations between competent authority and all interested stakeholders are a good practice example towards a proper and common sense implementation of new legal requirements. Described course of action is in any case a crucial first step before any enforcement activities aiming towards any conclusive case law in this area are initiated at national level. Additionally, any best practice approach will require close monitoring of important developments in the area, including innovative technological solutions towards meeting ePrivacy Directive requirements.

REFERENCES

[1] Article 2(5) and recitals 65-66: "Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws", OJ L 337, 18.12.2009., pp. 11-36; Article 5(3) and recitals 24-25: "Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector", OJ L 201, 31.7.2002., pp. 37-47.

[2] Recital 24 of ePrivacy Directive. Also see recital 25 of ePrivacy Directive and recitals 65-66 of Directive 2009/136/EC - in [1].

[3] R.Tirtea, C. Castelluccia, D. Ikonomou, "Bittersweet cookies. Some security and privacy considerations. European Network and Information Security Agency", 2.2.2011., http://www.enisa.europa.eu/act/it/library/pp/cookies.

[4] Articles 29-30 of "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data", OJ L 281, 23.11.1995, pp.31-50.

[5] Art. 29 WP, "Opinion 04/2012 on Cookie Consent Exemption", 00879/12/EN, WP 194, 7.6.2012. Other relevant work: "Opinion 16/ 2011 on EASA/IAB Best Practice Recommendation on Online Behavioural Advertising", 02005/11/EN, WP 188, 8.12.2011; "Opinion 2/2010 on online behavioural advertising", 00909/10/EN WP 171, 22.6.2010, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/index _en.htm.

[6] Article 15(3), recital 48 of ePrivacy Directive – in [1].

[7] Article 4(5) of ePrivacy Directive - Article 2(4c) of Directive 2009/136/EC; Article 15a (4) of ePrivacy Directive - Article 2 (10) of Directive 2009/136/EC, recital 52 of Directive 2009/136/EC - in [1].

[8] Art. 29 WP, "Rules of procedure", 15.2.2010, http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/rules-art-29_en.pdf.

[9] Art.29 WP opinions influence positions of national DPA's and national practice to a certain extent - this has so far not always been the case, see: European Commission, "Commission Staff Working Paper, Impact Assessment - Accompanying the document Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) and Directive of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data", SEC(2012) 72 final, 25.1.2012, p. 18, and "Annex 2 to the Impact Assessment: Evaluation of the implementation of the Data Protection Directive", at 10.13.

[10] Proposal is currently in legislative procedure: "Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)", COM (2012) 11 final, 2012/0011 (COD), Brussels, 25.1.2012, see esp. Chapter 7.

[11] Selected six examples of types of references to Art. 29 WP opinions on cookies include relevant documentation of: 1) Irish Data Protection Commissioner's Office - *Oifig an Chomisinéara Cosanta Sonraí,* http://www.dataprotection.ie/docs/Topical_Data_Protection_Issues/1241.htm#3; 2) Italian DPA - *Garante per la protezione dei dati personali,* http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/2146935; c) UK Information Commissioner's Office, http://www.ico.gov.uk/for_organisations/privacy_and_electronic_communications/the_guide/cookies.aspx; French DPA - *Commission nationale de l'informatique et des libertés*, http://www.cnil.fr/en-savoir-plus/fiches-pratiques/fiche/article/ce-que-le-paquet-telecom-change-pour-les-cookies/; and Art. 29 WP Opinions referred to at website of the Polish DPA - *Generalny Inspektor Ochrony Danych Osobowych,* available by search at: http://www.giodo.gov.pl, and the Slovenian DPA - *Informacijski pooblaščenec*, available by search at: https://www.ip-rs.si .

[12] Dutch telecoms regulator OPTA, for example, see (in Dutch): IAB Nederland, "Verslag OPTA rondetafelbijeenkomst cookiewetgeving i.s.m. DDMA en IAB", 6.8.2012, at p. 2, http://www.opta.nl/nl/download/publicatie/?id=3635.

[13] Electronic Communications Act, Official Gazzette of the Republic of Croatia nos. 73/08, 90/11 and 133/12.

[14] In addition to text of relevant rules see Art. 29 WP Opinion "16/2011" at p. 8 and "Opinion 2/2010" at p. 9. - in [5].

[15] Directive 95/46/EC - in [4].

[16] ePrivacy Directive: Article 1(1-2) and recital 10; for detailed analysis see Art.29 WP, "Opinion 2/2010" - in [5], at pp. 8 *et seq*.

[17] Art.29 WP "Opinion 2/2010", and "Opinion 16/2011" at p. 8. - in [5], Art. 29 WP, "Opinion 1/2008 on data protection issues related to search engines", 00737/EN, WP 148, 4.4.2008, http://ec.europa.eu/justice/data-protection/article-29/documentation/ opinion-recommendation/index_en.htm.

[18] Article 4 (1-2) of proposed General Data Protection Regulation – in [10]. See also recitals 23-24.

[19] Case C-70/10, "Scarlet Extended SA v Société Belge des auteurs, compositeurs et éditeurs (SABAM)", paragraph 51; C-360/10, "Belgische Vereniging van Auteurs, Componisten en Uitgevers (Sabam) v Netlog NV", paragraph 49 (cases not yet published in European Court Reports, text available at: http://curia.europa.eu).

[20] Also see Art. 29 WP "Opinion 1/2008" in [17] at p. 12, and "Opinion 2/2010" in [5] at p. 9.

[21] For concept of information society services and examples, see Article 2a and recitals 17-18 of "Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market", OJ L 178, 17.7.2000, pp. 1-16., in relation to Article 1(2) of Directive 98/34/EZ, OJ L 204, 21.7.1998., pp. 37-48 (as amended by Directive 98/48/EC, OJ L 217, 5.8.1998., pp. 18-26). Also, this directive does not cover questions relating to information society services covered by Directives 95/46/EC and 97/66/EC (today Directive 2002/58/EC), which mean these rules fully apply to providers of these services, see Article 1(5b) and recitals 14-15 of Directive 2000/31/EC.

[22] See Article 2 of ePrivacy Directive in [1] and recital 17 for reference to consent in DP Directive in [4]. Terminology of DP Directive is in Article 2 and information requirement in Article 10.

[23] Detailed guidance in the context of behavioral advertising (and presumed personal data processing) is available in: Art.29 WP, "Opinion 2/2010" in [5]. DPAs in some Member States already issued their own guidance on implementation, see e.g.: UK Information Commissioner's Office in 11 "Privacy and Electronic Communications Regulations - Guidance on the rules on use of cookies and similar technologies", V.3 May 2012, http://www.ico.gov.uk/for_organisations/privacy_and_electronic_communications/the_guide/~/media/documents/library/Privacy_and_electronic/Practical_application/cookies_guidance_v3.ashx. For example of guidance issued by national telecom authority see: Danish Business Authority - *Erhvervsstyrelsen*, "Guidelines on Executive Order on Information and Consent Required in Case of Storing or Accessing Information in End-User Terminal Equipment", 8.12.2011, http://www.erhvervsstyrelsen.dk/file/253400/cookie-exec-order-guidelines-english-version.pdf.

[24] Art.29 WP "Opinion 04/2012" in [5].

[25] Art.29 WP "Opinion 2/2010" and "Opinion 16/2011" in [5].

[26] Compare text of recital 66 of Directive 2009/136/EC and new Ar. 5 (3) of ePrivacy Directive and see statements by 13 Member States on new requirements: "Citizens Rights Directive, E-Privacy Directive", Council of the European Union, Addendum to "I/A" note, from General Secretariat of the Council, to: Permanent Representatives Committee (Part 1) / Council 2007/0247 (COD), Brussels, 18. 11. 2009, p. 3, http://register.consilium.europa.eu/pdf/en/09/st15/st15864-ad01re01. en09.pdf.

[27] For more details on this, see e.g. Art. 29 WP "Opinion 16/2011" in [5], esp. at pp. 9-11, and "Opinion 2/2010" in [5], at pp. 13-17;

advice on meeting local compliance can be found in relevant documentation issued by enforcing authorities, e.g. see in [23].

[28] Recital 66 of Directive 2009/136/EC – in [1].

[29] For more details see: "Opinion 16/2011" in [5], esp. at p. 10, and also see "Opinion 2/2010" in [5], at pp. 13-15.

[30] Blog of Neelie Kroes, "Why we need a sound Do-Not–Track standard for privacy online", http://blogs.ec.europa.eu/neelie-kroes/donottrack/; Neelie Kroes, "Online privacy and online business: An update on Do Not Track The Centre for European Policy Studies (CEPS)/Brussels", SPEECH/12/716, 11.10.2012, http://europa.eu/rapid/press-release_SPEECH-12-716_en.htm.

[31] "Tracking Preference Expression (DNT)", W3C Working Draft 2.10.2012, latest version at: http://www.w3.org/TR/tracking-dnt/.

[32] Actions initiated by European Commission before the Court of Justice of the EU (in the meantime legislation has been implemented) against Slovenia (C-406/12, OJ C 331, 27.10.2012, p. 15); Poland (C-331/12, OJ C 287, 22.9.2012, p. 24), and Portugal (Case C-325/12, OJ C 287, 22.9.2012, p. 22).

[33] For a quick example of nuances in adopted legislation during 2012, compare Article 11(7a) of Dutch Telecommunications Act – "Telecommunicatiewet", http://wetten.overheid.nl/BWBR000995, and Article 157 of Slovenian Electronic Communications Act - "Zakon o elektronskih komunikacijah (ZEKom-1)", Uradni list RS, št. 109/2012. EU-wide implementation chart with most recent update as of 02/13 is available here: Field Fisher Waterhouse LLP, "Cookie 'consent' rule: EEA implementation", 4.2.2013, http://www.ffw.com/pdf/cookie-consent-tracking-table.pdf.

[34] Slovenian Electronic Communications Act in [32]: Articles 157, 161, 235(1) point 13, and Article 235 (2-3).

[35] Informacijski pooblaščenec, "Stališče Informacijskega pooblaščenca glede predloga Zakona elektronskih komunikacijah (ZEKom-1) – osnutek za javno obravnavo", No. 007-21/2011/8, 28.11.2011, at p. 7, http://www.mizks.gov.si/fileadmin/mizks.gov.si/pageuploads/Informacijska_druzba/ZEKom1/ZEKom1_JO_12-07-11_IP.tif.pdf.

[36] See, e.g., guidance from UK Information Commissioner's Office in [23], at pp. 15 *et seq*, and from the Danish Business Authority Guidelines in [23], at pp. 14-15.

[37] Example of the UK and Irish DPAs, respectively: Dave Evans, "ICO blog: Education key to cookie law progress", 10.9.2012, http://www.ico.gov.uk/news/blog/2012/education-key-to-cookie-law-progress.aspx; "Data Protection Commissioner commences action on "Cookie" law", 19.12.2012, http://www.dataprotection.ie/docs/19/12/12_Press_Release_-_Data_Protection_Commissioner_commen/1274.htm.

[38] See e.g.: Information Commissioner's Office, "Tell us your concerns about cookies", https://www.snapsurveys.com/swh/surveylogin.asp?k=133707671186; OPTA - complaints submitted via helpdesk ConsuWijzer and with advice available at: http://www.opta.nl/en/about-opta/consuwijzer/; http://www.consuwijzer.nl/telecom-post/internet/privacy/uitleg-cookies.

[39] Quick-test was conducted on 30 most visited websites according to the monthly survey gemiusAudience for December 2012. Gemius d.o.o., "gemiusAudience", http://www.audience.com.hr.

[40] TRUSTe, "2012 UK TRUSTed Websites Index", http://www.truste.com/uk-privacy-index-2012-websites.

[41] D. Lancefield, M. Ambler, M. Rauber and R. Patel, "Research into consumer understanding and management of internet cookies and the potential impact of the EU Electronic Communications Framework", April 2011, http://www.culture.gov.uk/images/consultations/PwC_Internet_Cookies_final.pdf.

[42] Article 100(4) of ECA - in [13].

[43] Though very similar these definitions are not the same, compare Article 2(1) point 54 of ECA with Article 2(1) point 8 of the General Data Protection Act, Official Gazzette of Republic of Croatia no. 106/12 (clean integral text).

[44] Article 119 (1), point 52 and Article 119 (2-7)- in [13].

[45] Ministarstvo mora, prometa i infrastrukture, "Javna rasprava o izmjenama i dopunama Zakona o elektroničkim komunikacijama", http://www.mppi.hr/default.aspx?id=8981; Vlada Republike Hrvatske, "Prijedlog zakona o izmjenama i dopunama Zakona o elektroničkim komunikacijama, s Konačnim prijedlogom zakona, 7.7.2011.", http://www.sabor.hr/fgs.axd?id=18448 – with respect to proposals to amend Article 100 (4) of ECA.

The Cost of Information Security Management in Offshore SMB ICT Companies

Saša Aksentijević[1], Edvard Tijan[2], Dragan Čišić[3]
[1] Saipem SpA Croatian Branch
Alda Colonnella 2, Rijeka, Croatia
Tel: +385 51 65 17 00 Fax: +385 51 65 17 81 E-mail: axy@vip.hr
[2, 3] University of Rijeka, Faculty of Maritime Studies
Studentska 2, 51000 Rijeka, Croatia
Tel: +385 51 33 84 11 Fax: +385 51 33 67 55 E-mail: etijan@pfri.hr, dragan@pfri.hr

Abstract - Companies belonging to offshore SMB ICT segment are subjected to various costs arising from several sources like legal compliance, alignment with best practice guidelines and standards, employee education, basic computer and network infrastructure security and cost of SaaS/cloud solutions. Furthermore, such companies usually have very limited financial resources, yet they are often involved in large projects working for major offshore installation contractors. In this paper the authors will outline basic costs of information security management systems in offshore SMB companies and propose a simple model to continuously monitor and control them.

I. INTRODUCTION

Term "offshore" is nowadays usually used for "oil and gas drilling operations that are conducted in the ocean" [1]. However, they can also relate to such operations conducted in any large open or closed waters or lakes (for example, Mediterranean Sea or Caspian Lake). Offshore construction projects in oil and energy sector are usually executed by large companies called engineering and construction companies. In their form, they are usually corporations, or joint ventures/consortiums of such companies. Usual services that may be provided by such companies are "engineering, fabrication, transport, installation, procurement, research, manufacturing, environmental systems and project management" [2].

All these companies use very complex ICT systems in order to facilitate their core operations. Complexity of those systems is further compounded by the fact that operations are usually executed in difficult areas that are geographically remote, do not provide opportunity for adequate user support, aboard vessels, where it is difficult to obtain good quality hardware, data links and skilled personnel. Offshore ICT systems also include equipment that is seldom encountered in onshore or conventional ICT operations, like safety radio location beacon equipment in case of emergency situations or marine satellite equipment with self-pointing/auto-acquiring antennas [3].

For reasons that will be discussed in details, offshore construction companies usually subcontract local ICT companies belonging to SMB market segment to provide some of the ICT services required for successful completion of projects. Large offshore construction companies usually have well developed formal Information Security Management Systems (ISMS) [4], so it is a real challenge for smaller ICT companies working for them to keep up the pace with their client's information's security cost. This causes significant rise in information security costs for such companies and inherent need for proper management of that particular type of costs.

II. CHARACTERISTICS OF OFFSHORE PROJECTS

There are some characteristics of offshore projects in oil and energy sector, separating them from other large projects in other sectors, for example, civil engineering, road construction or dam construction. These characteristics are very important in order to understand the specific requirements of offshore ICT security that needs to be maintained and delivered by relatively smaller ICT companies belonging to SMB sector.

Some usual characteristics of offshore projects are as follows:

1. Offshore projects are typically very complex projects requiring mobilization of large capital base, human resources and usually, application of the most modern available technology,
2. They can be very diverse according to their length, from very short-term to long-term. Very often, large mega-projects are divided into smaller projects with different subcontractors,
3. Typically, there are several subcontractors working on a single installation project and their cohesion and cooperation is critical for successful execution of the project. These subcontractors use different methods, technologies, have diverse level of development of human capital and operate in different technical areas,
4. Offshore projects are projects connected with large risks that have to be properly quantified in order to be managed,
5. Environmental, health and safety and sustainability issues are some of the main considerations all companies operating offshore have to take into account. These issues are usually in the way of successful project execution,
6. Contracts for various phases of offshore contracts are typically stipulated very close to the moment when the project should start. This puts additional pressure to the management of projects,

7. Offshore projects are usually executed in very difficult areas, by one or several of the following criteria:

    - Harsh environment: extreme cold or warmth, deep sea, Arctic conditions
    - Politically unstable countries, sometimes even in war-stricken areas
    - Technically challenging environment: extreme depth, very shallow water, ice, mud, etc.
    - Logistics problems: remote areas that are typically away from main traffic routes, posing potential logistics problems

8. Political and sociological content of offshore projects is very high. Teams working on offshore projects are usually multinational and multicultural, which is an additional challenge to be tackled during long and exhaustive project planning and execution phases.

Due to all outlined characteristics of offshore projects, it is clear that they carry a large risk with them. This risk has to be properly managed. The goals of offshore project risk management process are the following [5]:

1. Setting realistic but reasonable cost and schedule contingencies,
2. Understanding the probability of cost overruns and delays of anticipated schedule,
3. Knowing the probability that the contracted cost and schedule will be achieved,
4. Understanding the accuracy of a cost estimate or project schedule, and
5. Ensuring that project teams identify and properly communicate risks and implement a risk mitigation plan.

III. SPECIFIC REQUIREMENTS OF OFFSHORE ICT SECURITY

SMB ICT companies are subject to quite specific requirements when it comes to ICT security. First and foremost, as already explained, they are operating on very complex projects in difficult areas that are geographically remote. In order to better understand what the specific requirements of offshore ICT security are, some specifics will be outlined:

1. Legal framework under which offshore projects are executed is very complex. It usually transcends a single nation, and refers to several countries. Also, considering that offshore projects are usually related to work conducted at the sea, maritime law is also applicable.
2. Existing legal requirements imposed in front of ICT security are very strict for business areas like financial sector or technical aspects of ISMS management. However, they are not easily applicable to ICT offshore operations of SMB companies due to different business context and available financial means.
3. Very often, investments in ICT security of SMBs is based on professional evaluation of cumulative risk or subjective evaluation of the owner or ICT project manager about justification of the investment compared to such risk. Best practice ISMS systems and frameworks do not evaluate influence of investments in SMS to company's or project's financial results.
4. One of very important restrictions for SMS ICT companies working on offshore projects in oil and gas sector is lack of internal human resources and financial strength that could adequately follow up growing requirements for ICT security solutions. Neglecting such requests usually results in increased levels of impact of security incidents and cost of remediation and opportunity cost.
5. ISMS management in offshore ICT operations is usually viewed as a technical discipline or as a "minimum cost" endeavor with unclear relation towards project cost or profit margin. Overall, a clear model does not exist that would put offshore ICT ISMS management in relation with business result of the offshore project.
6. One of the most common strategies used instinctively by small and medium businesses providing ICT services in offshore projects is accepting unreasonably high levels of risk and avoiding investments in offshore ICT security solutions. Investments in such ISMS solutions are usually perceived by those businesses as unnecessary or sunk cost.
7. Business financing sector, and especially banks, following the work of offshore ICT companies do not recognize the importance of ICT security for successful business models of such companies. No special analysis or economic impact of adopted ICT measures is required even though they are crucial for successful completion of offshore ICT projects and consequentially, both for the client's and ICT subcontractor's success. This way both the banks and SMB offshore ICT companies are facing unsorted, implicitly accepted risks.

From the outlined above, it is clear that the usage of economic criteria in decision making about investments in ISMS solutions for offshore ICT SMB companies is a very important factor. Successful ISMS systems of such companies have to include legal requirements, cost-benefit analysis of possible ICT security solutions, and risk based analysis. Such an approach has to be distant from usually adopted approach that includes only autonomous technical measures and haphazard risk assessment.

IV. PROPOSED MODEL OF OFFSHORE ICT COST MANAGEMENT IN SMB SECTOR

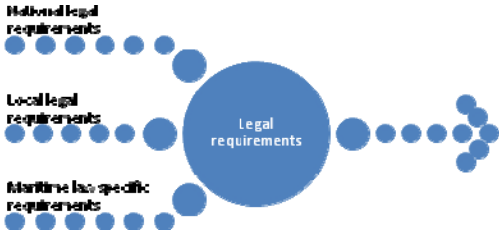
In order to propose a viable model of offshore ICT cost management in SMB sector, it is necessary to think of all possible requirements related to adopted ICT

security models. From what has already been outlined, they are divided into three sub-groups:

1. Legal requirements
2. Best practice requirements
3. Risk assessment requirements

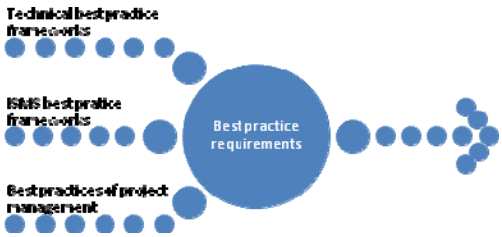
As shown in fig 1., a typical SMB ICT offshore company is subject to various national legal requirements (for example, requirements related to minimum of information security measures to be implemented, record retention, disaster recovery and business continuity). Furthermore, those SMB ICT companies providing network, radio or satellite communication hardware very often have to undergo a very strict process of local certification (equipment conformity, radio frequencies, and encryption systems). All these compliance requirements add up on running costs of a typical offshore ICT company. Finally, there are also specific requirements imposed by compliance with laws applicable to maritime and offshore operations.

Fig 1. Legal requirements of SMB ICT offshore companies



Typically, there are three sets of requirements related to best practice of execution of ICT projects. Primarily, they are technical best practice frameworks that already include certain levels of ICT security context. Their origins are usually best practice systems established by hardware or software manufacturers and specialized associations. Best practice frameworks for ISMS adoption and management are formalized and certifiable best practice standards endorsed by international bodies and typically well spread in the ICT business community. Last, but not least, best practices also relate to ICT project management, whose information security practices might prove to be of utmost importance for successful project completion.

Fig 2. Best practice requirements



One of the most widespread systems for ICT security risk management is the one that is risk based. Risk based approach usually lists all information assets, their vulnerabilities, matches them with applicable threats and end result is a matrix of assets, threats and vulnerabilities that carry certain level of risk that has to be mitigated using applicable list controls. This process should be endorsed by the top management in order to demonstrate its willingness for achievement of goals of excellence. In case of SMB companies, top management can be a single person, or in case of smaller companies, even an owner.

Fig 3. Risk assessment approach of SMB offshore ICT companies

Finally, inputs for baseline SMB offshore ICT security are cumulative requirements for applicable legal requirements, best practice requirements and risk assessment approach requirements, applicable for baseline operations, as shown in Fig 4.

Fig 4. Cumulative baseline SMB offshore ICT security

Cumulative baseline level of SMB offshore ICT security presents a set of expenditures, either in form of investments or costs that has to be maintained continuously. A similar exercise can be done on a project basis, where for a specific project, a matrix can be done with all possible risks related to that project. In that case, SMB offshore ICT security includes also temporary ICT security risks and measures that are existing only for the duration of the project, and after the project, the requirements for ICT security return to the baseline. Successful ICT security management in offshore operations manages to retain all mitigation measures and expenditures inside the baseline requirements and therefore avoid multiplication of the same expenditure through various ongoing projects, as show in Fig 5.

Fig 5. Portfolio structure of information security requirements and solutions for SMB ICT companies working on offshore oil and energy projects

Clearly, a portfolio approach [6] would be advisable to all companies operating in this segment, with well-developed portfolio analysis on both project, and basic ICT security foundations.

V. EXPECTED DEVELOPMENTS IN OFFSHORE SMB ICT SECURITY SYSTEMS IN NEAR FUTURE

There are several trends that can already be well identified and that have already started having impact on ICT security requirements of SMB companies providing services to offshore engineering and construction companies, and especially to their cost efficiency. These trends are the following:

1. Shift towards cloud based solutions as a cost enhancing solution is not always or easily applicable to offshore area. Data links that are usually satellite based do not allow for usage of public cloud based solutions as a viable option [7].
2. Offshore companies and large contractors typically prefer standard and well-proven solutions and measures to achieve goals of ICT security.
3. Host countries are likely to continue implementing more and more strict measures for control of information flow as the operations continue to move to more difficult and dangerous areas in terms of political, social and economic risk.
4. Local infrastructure in host countries that is often lacking in technology (hardware and software) used to achieve goals of ICT security and in lifeware will probably in the future present even bigger challenge for SMB ICT companies providing ICT project delivery offshore. This means additional pressure on profit margins and timely delivery as most solutions will have to be imported from other countries and not be obtained in host countries.
5. Cost of local certification of information security solutions and their maintenance continues having a big impact on the cost side of SMB companies providing ICT solutions to big contractors.
6. There is a number of hidden costs that have a large impact on operations too. These costs are typically constantly rising. Some of these costs are cost of equipment transport, import, storage, expediting, installation, cost of visas for the technicians, accomodation and personnel security, etc..

Therefore, there are a number of factors exerting influence on increase of cost of operations of information security of SMB ICT companies operating on offshore projects. In the near future, it cannot be expected that percentage of these costs in total operation costs will decrease. Also, major driver in enhancement of competitiveness of such companies will be portfolio approach to information security solutions that are applicable on various projects and regional orientation to certain areas that are more homogenous in requirements imposed on SMB ICT companies providing solutions.

VI. CONCLUSION

Offshore projects in oil and energy sector are very diverse, but are typically midterm to long term projects executed in remote and difficult areas. Engineering and construction companies and their assets are contracted to develop certain phases of the project, and they usually contract other companies for some phases of the project. Various companies belonging to SMB segment of the market are often contracted to provide specific services and installations of hardware, software and network infrastructure.

These companies are faced with large legal requirements, national and those in the territory where operations and installations are being executed, best practice and technology requirements and those arising from internal professional risk assessment. On the other hand, these companies usually have limited resources, financial and human; they can dedicate to this additional information security requirements. The number of such requirements is constantly rising.

The most appropriate approach to create and maintain solid ICT security systems for SMB ICT companies is to analyze its baseline requirements and create a portfolio satisfying both baseline requirements and project portfolio specific requirements. This will ensure that there is no duplication of costs and implementation of unnecessary solutions correlated with cost increase.

Near future will bring even more requirements for certifications, offshore ICT security solutions and locally imposed restrictions on the ways ICT companies may approach those issues. Only close following of the fixed and operative costs of ICT security solutions will prove to be a positive driver for cost optimization and improvement of provided services.

REFERENCES

[1] http://www.investopedia.com/terms/o/offshore.asp (15.12.2012.)
[2] http://www.offshoreguides.com/cptron/contact_engineering_construction.htm (15.12.2012.)
[3] "SENTINEL Auto Acquire Antenna Mobile Satellite Internet System", Owner's manual, General Dynamics, 4096-745 Rev. F, September 21, 2009.
[4] http://www.bsigroup.com/en-GB/iso-27001-information-security/ (15.12.2012.)
[5] Westney, E. Richard, “Managing the Cost & Schedule Risk of Offshore Development Projects”, Westney Project Services, Inc., Offshore Technology Conference, Houston, Texas, 30 April-3 May 2001.
[6] "Risk Management in International ICT Project Management",Global Sustainable Information and Communication Technology Management, University of East London, 2012., p. 12
[7] Armbrust, M; Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Zaharia, "A view of cloud computing.". Communication of the ACM 53 (4), 2010., p. 50.–58.

ICT Contribution to the Economic Development of some SEE Countries in Transition

Marijana Vidas-Bubanja, PhD[1]
Belgrade Business School and Faculty for Trade and Banking, Alpha University, Belgrade
Serbia
e-mail: marijana.bubanja@bbs.edu.rs

Abstract - The economic downturn has posed significant challenges to governments regarding the most appropriate policies to tackle the crisis. This paper analyzes the role of ICT in the economic development of South Eastern European countries focusing on ICT impact on GDP growth, unemployment, the organization and work of companies and state administration.

I. INTRODUCTION

During the past decade, south-eastern Europe (SEE) has undergone a dramatic transformation. The extent of the progress in economic development, democratic reforms, regional cooperation, and integration into global economic and financial markets was unthinkable even 10 years ago and is unprecedented in the region's modern history. For this analysis, SEE covers Albania, Bosnia and Herzegovina, Croatia, FYR Macedonia, Moldova, Montenegro, and Serbia. Despite their many differences, these countries share certain common characteristics: they are all transition countries; they are in the various stages of becoming members of the EU; and they are recipient members of the eSEEurope Initiative established by the Stability Pact for South Eastern Europe.

The financial crisis that began to affect western markets in the second half of 2007 took a while to be felt in SEE, but by the fourth quarter of 2008 it was clear that this region would also face a major economic slump. For all Balkan countries the period starting with 2009 has proved to be a difficult one. Growth in SEE countries in transition slowed down markedly, as a result of the widening spillovers from the Eurozone crisis.

This paper analyzes the possibilities offered by information-communication technologies (ICT) in overcoming the stagnant trends and supporting further economic development of SEE countries in transition.

II. ICT IN THE GROWTH MODEL OF SEE COUNTRIES

After the eurozone sovereign debt crisis deteriorated in the summer of 2011, growth prospects have again weakened and the economic outlook has worsened for SEE countries in transition. As a consequence, SEE governments have been faced with stagnant or decreasing growth rates, rising unemployment and worsening living standard of their citizens. Exports and capital inflows declined, causing difficult fiscal challenges and rising levels of public debt. This reflects an increasing awareness that the growth model on which much of the region had relied in the pre-crisis period, based on cheap inflows of foreign capital to fund credit booms, was inherently risky and unsustainable, and that developing local sources of funds and greater lending in local currency could lead to "safer" growth in the future [1].

TABLE I. ICT - SOLUTION FOR ECONOMIC CHALLENGES IN SEE

ICT as a part of solution for a number of problems and challenges for SEE	**SEE suffer from these key barriers:**	**Policy attention should be on:**
SEE needs new source of growth	Lack of policy attention	ICT should be a key priority for SEE
A severe lack of productivity, lack of competition and growth	Businesses find it difficult to operate in the new ICT based manner, markets are closed and monopolized	ICT transformation of companies and liberalization of markets
Low level of innovation	Decreasing R&D activities, number of researchers, brain drain,	Increase investment in education, research and development especially in ICT
High unemployment levels	Jobs lost in the process of privatization, young unemployment	Active labor market policies, skills, life-long learning and education
Inefficient and expensive state	Low productivity of public sector	Support e-government development
Transforming the SEE to a true knowledge economy	Lack of awareness about ICT potentials and risks	Create ICT supportive environment

Source: adapted by author according to [2].

Also, it was evident that the changed growth model in SEE countries needs new sources of growth that will enable

[1] This paper is the result of the research realized under the project 179038 –„ Modelling of the development and integration of Serbia into world trends in the light of the economic, social and political commotions", Ministry of science and education of the Rep. of Serbia.

further reform, modernization and innovation in order to boost growth rates towards their long-term potential. Opportunities offered by ICT in main areas in which SEE countries are facing stagnation or decreasing trends are in focus, as summarized in Table 1[2].

III. SEE COUNTRIES NEED NEW SOURCE OF GROWTH

GDP growth in the SEE transition region is expected to slow down substantially to 0.5 per cent in 2012 and 1.7 per cent in 2013, from 2.5 per cent in 2011 [3]. In the short term expectations are not very optimistic. Real activity in the Eurozone will continue to suffer both due to fiscal contraction and credit decline, although a full scale credit crunch in the single currency area should be avoided through a generally supportive policy stance by the ECB. This will have further negative impact to the SEE transition region's exports to the Eurozone, as well as availability of finance for the region's banks and therefore credit growth.

TABLE II. GDP GROWTH (AVERAGE -YEAR OVER YEAR PERCENT CHANGE)

	2008.	2009.	2010.	2011.	2012.F	2013.F
Albania	7.5	3.3	3.3	3.1	1.2	1.4
BIH	5.7	-2.8	0.6	1.8	0.3	0.6
Croatia	2.2	-6.0	-1.4	0.0	-1.2	1.5
FRY Maced.	5.0	-1.0	2.9	3.3	0.5	2.0
Monten.	6.9	-5.7	2.5	2.7	0.4	0.8
Mold	7.8	-6.0	4.1	5.2	2.5	4.0
Serbia	3.8	-3.5	1.0	1.6	0.1	1.7

F= forecast, Source: [1 and 3].

In the case of SEE, the fall in output has numerous causes, but three explanations stand out as particularly important. First, external demand from the main export markets has dropped sharply, especially for some key commodities that are now produced in the region. Second, access to finance became much more difficult or expensive, which is a serious impediment to the operation of businesses in the region. And third, the volume of remittances has fallen slightly, which in turn is depressing domestic demand as well as hindering the development of small businesses [4].

The drop in economic activity has had a significant impact on other macroeconomic variables, such as inflation, government deficits, and current account deficits. It is also expected to have a serious negative effect on unemployment and poverty in the forthcoming months, raising the possibility of social unrest [4].

It is also clear that although the integration process has brought enormous benefits to the region in the past decade, the downside is that crises that originate elsewhere will affect the region, no matter what actions are taken to mitigate these effects.

In order to overcome stagnant and decreasing economic growth, SEE countries have to consider the possibilities offered by ICTs as they now play a catalytic role in several aspects of economic and social life. In fact, technological progress has been driving economic growth since 1990. Over the past 15 years ICT and the Internet have delivered substantial economic growth accounting for 10 percent of GDP growth [5]. The ICT sector—infrastructure and networks, IT-related services and media—is regarded as an engine of growth and a source of innovation, with some of the highest growth in productivity. The sector also raises productivity throughout the economy by increasing efficiency across sectors. ICT and Internet maturity correlates with wealth creation and will remain the biggest drivers of global economic growth over the coming decades.

Based on this, SEE countries growth model need the following two elements: 1) the central role of ICTs in the economy, that is, to enable productive and innovative use of ICTs across the whole economy and society, and 2) the direct impact on economy and growth of the ICT sector itself (see Figure 1.)[5].

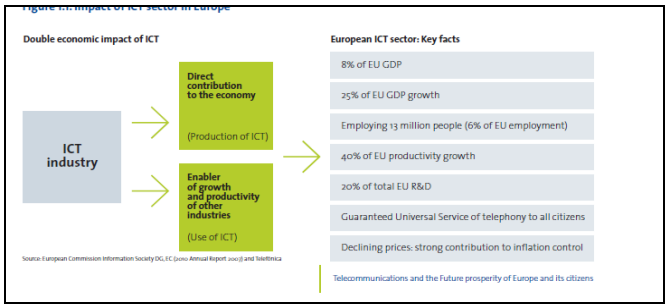


Figure 1. Impact of ICT sector, Source: [2]

As the main policy drivers for economic impact of the Internet and ICT, SEE countries should focus on "digital infrastructure", "digital readiness", and "digital content & services". Digital Readiness supports e-skills, e-awareness, ICT equipment, where as Digital Infrastructure comprises elements such as penetration, coverage, and bandwidth. On the other hand, activities in the area of "digital content& services" stress the impact of a well-functioning market, providing incentives for innovation in the services layer [2].

IV. THE IMPORTANCE OF ICT FOR SEE COMPANIES AND WORKERS

At the firm level, the response to the crisis has been broadly as expected. Many SEE enterprises have had to cut back on production and lay off workers. Others have resorted to wage cuts or freezes in an effort to contain costs. The big metals producers in the region – US Steel in Serbia, ArcelorMittal in Bosnia and Herzegovina, KAP in Montenegro and Silmak ferro-alloy furnace in FYR Macedonia for example – all had to cut production drastically for a period, and some firms resorted to introducing a reduced working week, as a way of retaining staff and minimizing job losses, presumably in the hope of a global upturn [4].

High energy prices and persisting difficulties in access to finance, due to uncertainties about the prospects of the European economy and the turmoil related to the debt crisis, are also adversely affecting the dynamics of the industrial production recovery in SEE countries.

Closely connected to the negative year-on-year growth in industrial output, monthly exports are volatile series in SEE countries. In order to get a better understanding of why exports have dropped, it is important to know *what* countries

are exporting, and *where* the exports are going to. In recent years, many countries of the region have developed a specialization in certain key industries. The steel sector is particularly important in Bosnia and Herzegovina, FYR Macedonia, and Serbia. Aluminium is the main export earner in Bosnia and Herzegovina and Montenegro. Tourism is an important part of the Croatian and Montenegrin economies. All of these industries have been particularly hit by the global recession, and this is a major factor behind the decline in output in the region. Exposure to the EU-27 has offered little protection; exports have tended to perform badly, whether or not a country exports mainly to the European Union (such as FYR Macedonia, where 78 per cent of exports are EU-bound or Serbia, where 54 per cent of exports go to EU countries) [4].

In order to face decreasing production and export challenges caused by inside or outside factors, SEE companies can use ICTs to reduce business costs, improve internal management, and expand access to new technologies and information on market opportunities. ICTs can also help optimize supply chains, making it easier to get goods and services to the market.

Empirical evidence shows that, on average, firms that use ICTs grow faster, invest more, and are more productive and profitable than those that do not. According to a World Bank study, sales growth and profitability are 3.4 and 5.1 percentage points higher, respectively, among firms that use ICTs effectively in their businesses [5].

For these SEE companies ICT innovation should be defined as introducing new or improved goods and services into the market—or as finding new ways to organize production or develop new markets with the help of ICTs. At the same time, it is important to notice that acquiring the necessary skills and embracing the digital opportunities is one thing. But integrating these into business and taking ownership of digital visions and decisions is a question of mind-set and leadership. Thinking digital will give SEE companies a competitive edge. They can start by procuring and invoicing digitally instead of using paper and making sure their business is visible and accessible on the Internet.

Rejuvenating traditional activities has been the ICT and Internet's main impact and that is very important for the structure of the majority of SEE economies. The ICT has enabled fundamental business transformation that spans the entire value chain in virtually all sectors and types of companies. This shift includes wholesale changes not only in how products are brought and sold, but also in how product and services are designed, produced and distributed.

Another important Internet effect for the SEE economies is connected with SMEs as such companies are fundamental for these countries. The Internet's biggest impact on SMEs has been as a great leveler, making it possible for a small firm to be a global company from day one, with the reach and capabilities that once only large companies could possess. SMEs that use ICT and the Internet brought in more than twice as much revenue through exports as a percentage of total sales that those that use the Internet sparingly. These web-knowledgeable SMEs create more than twice as many jobs as companies that are not heavy Internet users. This is true across sectors from retail to manufacturing [5].

As SEE countries have the highest unemployment and poverty rates in Europe, it is worth noticing that the Internet and ICT can also serve as a powerful catalyst for job creation. McKinsey`s global SME survey found 2.6 jobs were created by ICT and the Internet for every one destroyed [5].

TABLE III. UNEMPLOYMENT %, 2011

Albania	BIH	Croatia	FYR Macedonia	Monte negro	Moldova	Serbia
13.3	27.6	18.0	31.6	12.7	7.2	26.1

Source: [3 - National authorities via CEIC data service].

Moreover, what growth there was in SEE countries during the nascent recovery in 2010-11 was largely jobless. At about 23 percent, the average unemployment rate in SEE is more than twice the Western Europe average, and is highly concentrated among youth and long-term unemployed, with a devastating impact on human capital. Pre-crisis poverty reduction gains are being reversed, and after large shocks and depleted household buffers and savings, the middle class has become more vulnerable. With growth prospects much more moderate than before the crisis and with social pressures high, it is urgent that SEE country governments adopt a more ambitious structural reform agenda for growth and jobs and use potentials offered by ICT in this area [6].

The high unemployment underscores the need for the countries in the region not only to pursue pro-growth policies, but also to improve labor market flexibility and mobility through reforming, among others, the labor regulations to make their economies more contestable. [6].

Most welcome in SEE are initiatives that facilitate the development of sectors with the highest employment potential, including the low-carbon, resource-efficient economy ("green jobs"), health and social sectors ("white jobs") and the digital economy. In fact, ICT supported production and IT-based services have the potential to create jobs and advance the economic goals of SEE countries. Direct and indirect employment opportunities, particularly for youth and women created in this way are a chance for SEE countries to reduce high unemployment rates [5].

Increasing employment levels and enhancing labor productivity are two key sources of growth, together with capital investment and innovation. Education and life-long learning activities are fundamental in this respect. Skills are essential to grow innovative IT-based companies that create jobs and generate income in SEE countries. For SEE economies in transition, it is important to focus on the quality of education in close partnership with the private sector and education institutions. Regional centers of excellence and distance learning can play a role.

SEE government should also create ICT innovation ecosystems across the economy by supporting policies that promote ICT innovation, developing local business incubation ecosystems, and promoting light innovation as a

bottom-up and user-centric approach to drive ICT innovation across the economy.

The role of innovation in the SEE countries, however, remains very limited. Public sector expenditure on R&D was well below 1 percent of GDP in all SEE countries, except Montenegro (1.1 percent of GDP). Businesses in SEE spent less on innovation than any other region in Europe. Similarly, tertiary education attainment among the population aged 30-34 was the lowest in Europe, amounting to 15 percent in comparison to around 30 in the EU15 Continental; and above 35 in the EU15 North. This is also reflected in the exports of innovation products. Reflecting this limited role of innovation, SEE's medium and high-tech products and knowledge intensive services accounted for 17 percent of total goods exports in 2010, the lowest ratio in Europe. Consequently, SEE receives marginal revenues from royalties and license fees from abroad (0.06 percent of GDP) [6].

V. USE OF ICT IN SEE GOVERNMENTS

SEE governments themselves could benefit from increasing their own use of the ICT and the Internet to transform the efficiency and effectiveness of the public sector. In many countries parts of the public sector continue to suffer from low productivity and the adoption of technology in the public sector often lags well behind what has been accomplished in the private sector. This can lead to savings in public expenses and reduce public debt [2].

This is especially important, having in mind that SEE faces one of the worst demographic outlooks in the next four decades, with significant population aging, a shrinking labor force (by a quarter), and outward migration–all imposing major stresses on social protection and pension systems, health care costs, and public finances. The pension systems in SEE are likely to be burdened heavily, as fewer workers are called on to support many more elderly reflecting higher dependency ratios. Healthcare costs are also likely to increase significantly, as the countries grapple with healthcare for the elderly. This means that big fiscal adjustments are needed to address both current imbalances and future costs related to the aging population, including reforms of the pensions and social transfers systems, and healthcare. Improving efficiency and effectiveness of public spending will be critical to addressing these challenges. Reforming social protection and health care today will be essential for the countries in the region to be better prepared for the demographic transition and aging population of tomorrow; and these reforms in turn are critical to making fiscal adjustment needed to keep public finances, including public debt, in order.

Governments in SEE countries can use ICTs to transform public service delivery across sectors—health, education, social protection, justice, agriculture, water, energy, and transport—both central and local. ICTs offer an opportunity to revolutionize public service management by capturing information efficiently in shared databases accessible to all government agencies (such as a national database of electronic identification of citizens). ICTs help governments to simplify administrative procedures and share services (such as citizen authentication and payment systems). And governments can pool resources, reduce costs, adopt common standards for information exchanges, and perform government functions more efficiently (with faster response time and fewer errors).

Upgrading economic management and public administration back-end functions through IT systems is a core activity of the public sector reform, mostly through public management information systems. These typically include customs and taxation, electronic procurement, human resource management (including payroll), financial management (accounting, budgeting, control, and treasury functions), and social security management. Governments may, however, need to develop capacity to manage the transition.

VI DIGITAL ECONOMY IN SEE AS A PRECONDITION FOR EU MEMBERSHIP

According to EU, the digital economy is a major source of growth and innovation. The digital economy is an important priority for the Barosso II Commission and it is at the core of the new Europe 2020 strategy. Under the heading Digital Society, the initiative "A digital agenda for Europe" is one of the three proposed flagship initiatives to generate smart growth in Europe. The purpose is to speed up the roll-out of high-speed Internet and it specifically sets out to"reap the benefits of a digital single market (DSM) for households and firms" [7].

The DSM can be seen as the fourth wave of economic integration of the EU market. After the 1992 Single Market Programme, EMU and the common currency (Euro) in 1999/2002 and the 2007 Services Directive, the creation of a single market for digital goods and services can be seen as a fourth step of the European Union to foster growth, jobs and innovation. A Digital Single Market which ensures the free movement of information and knowledge could bring benefits to European consumers and business, because knowledge and information move online and are the key drivers of modern growth.

EU estimate that at least 4 percent additional GDP (EU27) can be gained in the longer run by stimulating further adoption of ICT and digital services through the creation of a DSM. This will also imply job creation. One estimate shows that in Germany alone, the improvement of digital infrastructure will trigger innovation and growth leading to an additional 427,000 jobs over the period 2015-2020. Building a Digital Single Market in Europe requires an ambitious policy response, of all Member States and potential members as well [7].

As accession to the EU is high on the agenda of countries in SEE[2], they have to develop information society and implement ICT in their economies in accordance with European models and standards.

[2] All SEE countries except Moldova, have signed the Stability and Association Agreements with the European Union. Croatia completed the accession negotiations and stands ready to become full EU Member in July 2013, Montenegro opened negotiations, Serbia received a candidate status.

Since 2000, the Western Balkans countries have achieved a remarkable progress in Information Society development through regional cooperation within the eSEE Initiative. The Initiative was formally established in 2002 under the umbrella of the Stability Pact by signing of the eSEE Agenda at the ministerial conference in Belgrade. eSEE Agenda aimed at creating legal and institutional framework for the development of information society in the South East Europe. Regional Cooperation Council's (RCC) succeeding of the Stability Pact reaffirmed activities of the eSEE Initiative at the ministerial level by the signature of eSEE Agenda Plus [8] in Sarajevo in 2007. The eSEE Agenda Plus widened the scope of eSEE Agenda activities defining a large list of ambitious objectives aiming to establish a strong regional market for electronic communication services, to create information society infrastructure in the public sector, encourage development of rich digital content and innovative services and promote eInclusion, eParticipation and eDemocracy.

The global economic crisis had its impact on the implementation of the eSEE Agenda Plus. In November 2011 ministers responsible for the development of information society in the region met in Tirana to reaffirm their commitment to the implementation of the eSEE Agenda Plus, and reassess the deadlines for reaching the targets taking into account the reduced resources [9].

According to the eSEE Initaitive report of September 2011 [10] Internet penetration in households in the region ranges from 31% in Bosnia and Herzegovina to 55% in Croatia. Broadband penetration is still very low but is increasing all over the region and has reached the level of 40% in Montenegro, and 37% in Macedonia (FYR). All countries have established at least the basic benchmarking of availability of eGovernment services. Most elementary and high schools are connected to the Internet and equipped with computer labs with the number of pupils per computer ranging from 32 in Albania down to 1.5 in Macedonia. However, Albania and Bosnia and Herzegovina, still have to establish their academic and research networks. Countries do not track funding of ICT research, research professionals or intellectual capacity in IT industry, which are obligations foreseen in the eSEE Agenda Plus. On the other hand, some countries have introduced favorable tax schemes for IT sector (Moldova), and have reduced tax rates for computer equipment (Macedonia, Montenegro and Serbia). In the majority of the countries there are also activities for stimulating business incubators, techno-parks and business start-up centers.

While overall eSEE Initiative countries do show a remarkable progress, most are still missing the ambitious deadlines set in the eSEE Agenda Plus. Progress in improving SEE countries‘capacity to absorb and use ICT technologies throughout their economies has been slower due to the weakness of domestic skills and competencies, as well as rigidities in the regulatory environment, which prevent the creation and expansion of innovative firms. Despite their technologically sophisticated cities and world-class firms, many SEE countries fall far below the technological achievement in more ICT-savvy economies.

This is also reflected in countries' performance in global rankings. The World Economic Forum and INSEAD have been monitoring the global network readiness since 2002. The aim of the Network readiness index (NRI) is to measure the “…degree of preparedness of a nation or community to participate in and benefit from ICT developments”

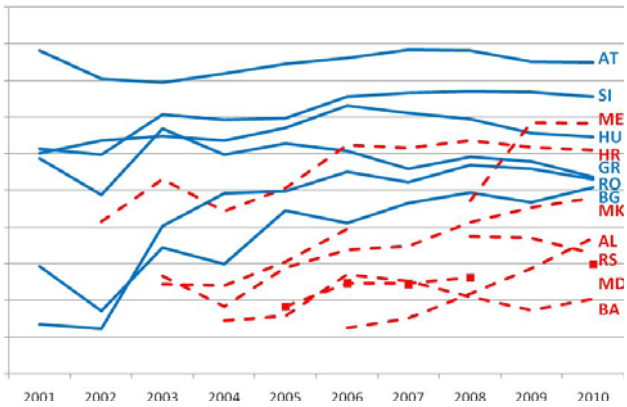


Figure 2. World Economic Forum Network Readiness Index-Position in the renking of the selected countries
Percentage represents proportion of all countries ranking below each country for the referent year. Solid blue lines represent EU member states, and dashed red lines eSEE Initiative members. Based on data from [11-20] undertaken from [9].

Figure 2 presents the relative position of eSEE and neighboring countries in the global Network Readiness Index ranking during the last decade [11-21]. Each country line represents the proportion of countries ranking below it in the given year. Thus, Austria ranks steadily among the top 20% of the countries, and Bosnia and Herzegovina remains among the bottom 30%. The positive trend seen until 2009 has continued for Macedonia and Albania, but leveled off for Montenegro, without jeopardizing its leading position in the region. Overall, with the exception of Croatia and Montenegro, the region remains firmly within the lower half of the ranking [9].

VII CONCLUSION

The economic downturn has posed significant challenges to governments regarding the most appropriate policies to tackle the crisis. In major advanced economies, the key policy requirements are to address medium-term fiscal imbalances and to repair and reform financial systems, while sustaining the recovery. In transition and developing SEE economies, near-term policy should focus on responding to moderating domestic growth and to slowing external demand from advanced economies.

Yet even with the difficult short-term situation, SEE countries now have a historic opportunity to board the European “convergence train” and over the long term reduce their per capita income gap with developed European Union countries. All earlier entrants were able to “catch up quickly.” In principle, the same “convergence train” is now pulling into the EU candidate countries in SEE; but these gains are not automatic, they will materialize only if country policies and reforms facilitate them. The long-term SEE structural reform agenda must leverage greater trade and financial integration and reform labor markets and the public sector [6]. This structural reform in SEE countries should also include the Internet’s and ICT`s huge potentials for

generating economic and social benefits.

Government, policy makers and business in SEE region must recognize and embrace the enormous opportunities offered by ICT and the Internet by undertaking the following activities [5]:

- *Firstly, foster competition (promote deregulation)* – a competitive business environment benefits users and lets innovators and entrepreneurs thrive. Countries that make their markets more open and competitive achieve greater productivity. Such competition ensures that the most innovative and productive companies create more attractive products and services for users and gain market share at the expense of the less productive.
- *Secondly*, *encourage innovation* – an environment that encourages innovation and entrepreneurship is vital to capture ICT-based growth. Features of such environment include: 1) access to start-up capital, 2) protection of intellectual property rights, 3) support for research and development and 4) the availability of world-class "common research" conducted by universities and government-funded teams.
- *Thirdly*, *develop human capital* – talented and educated human capital is the key to success in digital economy. This capital is attracted, generated and nurtured by universities, corporate R&D centers. Developing countries should build policies that both enhance the technological absorptive capacity in firms and accelerate the general population's acquisition of technologies.
- *Fourthly, build infrastructure* – Internet infrastructure is a prerequisite for economic and social growth and development. It is a platform upon which users and organizations experience the Internet and upon which entrepreneurs and businesses innovate. As the Internet continues to expand in an era of cloud computing, big data and other innovations that engage users more deeply, infrastructure requirements will continue to grow. Concern about privacy, online fraud, identity theft and hacking of sensitive materials and databases should capture significant attention.

SEE government and business should also be aware of the following facts. The developmental impacts that result from proper ICT implementation are not the product of changes in technology, connectivity or access alone, but of the conjunction of those changes with human development activity— the choices people, businesses and Governments make, the ways in which they connect with new opportunities, the extent to which they seek to change their prospects in the longer term.

In order to achieve joint goals of using opportunities offered by ICT and the Internet for economic growth, enrich lives, build business and give customers the enhanced choices, collaboration across the private, public and nonprofit sectors in SEE countries is required. Public –private dialog should result in the creation of a constructive environment in which the benefits of the Internet and ICT can be better understood and properly used in the economy as a whole, enabling more competitive work of domestic enterprises and higher growth rates of national economy [21].

As more people, companies, government institutions in SEE region get connected, it will trigger massive implications for productivity and will open all kinds of new opportunities for countries and individuals. Regional cooperation through the experiences of eSEE Initiative proved to be a good way to achieve information society development in short term and in a more efficient manner.

The challenge is even more relevant now, because many countries are rethinking their growth and development models due to global economic crisis. This is an area in which technology—and particularly ICT—can play a crucial role, even if there are no solutions that guarantee success.

REFERENCES

[1] EBRD, Transition Report 2011, London, 2011.

[2] European Policy Centre, The Economic Impact of a European Digital Single Market, Final report, Copenhagen Economics, March 2010.

[3] EBRD Office of the Chief Economist, Regional Economic Prospects in EBRD Countries of Operations: July, 2012.

[4] Sanfey, P., South-eastern Europe: lessons from the global economic crisis, WP 113, February, 2010.

[5] Pelissie du Rausa, M. Manyika, J. et al, Internet matters: The Net`s sweeping impact on growth, jobs and prosperity, McKinsey Global Institute, May, 2011.

[6] The World Bank, South-east Europe regular Economic Report, June 5, 2012.

[7] European Commisson, Annual Growth Survey 2012, COM(2011) 815 final, VOL. 1/5, Brussels, 23.11.2011.

[8] eSEE Agenda Plus, Sarajevo 2007 (http://www.eseeinitiative.org accessed 31 January 2013)

[9] Šimić, D., ICT R&D Challenges for the Western Balkans, Budapest, November 2011.

[10] Matrix of Fulfillment of eSEE Agenda Plus, September 2011

[11] Kirkman G, Cornelius P, Sachs J, Schwab K (eds.) The Global Information Technology Report 2001–2002: Readiness for the Networked World. New York: Oxford University Press, 2002.

[12] Dutta S, Lanvin B, Paua F (eds.) The Global Information Technology Report 2002-2003 - Readiness for the Networked World. New York: Oxford University Press, 2003.

[13] Dutta S, Lanvin B, Paua F (eds.): The Global Information Technology Report 2003-2004 – Towards an Equitable Information Society. New York: Oxford University Press, 2004.

[14] Dutta S, Lopez-Claros A (eds.) The Global Information Technology Report 2004-2005: Efficiency in an Increasingly Connected World. New York: Palgrave Macmillan, 2005.

[15] Dutta S, Mia I, Lopez-Claros A (eds.) The Global Information Technology Report 2005-2006: Leveraging ICT for Development. New York: Palgrave Macmillan, 2006.

[16] Dutta S, Mia I (eds.) The Global Information Technology Report 2006-2007: Hampshire: Palgrave McMillan,2007.

[17] Dutta S, Mia I (eds.) The Global Information Technology Report 2007–2008: Hampshire: Palgrave McMillan, 2008.

[18] Dutta S, Mia I (eds.) The Global Information Technology Report 2008-2009. Geneva: World Economic Forum, 2009.

[19] Dutta S, Mia I: The Global Information Technology Report 2009-2010,World Economic Forum, 2010.

[20] Dutta S, Mia I: The Global Information Technology Report 2010-2011 – Transformations 2.0 10th Anniversary Edition. Geneva: World Economic Forum, 2011.

[21] Vidas-Bubanja, M., "Information society as a key factor of the Serbian development strategy" -Informaciono društvo kao ključni faktor strategije ekonomskog razvoja Srbije, Ekonomski vidici, br.2, 2011, pg. 247-260.

Optimizacija ICT kapaciteta po filozofiji modela JIT (UNV)

Zaim Buljubašić, BH Telecom, Aleja Alije Izetbegovića 29, 75000 Tuzla, Bosnia & Herzegovina, e-mail: zaim.buljubasic@bhtelecom.ba
Izudinka Kapetanović Serdarević, BH Telecom, Aleja Alije Izetbegovića 29, 75000 Tuzla, Bosnia & Herzegovina, e-mail: izudinka.kapetanovic@bhtelecom.ba
Nermin Buljubašić, BH Telecom, Aleja Alije Izetbegovića 29, 75000 Tuzla, Bosnia & Herzegovina, e-mail: nermin.buljubasic@bhtelecom.ba

Sažetak - Integracija baza podataka, odnosno aplikacija koje podupiru pojedine faze rada treba biti sinkronizirana s integracijom poslovnih faza rada u procese, kao i integracijom samih procesa. Informacije se ne daju, nego se povlače po kanban sistemu tako da ni jedna faza procesa nije zatrpana „zalihama" informacija zbog neizvršenih poslova. Tržište, odnosno prodaja određuje kroz terminski plan količinu potrebnih usluga, a samim tim količinu i kvalitetu potrebne podrške informacijsko komunikacijske tehnologije. Visok stupanj neiskorištenosti kapaciteta u informacijsko komunikacijskoj tehnologiji obavezuje racionalniji pristup u dizajniranju arhitekture e-poslovanja. U suvremenom svijetu informacija ima tretman robe, ona se pojavljuje na tržištu ponuđena kroz formu usluge od strane ponuđača-davatelja usluge i kupaca-korisnika usluge. Sredstvo koje omogućava kupoprodajne transakcije je novac i u principu razmjena se odvija uz djelovanje klasičnih tržišnih zakonitosti. Pretpostavimo li da svi resursi koji u određenom periodu nisu u poslovnoj funkciji čine neiskorištene rezerve ili zalihe, te je neophodno učiniti maksimum napora kako bi se one minimizirale, jer već je dokazano na primjeru „klasičnih zaliha" koliko su one opterećenje za uspješno poslovanje. U ovom trenutku rješenje treba tražiti u optimiziranju kapaciteta, arhitekture e-poslovanja i primjene cloud-a.

Ključne riječi: IT resursi, stupanj iskorištenosti, tradicionalno računarstvo, oblačno računarstvo, poslužitelji.

I. UVOD

Uvijek, a pogotovo u suvremenom dobu kompletna društva, države pa i same kompanije susreću se s ogromnim problemima u poslovanju koji su u dobroj mjeri izazvani svjetskom ekonomskom krizom, značajnim problemima u nedostatnim budžetima u financiranju osnovnih obima proizvodnje s minimalnim stopama rasta i razvoja bazičnih djelatnosti, te očuvanja tržišnih pozicija i nivoa konkurentnosti bar na postojećoj razini. Svi segmenti društva su manje više potpuno svjesni da funkcionalnost i efikasnost svojih poslovnih procesa mogu osigurati samo uz značajnu potporu ICT (*Information and communications technology*). Kompanije već imaju problema sa osiguranjem potrebnog obima sredstava za financiranje osnovne djelatnosti, a primorane su ulagati značajna sredstva u najnovija i adekvatna rješenja u upravljanju, planiranju i nadzoru poslovnih procesa koji moraju biti još učinkovitiji kako bi se mogla osigurati veća konkurentnost i značajniji iskorak u razvoju. Problemi se još više umnožavaju doda li se tome činjenica da nivo ulaganja u ICT nije uopšte zanemariv, te da je nivo iskorištenosti kapaciteta u ovom segmentu dosta niska (u računarstvu od 5-10%). Jasno je da se nameće potreba drugačijeg pristupa u oblikovanju pojedinih poslovnih procesa, s jedne strane njihovom usitnjavanju pa ponovnom integriranju, a s druge strane drugačijem dizajniranju ukupne informatičke potpore. Informacije u svakom svom pojavnom obliku u današnje vrijeme imaju karakter robe, pa se i o „jalovom" skladištenju informacija treba povesti računa, ukoliko se žele premašiti navedene ekonomske teškoće. Kada pričamo o obimu ulaganja u ICT treba uzeti u obzir i nužna ulaganja u kvalitetne građevinske objekte pristojne površine, antistatičkim podovima, uređajima za neprekidno napajanje, odgovarajućim klimatizerima i još dosta elemenata koji ne čine direktnu informatičku ili komunikativnu opremu. U ovom radu se pošlo i od realne pretpostavke da sve ono za što su utrošena određena novčana sredstva, a nije u tom trenutku u funkciji odvijanja poslovnih procesa, čini zalihe, te će kao takve biti i tretirane. Kakav su negativan efekt imale i još uvijek u nekim procesima imaju materijalne zalihe po ukupnost rezultata poslovanja manje više je već dobro poznato, pa se i kraj prošlog stoljeća između ostalog karakterizira i po uvođenju različitih modela upravljanja zalihama, a sve sa ciljem njihovog optimiziranja ili bolje rečeno njihovog minimiziranja. Saznanja do kojih se došlo u primjeni modela upravljanja zalihama kod materijalnih resursa u dobroj mjeri su primjenjiva i u sinkronizaciji korištenja IT (*Information technology*) resursa, kako bi se premašio problem niskog stupnja iskorištenost ICT kapaciteta.

II. TRENUTNO STANJE PRIMJENE IT U POSLOVNOJ PRAKSI

Značaj primjene ICT u optimizaciji poslovnih procesa, u kolikoj mjeri primjena ICT čini te poslovne procese efikasnijim i podiže ukupni nivo produktivnosti ne može se reći. U realnom životu i praksi ta primjena, kada su u pitanju poslovne mreže putem kojih se upravlja uslužno-

proizvodnim procesima je doživjela potrebnu afirmaciju. Prema nekim istraživanjima uslužne djelatnosti su otišle korak dalje od proizvodnih djelatnosti, ali to treba prije svega zahvaliti utvrđenoj korelaciji između višeg nivoa obrazovanosti uposlenika u uslužnim djelatnostima. Ono što čini dodatni problem je potreba za dodatnim kapitalom za investiranje u ICT o čemu je bilo riječi u uvodu i svakako niskim stupnjem iskorištenosti kapaciteta IT. Što se ovdje može učiniti? Vrijeme *mainframe* računarstva je poznato po vrlo visokim troškovima prilikom uspostave sustava, te potrebe za zapošljavanjem specijaliziranih djelatnika koji su mogli upravljati tim sustavima. Međutim, ovakvo računarstvo su si mogli priuštiti samo velike organizacije s centraliziranim IT sustavima koji su mogli skupiti dovoljno radnih zahtjeva kako bi opravdali investiciju u sustav. Zbog visokih troškova, upotreba resursa se uglavnom vršila po prioritetima, te su korisnički zahtjevi čekali u redu i bili obrađeni tek onda kada su resursi bili dostupni. Kako se zahtijevana računarska snaga (mjerena u MIPSovima – milijunima računarskih instrukcija u sekundi) povećavala, troškovi su drastično padali. S napretkom miniračunala i kasnije klijent/server tehnologija, značajno se smanjio minimalni resurs kojeg je potrebno (za)kupiti, a resursi su postali jednostavniji za upravljanje i održavanje. Ovakva modularizacija resursa značajno je smanjila barijere u pružanju IT usluga, povećavajući dostupnost krajnjim korisnicima. No i dalje je ostao problem učinkovitog iskorištavanja resursa koji je rezultirao datacentrima koji su bili puni poslužitelja za razne namjene, a njihova iskorištenost bila je teksa 5-10% kapaciteta [1].

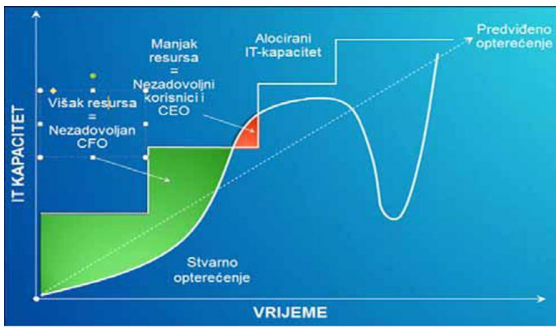
Međutim, za mnoge kompanije, informatička infrastruktura blisko je povezana s centralnom oblašću njihovog poslovanja i eksternalizacija računarskih kapaciteta, te bi predstavljala veliki poslovni rizik. Sigurnost, pouzdanost, performanse i usuglašenost sa standardima predstavljaju najvažnija pitanja za rukovodioce informatičkih odjeljenja, prilikom procjenjivanja novih tehnologija. Na primer, biotehnološka kompanija možda neće željeti da vrši modeliranje novih lijekova izvan vlastitih prostora, već će za naprednu simulaciju novih jedinjenja iskoristiti vlastite servere. Kako bi iskoristila svu fleksibilnost i efikasnost arhitekture *cloud computing*-a, ta kompanija može razviti internu, privatnu *cloud computing* mrežu nad kojom će imati veću kontrolu i ostvariti veću sigurnost računarskih resursa. Kod tradicionalnih sustava, slika 1., opterećenje se obično predviđa kao pravac koji pokazuje povećanje zahtjeva na resurse kako vrijeme protječe. Kako bi zadovoljili potrebu za IT-resursima, u tradicionalne se sustave obavljaju ulaganja koja su više diskretne prirode koja su na slici 1. prikazana kao stepenice koje nastaju kao rezultat predviđanja potreba za IT-resursima, kao i zbog činjenice da je potrebno neko vrijeme da se oprema nabavi, osobito u velikim organizacijama. Kao posljedica takve „nabave“ računarskih resursa dolazi do najmanje tri situacije koje su nepovoljne što se tiče ulaganja ili raspoloživih IT-resursa, zato što je kod tradicionalnih IT-sustava:

1. potrebno veće inicijalno ulaganje u IT-resurse, nego što nam stvarno u tome trenutku treba, kako bi predvidjeli buduće opterećenje/zahtjeve za IT-resursima. Na slici 1. to je slikovito okarakterizirano kao *Višak resursa* - neiskorišten kapital,

2. moguća je situacija kada nedostaje IT-resursa zbog sporije dinamike nabave, odnosno veće dinamike poslovanja, te se može jednome trenutku dogoditi da zahtjevi korisnika za IT-resursima budu veći od onih koji su trenutačno dostupni i onda imamo nezadovoljne korisnike,

3. kada prestane zahtjev za IT-resursima oni se ne mogu „vratiti“ i dalje ostaju kao trošak. Ako je riječ o sezonskim opterećenjima ili jednostavno padu potrebe za IT-resursima oni ostaju i dalje alocirani i stvaraju troškove, kako amortizacije, tako i za održavanje, bez obzira na to što nam više ne trebaju.

Postavlja se još nekolicina otvorenih pitanja: da li nabavljati i izgrađivati vlastiti informacijski centar i izdvojiti dodatna investicijska sredstva u tu svrhu, te u fazi eksploatacije uvažavajući i ostale momente (sigurnost, pouzdanost, itd.) ostvarivati značajne pozitivne ekonomske efekte nego što bi to bilo da se pristupilo cloud computing-u. Treba još posebno naglasiti i visok nivo fiksnih troškova, te nisku stopu iskorištenosti kapaciteta koja nužno determinira vrlo visoke troškove poslovanja.



Slika 1. Ulaganje i raspoloživi resursi kod tradicionalnih IT-sustava [1]

III. CLOUD COMPUTING RACIONALNI INFORMACIJSKI PRISTUP

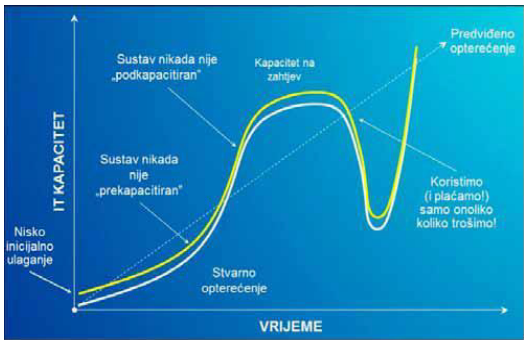
Cloud computing predstavlja “fenomen u nastajanju” [2], odnosno fenomen koji nastupa u trenutku kada suuvjeti za to povoljni i kada brojni faktori to dozvoljavaju – u ovom slučaju, orijentacija ka servisima, virtualizacija i standardizacija računarstva putem interneta. U kombinaciji s globalnom ekonomskom krizom, cloud computing omogućava smanjenje određenih troškova vezanih za informatičke tehnologije. Apstrahirana, visoko skalabilna i kontrolirana računarska infrastruktura koja iznajmljuje aplikacije namìnjene krajnjim korisnicima i čije se usluge

naplaćuju na bazi ostvarene potrošnje [3]. Postoje tri kategorije visokog nivoa usluga *cloud computing*-a: softver u vidu servisa (*software-as-a-service*, SaaS)- softver koji je implementiran u obliku iznajmljenog servisa kome se pristupa putem interneta, platforma u vidu servisa (*platform-as-a-service*, PaaS) - platforme koje mogu biti korištene za realizaciju aplikacija osiguranih od strane klijenata ili partnera provajdera platforme i infrastruktura u vidu servisa (*infrastructure-as-a-service*, IaaS) – računarska infrastruktura, kao što su serveri, skladištenje podataka i umrežavanje, ostvarena u vidu cloud computing-a, obično korištenjem virtualizacije. Infrastruktura u vidu servisa može obezvjeriti pristup serveru, dijelu servera ili skladišnom prostoru, kao što je, na primjer, sistem Backspace Cloud Servers. Amazon-ov Elastic Compute Cloud (EC2) predstavlja primjer platforme u vidu servisa. Amazon omogućava dizajnerima da koriste njegove virtualizirane i udaljene servere za razvoj aplikacija i njihovo iznajmljivanje uz upotrebu modela plaćanja na bazi ostvarene potrošnje. Google-ov App Engine predstavlja još jedan primjer platforme u vidu servisa. Salesforce.com je primjer softvera u vidu servisa. Salesforce.com već skoro deset godina implementira rješenja za menadžment odnosa s klijentima (*Customer Relationship Management*, CRM) uz upotrebu modela softvera u vidu servisa i smatra se da predstavlja jednog od pionira upotrebe modela isporuke u vidu servisa u oblasti računarske tehnologije. Giganti kao što su IBM i Microsoft načinili su vlastite investicije u oblasti usluga cloud computing-a. Uz to, SAP i Oracle tvrde da će sljedeća generacija njihovih poslovnih aplikacija moći da koristi cloud computing i da će biti dostupna u formi softvera u vidu servisa (SaaS). Kompanija 3Tera razvila je grid operativni sistem AppLogic koji omogućava svakom hosting provajderu da ponudi i raznorazne *cloud computing* usluge i da ih povežu u jednu *cloud computing* platformu. Iako Google i Amazon raspolažu ogromnim *on-line* resursima koji su dostupni iz svake točke svijeta na kojoj postoji pristup Internetu, pitanje je da li će ove kompanije moći da budu dovoljno fleksibilne za zahtjeve raznorodnih korisnika. Procjenjuje se da će AppLogic rješenje imati na duži rok veći uspjeh jer omogućava instalaciju i korištenje na raznorodnom hardverskom sistemu, bez obzira na specifične zahtjeve hosting aplikacije. Iza ovog rješenja stoji logično razmišljanje da cloud computing tehnologija treba da se širi poput Interneta koji je dozvolio svakom pojedincu i firmi da sistemu doda svoje jedinstvene mogućnosti. AppLogic upravo to omogućava - kreiranje *cloud computing* sistema nevezanog za neku konkretnu infrastrukturu, što dozvoljava programerima da svoje aplikacije i usluge pripremaju za one arhitekture koju sami odaberu. Kod IT-sustava zasnovanih na računarskom oblaku imamo bolju situaciju kada promatramo potrebu za IT-resursima tijekom vremena:

1. potrebno je nisko inicijalno ulaganje u IT-resurse. Riječ je o tzv. postavljanju (*setup*) sustava u računarskom oblaku. Zakupljuje se onoliko IT-resursa koliko je u tome trenutnu potrebno.

2. kada nam treba više resursa, jednostavno možemo kupiti i alocirati automatski, bez asistencije pružatelja usluge,

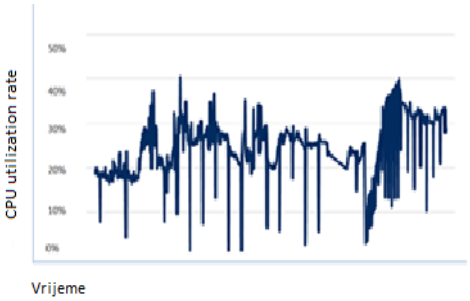
3. kada prestane potreba za određenim dijelom IT-resursa, jednostavno ih prestanemo koristiti: prestankom korištenja plaćamo manje, a samim time smo i konkurentniji i „štedimo“; kada nam dodatni IT-resursi zatrebaju, dostupni su nam.



Slika 2. Dostupnost IT resursa kod *cloud computing*-a

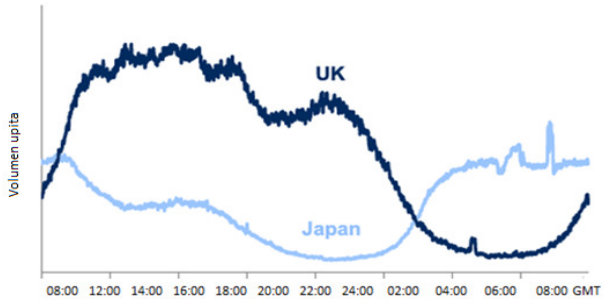
Virtualizacija omogućava da se višestruke aplikacije pokreću na jednom fizičkom poslužitelju unutar posebno prilagođene instance operacijskog sustava. Primarna prednost virtualizacije je mogućnost da manje poslužitelja može podnijeti isti broj radnog opterećenja. Kako ovo utječe na ekonomiju velikih brojeva? Kada bi sva radna opterećenja imala istu iskoristivost, ovo bi se pretvorilo u jednostavno sažimanje potreba s predvidivim opterećenjima. No, u stvarnosti se opterećenja značajno mijenjaju tijekom vremena, često zahtijevajući velike resurse u jednom trenutku, dok u sljedećem ne zahtijevaju gotovo ništa. Ovo otvara značajne mogućnosti za poboljšanje iskoristivosti putem zbrajanja i podjele (*aggregation and diversification*) zahtjeva od strane kupaca. Analiza opterećenja je pokazala da postoje različiti izvori varijabilnosti iskoristivosti. Zanimljivo je pogledati na koji način računarstvo u oblacima može odgovoriti na njih i time smanjiti troškove.

1. **Slučajnost**. Uzorci opterećenja krajnjeg korisnika sadrže određene elemente slučajnosti. Na primjer, ljudi provjeravaju svoju elektroničku poštu u različita vremena kao što je prikazano na slici 3. Kako bi se zadovoljili pragovi *service-level* ugovora, potrebno je ugraditi kapacitet (*buffer*) koji može predvidjeti hoće li pojedini korisnici iste aktivnosti provesti u isto vrijeme. Ako je poslužitelje moguće dodavati prema potrebi, varijabilnost je moguće smanjiti.



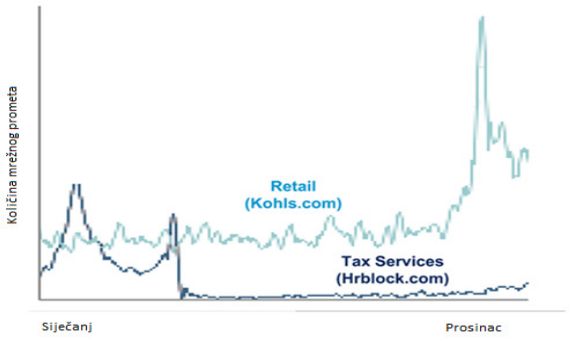
Slika 3. Opterećenje Exchange poslužitelja tijekom vremena (izvor: Microsoft)

2. **Dnevni uzorci**. Postoje predvidivi dnevni uzorci u ponašanju korisnika: servisi koji poslužuju krajnje korisnike (*consumer*) uobičajeno su najopterećeniji krajem dana, dok servisi koji poslužuju poslovne korisnike uobičajeno su najopterećeniji tijekom radnog dana.



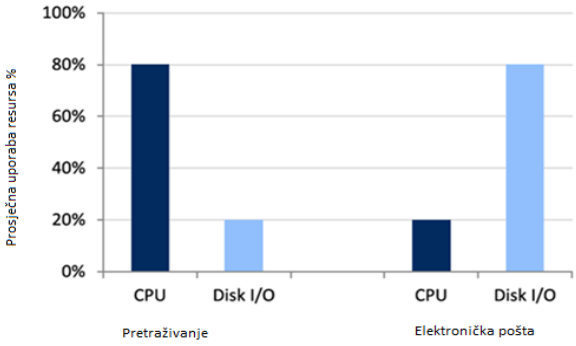
Slika 4. Dnevni uzorci volumena pretraživanja, izvor: Bing Search volumeni u 24 sata

3. **Varijabilnost ovisno o industriji**. Ponekad je varijabilnost opterećenja ovisna o dinamici pojedine industrije koja stvara opterećenje. Firme koje prodaju robu široke potrošnje najveće opterećenje imaju tijekom praznika, dok institucije koje obrađuju porezne prijave najveće opterećenje imaju u drugom ili trećem mjesecu u godini.



Slika 5. Varijabilnost po industriji (izvor: Alexa Internet)

4. **Varijabilnost upotrebe resursa**. *Compute, storage, input/output* (I/O) resursi uobičajeno se kupuju zajedno. Poslužitelj sadrži određenu količinu računalne snage (CPU), prostor za pohranu, te I/O mogućnosti (pristup prostoru za pohranu). Pojedina radna opterećenja, kao što su pretraživanje koriste veću računarsku snagu, ali relativno malo prostora na disku ili kapaciteta I/O, dok druga radna opterećenja kao što je *email* koriste dosta prostora za pohranu, ali malo računalne snage, slika 6.



Slika 6. Varijabilnost različite uporabe resursa, izvor: Microsoft

5. **Nepredvidivi uzorci rasta**. Problemi predviđanja buduće uporabe računarskih resursa i dugotrajno vrijeme da se novi resursi stave u pogon još jedan od razloga neiskorištenja resursa, slika 6. Organizacije moraju osigurati odobrenje za IT resurse daleko prije no što se potreba dogodi ili prije no što razumiju da moraju resurse osigurati. Čak i velike privatne kompanije imaju ove izazove, jer nije uobičajeno da tvrtke planiraju svoje nabave 6 ili 12 mjeseci unaprijed. Upravljajući radnim opterećenjima raznih korisnika, pružatelji usluga računarstva u oblacima mogu smanjiti ovu varijabilnost, jer se zahtjevi za resursima mogu lako međusobno kompenzirati.

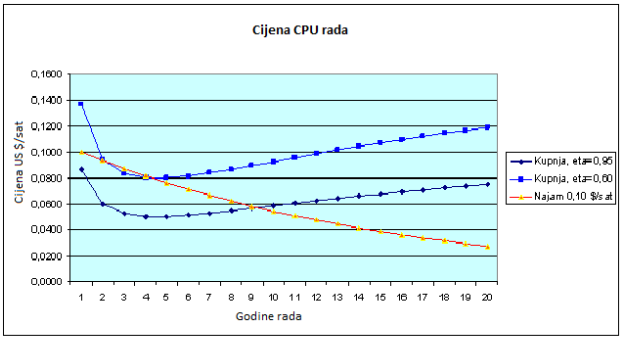
IV. EKONOMSKI EFEKTI RAZLIČITE ISKORIŠTENOSTI IT RESURSA

U nastavku će se dati prikaz komparativnih ekonomskih istraživanja obavljenih u eminentnim svjetskim istraživačkim institucijama iz ove oblasti, u dijelu koji preciznije daje odgovor na pitanje da li investirati uvlastitu IT infrastrukturu ili koristiti koncept oblačnog računarstva u bilo kojoj predviđenoj formi. Kako se može i pretpostaviti rezultati nisu jednoznačni i svakako su u znatnoj mjeri određeni sljedećim parametrima: samom prirodom i veličinom razvijenog biznisa i razvijenosti same kompanije, intenzitetom potreba za sigurnosnim karakteristikama određenih informacija vezanih za poslovanje kompanije i stupnjem potencijalne iskorištenosti IT resursa. Posljednji aspekt je centralno pitanje na koje se traži odgovor, jer on će reći da li iznajmljivati IT resurse i plaćati ih u mjeri koliko se koriste ili graditi vlastitu IT infrastrukturu, slika 7. Sa slike se mogu izvućisljedeći zaključci:

- Tamnoplava crta označava cijenu CPU sata kroz godine uz $\eta = 0,95$ što je iznimno dobro iskorištenje.

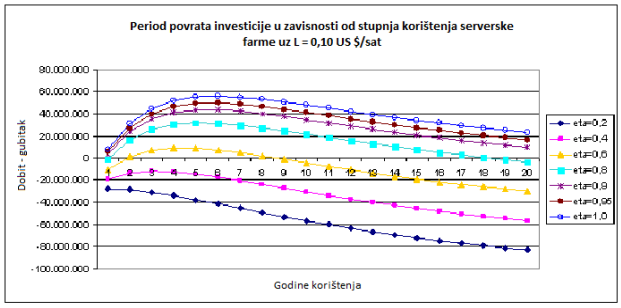
• Svijetloplava crta označava cijenu CPU sata kroz godine uz η = 0,6 što je, IMHO, i dalje jako dobro i rijetko gdje dosegnuto iskorištenje.

• Crvena crta sa žućkastim oznakama označava cijenu najma, formalno 10 centi po satu (kako ju otprilike naplaćuje većina oblačnih poslužitelja poput Amazona i Microsofta), no diskontiranu na netto sadašnju vrijednost.



Slika 7. Racionalnost kupovine i najma IT resursa

Povrat investicije (*Return On Investment* –ROI) izrazito ovisi o iskorištenju serverske firme η kako je prikazano na slici 8.



Slika 8. ROI serverske farme u ovisnosti o η

Dakle, uz visoki η, ulagač u računalni oblak u prvih 6 godina će imati povrat sredstava prosječno 12,6% godišnje u rangu ulaganja u vrlo uspješne dioničke fondove. Znatno bolje od običnog oročavanja u banci!

Može se zaključiti da sadašnja cijena najma od 10 centi po CPU satu omogućava dobru zaradu poslužiteljima koji znaju iskoristiti svoje servere, dok sam najam predstavlja isplativo rješenje za one korisnike koji ne mogu postići visok stupanj prosječnog iskorištenja vlastitih računalnih resursa. Stoga se može predvidjeti da će u budućnosti uspješno poslovati samo oni oblačni poslužitelji koji će na vrijeme usklađivati svoja ulaganja u serversku opremu s povećanjem potražnje CPU vremena i pri tome, na tehničkoj razini, pružati svoje usluge uz maksimalni stupanj iskorištenja servera.

V. ZAKLJUČAK

Vlastite računarske kapacitete isplativo je kupovati onda ako su aplikacije tako dobro pripremljene da će nabavljeni računarski resursi biti iskorišteni u vrlo visokom stupnju. Ako pak organizacija nije spremna za intenzivno korištenje suvremenih informacijskih tehnologija u svim svojim poslovnim procesima, onda je za nju povoljnije da za izabrani poslovni proces unajmi u računarskom oblaku točno toliko računarskih kapaciteta, koliko treba za pripremljene i uvedene aplikacije. Ova opcija je osobito povoljna za mala i srednja poduzeća koja uglavnom nemaju dovoljno stručnjaka informatike za organizaciju i pogon vlastitog računarskog centra, ali žele koristiti suvremene informacijske tehnologije za potporu nekog od svojih strateških poslovnih procesa.

Cloud computing omogućava pristup računarima i njihovim funkcionalnostima putem Internet mreže ili lokalne mreže. Korisnici oblaka imaju pristup putem niza web usluga koje upravljaju računarskim resursima kao što su na primjer računari, mreža, sustav za pohranu podataka, operativni sustav, okružje za razvoj aplikacija, aplikacijski programi, te mnogi drugi. Računarski resursi su dodijeljeni korisniku sve dok ih korisnik ne oslobodi. Uz *cloud*, troškovi interne infrastrukture smanjuju se i do 80 %, a pada i niz operativnih troškova poput troškova održavanja, troškova administracije. Padaju i troškovi skaliranja informatičke infrastrukture. Jednostavno se može iznajmiti više prostora u *cloud*-u umjesto kupovine posve novih, skupih servera.

LITERATURA

[1] The Economics of Virtualization: Moving Toward an Application Based Cost Model, IDC, November 2009.
[2] . http://www.internetworldstats.com
[3] . http://www.gartner.com
[4] Kalakota R., Robinson M (2002), E-Poslovanje2.0, Vodič ka uspjehu, Zagreb: Mate
[5] Tomislav Brozin, Dobriša Adamec : „Uzlet u oblake“, Infotrend, Zagreb, srpanj 2011.

Analiza problema repova kod telekomunikacijskih operatera

Zaim Buljubašić, BH Telecom, Aleja Alije Izetbegovića 29, 75000 Tuzla, Bosnia & Herzegovina,
e-mail: zaim.buljubasic@bhtelecom.ba
Nusret Salkić, BH Telecom, Aleja Alije Izetbegovića 29, 75000 Tuzla, Bosnia & Herzegovina,
e-mail: nusret.salkic@bhtelecom.ba
Nermin Buljubašić, BH Telecom, Aleja Alije Izetbegovića 29, 75000 Tuzla, Bosnia & Herzegovina,
e-mail: nermin.buljubasic@bhtelecom.ba

Sažetak: Kod velikog broja poslovnih procesa koji u temelju imaju uslužni karakter javljaju se problemi oko linija čekanja (repova), odnosno optimizacije tih procesa, vodeći računa o kvaliteti usluge, zadovoljstvu korisnika i naravno ukupnim troškovima tih procesa. Osnovni elementi repova su: dolasci, poslužitelji i vrijeme posluživanja. Sam proces je iniciran dolascima, ograničen kapacitetima poslužitelja i troškovima koje on proizvodi u vremenu opsluživanja koje je također od velikog značaja za ukupno vrijeme trajanja „repa". Problem repova se može riješiti uz pomoć analitičkih formula i metodama simulacije. Primjena analitičkih formula daje početni (aproksimativni) model repa iz kojeg se mogu vršiti daljnja proračunavanja i željene analize. Ovaj model je dosta jeftiniji, ali su i rezultati manje pouzdani zbog potrebnog uopštavanja i prilagođavanja matematskog modela. Temelj za utvrđivanje analitičkog modela čini Poissonova distribucija vjerojatnoća, jer je najprimjerenija ditribuciji dolazaka korisnika usluga. Definiranje modela repova, bez obzira kojom metodom se radilo, daje menadžeru mogućnost da prilikom oblikovnja ovih poslovnih procesa, ustvari ih može planirati i upravljati njima, što je naročito značajno za zadovoljenje i pridobijanje novih klijenata, zadržavanje postojećih i povećanja ukupnog korisničkog zadovoljstva.

Ključne riječi: linije čekanja, zadovoljstvo korisnika, Poissonova distribucija vjerojatnoća, poslužitelji, dolasci

I. UVOD

Linije čekanja su kontinuirani problem većeg dijela poslovnih procesa, a prije svega uslužnog karaktera, jer su uvjetovani činom podnošenja zahtjeva za određenom uslugom, a zatim i ograničenim potencijalima za realizaciju istih, prije svega od strane poslužitelja. Treba odmah reći da pomenuta ograničenost u najvećem broju slučajeva je određena troškovnim, odnosno ekonomskim motivima. Nije namjera autora između pomenuta dva „ograničenja" stavljati znak jednakosti, ali je praksa pokazala da su u velikom broju slučajeva oni međusobno uvjetovani. Za potpunu analizu ovog problema potreban je višekriterijalni pristup definiranja integralnih ciljnih funkcija što zahtijeva primjenu moćnih i prije svega skupih alata u njegovom definiranju i implementaciji. U ovom radu će biti predstavljen jedan od mogućih matematskih modela uobličavanja i rješavanja ovog problema uz skromna ulaganja, a koji s druge strane može poslužiti kao dobro polazište kompleksnijim i zahtijevnijim simulacionim ili nekim drugim metodama. Sva istraživanja i testiranja mogućnosti primjene ovih rješenja su vršena u praksi stvarnim veličinama i rezultati su povrdili svoju primjenjivost, što im daje dodatnu validnost uvažavajući navedena ograničenja. Zašto je ovo pitanje uopće bitno? Činjenice nas uvjeravaju da ekonomija svih zemalja svijeta, a prije svega razvijenih zemalja postaje pretežno uslužnog karaktera. U to nas nedvojbenu uvjeravaju podaci o učešću uslužnog sektora u ukupnom BDP svake zemlje, te kretanja o broju uposlenih po pojedinim sektorima. Dalje, svako od nas ima sliku ili percepciju odlične usluge i one koja to nije. Značaj analize repova kod telekomunikacijskih operatera ogleda se prije svega u optimiziranju angažovanih resursa, gledano prije svega dovodeći u sklad nivo angažovanih sredstava na ekonomski prihvatljivom nivou i zadovoljavajućeg stepena korisničkog zadovoljstva koji će biti projektovan tako da nivo zadržavanja korisnika u sistemu, odnosno repu bude takav da korisnici po ovom osnovu ne prelaze kod drugih poslužitelja.

II. KARAKTERISTIKE REPOVA

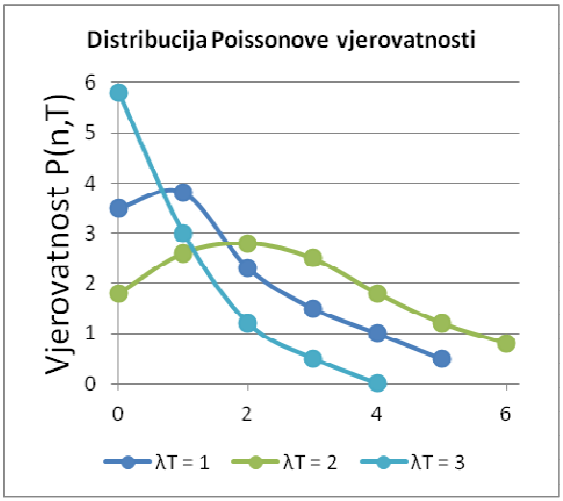
Svi repovi imaju neke svoje zajedničke karakteristike. Oni se formiraju na sličan način i i njihovo iniciranje kroz podnošenje zahtjeva ili dolaske se ustvari odvija po određenoj distribuciji vjerojatnoća. Može se također zaključiti da vrijede manje-više ista pravila bez obzira da li čekanje bilo u redovima trgovinskih objekata na kasama roba široke potrošnje, čekanja u redovima banaka u cilju korištenja bankarskih usluga, distribucija polijetanja ili slijetanja pojedinih aviona, popravke i stavljanja u ponovnu funkciju pojedinih strojeva, traženja i korištenja telekomunikacijskih usluga od telekomunikacijskih operatera itd. Svaki problem

repa se može opisati sa slijedeće tri karakteristike: dolazak, rep i poslužitelj.

1. Dolazak. Dolasci su opisani njihovom statističkom distribucijom dolazaka, koja se može specificirati na dva načina: dolascima u jedinici vremena ili distribucijom vremena između dolazaka. Ako se dolasci specificiraju po prvom principu, onda je neophodno definirati broj dolazaka u jedinici koja može biti npr. 1h (jedan sat), jedan dan ili bilo koje drugo vremensko razdoblje. Kako se dolasci događaju slučajno, značajno je odrediti vjerojatnost n dolazaka u određenom vremenskom razdoblju pri čemu n može biti n = 0,1, 2,3,.....n. Ako se dolasci promatraju po drugom načinu, tada se dolasci događaju u prosječno konstantnoj veličini u određenom vremenu, s tim što su oni međusobno neovisni i distribucija vremena između dolazaka je primjerena Poissonovoj distribuciji vjerojatnoća. U tom slučaju je vjerojatnost n dolazaka u vremenu T određena slijedećim izrazom:

$$P(n,T) = \frac{e^{-\lambda T}(\lambda T)^n}{n!} \qquad n = 0,1,2,\ldots$$

pri čemu je: λ – prosječan broj primljenih zahtjeva (dolazaka); T = vremensko razdoblje; P(n,T) = vjerojatnost dolazaka u vremenu T. U nastavku se mogu vidjeti n grafičkom primjeru tri tipična modela Poissonove distribucije vjerojatnosti, te svu logičnost u primjeni baš ovog modela distribucije vjerojatnoća kod definiranja analitičkog modela linija čekanja, odnosno repova:

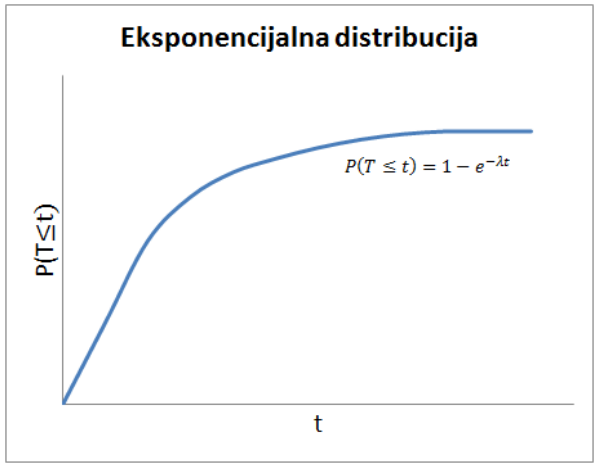


Slika 1: Distribucija Poissonove vjerjatnosti [1]

Iz ovog grafikona se jasno vidi da je za vrijednost λ T = 1 najveća vjerojatnost dolaska prishvatljiva ka i za vrijednosti λ T=2 i 3, s tim što je sasvim vidljivo i jasno da će za veći broj dolazaka od navedenih u istom vremenskom razdoblju biti mala vjerojatnost. Ako se specificiraju vremena između dolazaka i pokuša se utvrditi pravilnost u distribuciji međudolazaka, jasno se može reći da čak i sa većim stupnjem pouzdanosti opravdanost primjene modela Poissonove distribucije vjerojatnosti, te da ona u vremenu poprima eksponencijalni karakter:

$$P(T \leq t) = 1 - e^{-\lambda t} \qquad 0 \leq t < \infty$$

pri čemu je: $P\ (T \leq t)$= vjerojatnost da je vrijeme između dolazaka T ≤ od zadane veličine t; λ= srednje vrijeme dolazaka u jedinici vremena; t= zadana veličina vremena. Grafički to izgleda ovako:



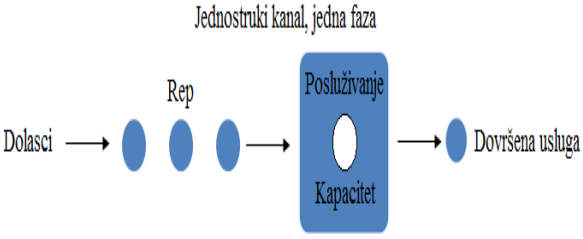
Slika 2:Eksponencijalna distribucija [2]

Eksponencijalna i Poissonova distribucija potpuno su ekvivalentne u pretpostavkama o dolascima, pa se i jedna i druga mogu iskoristiti za specificiranje dolazaka ovisno o tome želi li se dobiti distribucija vremena između dolazaka ili broj dolazaka u određenom vremenskom razdoblju. Postoje i druge distribucije koje se mogu koristiti za specificiranje dolazaka, kao što je npr. Erlangova distribucija i druge.

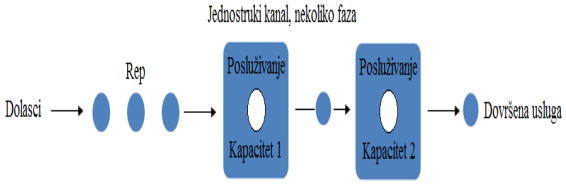
2. Rep. Za definiranje modela repa moraju se odrediti i određene njegove karakteristike, koje se najčešće nazivaju disiplina repa. Takve pretpostavke se moraju unaprijed postaviti, pa se recimo zna da li se opluživanje dolazaka vrši po pravilu „prvi došao-prvi opslužen“ ili se pojedini zahtjevi prioriteziraju pa se njihovo vrijeme opsluživanja skraćuje i pomjera na početak repa. Dužina repa također treba biti određena, jer u nekim slučjevima čak i kada su repovi neograničeni to nema nikakvu praktičnu važnost, dok opet u drugim slučajevima dužina zadržavanja u repu može biti od presudnog značaja za poslužitelja. U definiranju samog analitičkog modela ovi elementi moraju biti unaprijed određeni.

3. Poslužitelj. Poslužitelji kao dio ovog poslovnog procesa, također imaju određene karakteristike koje je potrebno znati i one se ogledaju u više elemenata. Jedna od tih karakteristika je svakako distribucija vremena opsluživanja za koje je logično

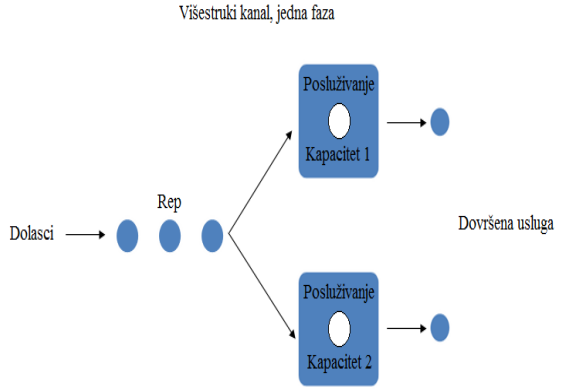
pretpostaviti da se odvija po principima eksponencijalne distribucije, jer je realna pretpostavka da će u dužem vremenskom razdoblju eksponencijalno rasti vjerojatnost povećanog broja opsluživanja. Međutim treba isto tako znati da i distribucija vremena opsluživanja, kao i distribucija dolazaka može varirati od kupca do kupca, a i od poslužitelja do poslužitelja. Druga karakteristika je broj poslužitelja koje često nazivaju kanalima opsluživanja, jer i njih može biti različit broj, a svaki dodatni kanal opsluživanja u analitičkom smislu model čini kompleksnijim. Slijedeća karakteristika je da se usluga može pružati kroz jednu fazu i jedan kana, dok imamo slučajeva kada to nije moguće. Neke od mogućih kombinacija različitog broja kanala date su na narednim slikama:



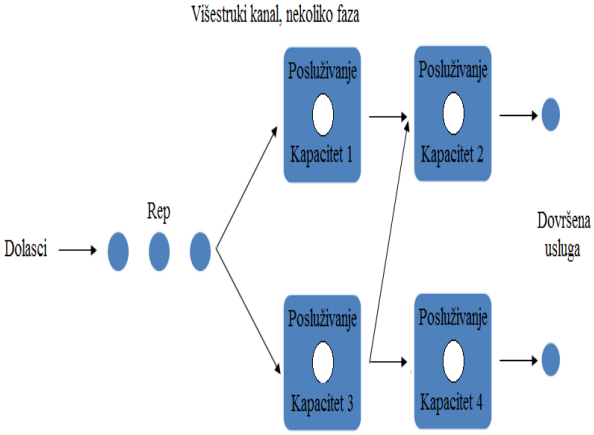
Ovaj tip zahtijevane usluge je najjednostavniji kao primjer može poslužiti čekanje u redu za plaćanje u trgovinskom objektu, plaćanje računa komunalnih usluga na specijaliziranim punktovima i slično.



Ovaj model predstavlja prikaz jednostrukog kanala dolaska i dovršetka usluge; u procesu pružanja se traži višefazni pristup. Primjer: zahtjev za uslugama fiksne telefonije kod telekomunijacijskih operatera i sl.



Ovaj model predpostavlja složenije poslovne procese koji se pružaju uz više specijaliziranih poslovnih faza. Primjer: pojedine usluge iz oblasti ICT, zdravstva, pravosuđa i slično.



Ovaj model predpostavlja iznimno komplekne, visokostručne, često puta i prostorno odvojene procese.

III. FORMULIRANJE PROBLEMA REPOVA

Uvažavajući neke od karakteristika repova, u nastavku će im se pridodati određene pretpostavljene veličine, kako bi se mogli izvući zaključci o različitim ponašanjima sustava repova u različitim uvjetima. Prije svega te pretpostavljene performanse se ogledaju u različitom prosječnom broju dolazaka, različitom vremenu čekanja i različitim stupnjem nerada poslužitelja.Te različite pretpostavljene performanse mogu dati odgovor na potreban broj angažiranih kanala posluživanja, cijene usluge, marketinške strategije i druga značajna pitanja iz oblasti tržišnog pozicioniranaja. Mjerila performansi su data u slijedećem lako mjerljivom sustavu simbola i njihovim pretpostavljenim iznosima:

<u>Pretpostavka 1</u>

λ – prosječan broj primljenih zahtjeva (dolazaka)

$\lambda = 60{,}18$

μ – prosječan broj realizovanih zahtjeva (usluženja)

$\mu = 46{,}65$

$\lambda > \mu$ => vodi u sistem nestabilnih repova koji teži beskonačnosti

$$\rho = \frac{\lambda}{\mu} = \frac{60{,}18}{46{,}65} = 1{,}29$$

Da bi se odradili svi poslovi, poslužitelji bi trebali raditi sa najmanje 29 % više kapaciteta (produktivnosti) da bi bio jednak broj primljenih zahtjeva, čime se ne bi eliminisali repovi.

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - 1{,}29 = -0{,}29$$

U prvom trenutku najmanje 29 % korisnika će biti u repu, što svakako nije poželjno, a u n-tom trenutku relativni broj korisnika u repu se izračunava po slijedećoj formuli:

$$P_n = P_0\left(\frac{\lambda}{\mu}\right)^n$$

$$P_1 = P_0\left(\frac{\lambda}{\mu}\right) = -0{,}29\left(\frac{60{,}18}{46{,}65}\right) = -0{,}3741$$

Poslije prvog perioda mjerenja u repu će biti najmanje 37,41 % korisnika

$$P_2 = P_0\left(\frac{\lambda}{\mu}\right)^2 = -0{,}29\left(\frac{60{,}18}{46{,}65}\right)^2 = -0{,}4825$$

- što slijedi navedenu logiku.

$$P_3 = P_0\left(\frac{\lambda}{\mu}\right)^3 = -0{,}29(1{,}29)^3 = -0{,}6225$$

- logično je zaključiti da u ovako pretpostavljenim uvjetima se ide u sistem nestabilnih repova.

Pretpostavka 2
λ = 60,18
μ = 70,00

$$\rho = \frac{\lambda}{\mu} = \frac{60{,}18}{70{,}00} = 0{,}8597 <=> 85{,}97\%$$

$$P_0 = 1 - \rho = 1 - 0{,}8597 = 0{,}1403$$

- vremena neće biti korisnika u repu

$$P_1 = P_0\rho = 0{,}1403 \times 0{,}8597 = 0{,}1206$$
$$P_2 = P_0\rho^2 = 0{,}1403(0{,}8597)^2 = 0{,}1037$$
$$P_3 = P_0\rho^3 = 0{,}1403(0{,}8597)^3 = 0{,}0891$$

$$\text{Broj izbug. korisnika} = 1 - (P_0 + P_1 + P_2 + P_3) = 1 - 0{,}4537 = 0{,}5463 <=> 54{,}63\%$$

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{3621{,}63}{70(70 - 60{,}18)} = 5{,}268$$

- prosječno korisnika u repu

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{60{,}18}{9{,}82} = 6{,}128$$

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{60{,}18}{70(70 - 60{,}18)} = 0{,}0875$$

- prosječno dana u repu

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{9{,}82} = 0{,}1018$$

- prosječno dana u sistemu

Pretpostavka 3
λ = 60,18
μ = 80,00

$$\rho = \frac{\lambda}{\mu} = \frac{60{,}18}{80{,}00} = 0{,}7522 <=> 75{,}22\%$$

$$P_0 = 1 - \rho = 1 - 0{,}7522 = 0{,}2478$$
$$P_1 = P_0\rho = 0{,}2478 \times 0{,}7522 = 0{,}1863$$
$$P_2 = P_0\rho^2 = 0{,}2478(0{,}7522)^2 = 0{,}1402$$
$$P_3 = P_0\rho^3 = 0{,}2478(0{,}7522)^3 = 0{,}1054$$

$$\text{Broj izgub. korisnika} = 1 - (P_0 + P_1 + P_2 + P_3) = 1 - 0{,}6797 = 0{,}3203 <=> 32{,}03\%$$

$$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{3621{,}63}{80(80-60{,}18)} = 2{,}284$$

$$L_s = \frac{\lambda}{\mu-\lambda} = \frac{60{,}18}{19{,}82} = 3{,}036$$

$$W_q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{60{,}18}{1585{,}6} = 0{,}0379$$

$$W_s = \frac{1}{\mu-\lambda} = \frac{1}{19{,}82} = 0{,}0504$$

Pretpostavka 4
λ = 60,18
μ = 93,30 (46,65 x 2)

$$\rho = \frac{\lambda}{\mu} = \frac{60{,}18}{93{,}30} = 0{,}6450 <=> 64{,}50\%$$

$$P_0 = 1 - \rho = 1 - 0{,}645 = 0{,}355$$
$$P_1 = P_0\rho = 0{,}355 \times 0{,}645 = 0{,}2289$$
$$P_2 = P_0\rho^2 = 0{,}355 \times 0{,}416 = 0{,}1476$$
$$P_3 = P_0\rho^3 = 0{,}355 \times 0{,}2683 = 0{,}0952$$

$$\text{Broj izgub. korisnika} = 1 - (P_0 + P_1 + P_2 + P_3) = 1 - 0{,}8267 = 0{,}1733 <=> 17{,}33\%$$

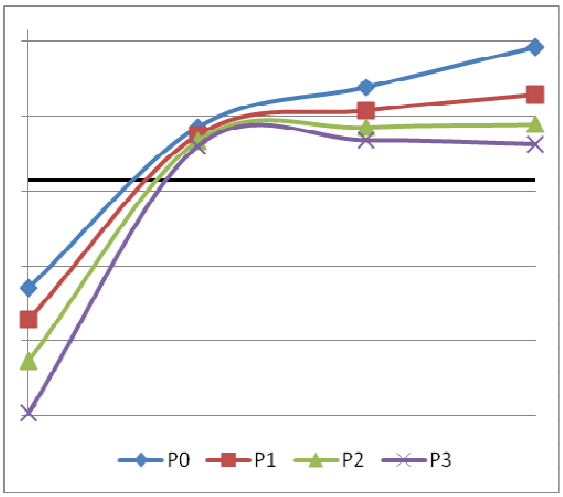
$$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{3621{,}63}{93{,}3(93{,}3-60{,}18)} = 1{,}172$$

$$L_s = \frac{\lambda}{\mu-\lambda} = \frac{60{,}18}{93{,}30-60{,}18} = 1{,}817$$

$$W_q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{60{,}18}{3090{,}09} = 0{,}0194$$

$$W_s = \frac{1}{\mu-\lambda} = \frac{1}{33{,}12} = 0{,}0301$$

U nastavku će biti dat grafički prikaz kretanja, odnosno stepen ovisnosti promjena iskorištenosti radnika i vjerojatnoće pojavljivanja broja korisnika u n vremena.



Slika 7. Vjerojatnoća broja korisnika u n-tom vremenu

IV. ZAKLJUČAK

Iz naprijed rečenog mogu se izvući slijedeći zaključci:

-za postavljanje analitičkog modela u uvjetima dolazaka koji su slučajnog karaktera i poštovanje osnovnih disciplina ponašanja u repu što se tiče redoslijeda i dinamike opsluživanja, kao jako primjenjiva se čini Poissonova distribucija vjerojatnsti

-za uspješno upravljanje poslovnim procesima prije svega u uslužnim djelatnostima neophodno je utvrditi model koji će omogućiti menadžerima tih procesa donošenje optimalnih odluka sa aspekta kvalitete usluge i ekonomskih motiva,

-u uslužnim djelatnostima je postojanje repova objektivna nužnost, obzirom da je sam proces iniciran dolascima, pa uz izvjesno vrijeme i isporuku usluge. Tu se ne želi poništiti karakteristika marketinga usluga o ostovremenosti korištenja i davanja usluge. Naprotiv, ovdje se dodaje specifičnost karaktera usluge i produbljuje poimanje marketinga usluga u specifičnim uvjetima,

-racionaliziranje same usluge uz zadržavanje projicirane kvalitete je moguće realizirati i uz pomoć drugih modela osim ovog koji je predstavljen, ali su oni u principu znatno skuplji i zahtijevaju više vremena.

V. LITERATURA

[1] Wagner, Harvey. Principles of Operation Research, 2.izd.San Francisko,2007.g

[2] Roger Schroeder, Upravljanje proizvodnjom, Mate, Zagreb, 1999.g

[3] Đurić, Jan, Marketing usluga, Nova misao, Beograd, 2009.g

[4] Buljubašić,Z,Bajrić,H;Spesifičnosti marketinga usluga, Mipro,Zbornik radova,2007.g

Od sustava e-poslovanja klastera do e-klastera

Milan Davidović
VSITE – Visoka škola za informacijske tehnologije, Zagreb, Hrvatska
milan.davidovic@vsite.hr

Sažetak - U svijetu digitalne ekonomije male tvrtke moraju se umrežiti i povezivati se horizontalno i/ili vertikalno u (poslovne) klastere, integrirati u opskrbne lance svojih dobavljača, trgovaca ili kupaca, te uvoditi nove direktne i indirektne distribucijske e-kanale i stvarati sustav e-poslovanja koji omogućava sigurnu razmjenu dobara, usluga i informacija te razvoj kooperativnih odnosa koji su uvjet održivosti klastera. Da bi bio djelotvoran i učinkovit, sustav e-poslovanja klastera treba biti interoperabilan na svim razinama kako bi se horizontalno i vertikalno (informacijski) integriralo poslovanje klastera radi zadovoljavanja potreba i zahtjeva kupaca te obavljale elektroničke transakcije između svih učesnika e-poslovanja. Ovakav složeni sustav, koji je po svom karakteru otvoren jer treba omogućiti i integraciju opskrbnih i distributivnih e-lanaca, praktično je vrlo teško realizirati primjenom postupaka uvođenja e-poslovanja koje se primjenjuju u većim poslovnim sustavima. Kao izvedivo rješenje predlaže se e-klaster kao oblik poslovne mreže koja omogućava malim i srednjim tvrtkama, u uvjetima digitalne ekonomije, da se fokusiraju na svoje ključne sposobnosti i kroz lanac stvaranja nove vrijednosti u klasteru ostvaruju održivi rast i istovremeno doprinose razvoju klastera i njegovih članova.

I. UVOD

Mali proizvođači, mali trgovci i druge male tvrtke horizontalno se povezuju na određenom geografskom području i formiraju klastere kako bi se uključili u velike opskrbne lance i ostvarili bolju pregovaračku poziciju. Vertikalnim povezivanjem malih poduzeća i njihovih partnera u klastere uspostavljaju se novi lanci stvaranja vrijednosti i vrijednosni lanci koji mogu konkurirati drugim opskrbnim lancima ili postati njihov partner u tržišnom nadmetanju između opskrbnih lanaca. Uključivanjem u klastere znanja (eng. *knowledge cluster*) male tvrtke razvijaju i implementiraju nove tehnologije i stvaraju inovativne proizvode i usluge kojima mogu osvojiti nova tržišta i ostvariti brži rast. Velike tvrtke osnivaju klastere u koje uključuju male tvrtke zbog znanja koje posjeduju i smanjenja troškova, te tako stvaraju novu dodanu vrijednost važnu u tržišnom nadmetanju.

Klasteri (eng. *business cluster*, *cluster*) su regionalne grupacije poduzeća i ustanova na određenom području koje omogućavaju da se konkurentska prednost na globalnom tržištu ostvaruje korištenjem lokalnih snaga – znanja, povezanosti i motivacije, koje udaljeni konkurenti nemaju. Povećanje konkurentnosti klastera zasniva se na rastu produktivnosti njegovih članova, inovacijama koje podupiru budući rast produktivnosti, širenju poslovanja i stalnom rastu klastera [1]. Sustav e-poslovanja i korištenje e-dobavnih lanaca te indirektnih i direktnih e-distribucijskih kanala postali su preduvjet opstanka klastera i njegovih članova u uvjetima koji vladaju u digitalnoj ekonomiji, jer djelovanje klastera izvan sustava e-trgovine više nema perspektivu [2][3][4]. No, kako je pokazano u ovom radu, na razvoj sustava e-poslovanja klastera utječu model organizacije klastera i kooperacijski odnosi koji vladaju u klasteru, a uspostavljeni sustav e-poslovanja neće omogućiti iskorištavanje svih prednosti klastera dok nije uspostavljen sustav upravljanja znanjem klastera koji obuhvaća inoviranje proizvoda i procesa kooperacije. A to nas dovodi do e-klastera kao novog koncepta razvoja klastera.

II. SUSTAV E-POSLOVANJA POSLOVNIH KLASTERA

Elektroničko poslovanje ili **e-poslovanje klastera** (eng. *cluster e-business*) je

- **Skup aktivnosti** koje poduzimaju menadžment, članovi i poslovni partneri klastera uz informacijsko-komunikacijsku (IKT) potporu radi sigurne razmjene dobara, usluga i informacija,
- **Skup kooperacijskih odnosa** koje grade menadžment, članovi i poslovni partneri klastera kroz preustroj i unaprjeđenje poslovnih procesa i funkcija klastera radi djelotvornog i učinkovitog korištenja IKT u provedbi kooperacijskih projekata i nastupu klastera na tržištu te
- Razmjena vjerodostojnih informacija kroz **elektroničke transakcije koje su normirane (po sadržaju i značenju)** koje prate razmjenu dobara i usluga među članovima klastera, te između klastera i ostalih učesnika u realizaciji kooperacijskih projekata i zajedničkih proizvoda klastera, te u nastupu klastera na tržištu [3].

Iz definicije e-poslovanja klastera proizlazi da e-poslovanje klastera može obuhvatiti samo ograničeni skup aktivnosti i elektroničkih transakcija u razmjeni dobara, usluga i informacija uz potporu IKT koje se obavljaju u skladu s utvrđenim kooperacijskim odnosima u klasteru. Dakle, informatizirati se mogu samo aktivnosti kojima se realizira neki **poslovni slučaj e-poslovanja klastera** (eng. *e-business case*) [2][4].

Aplikacijski sustavi (e-rješenja) koji podržavaju funkcije klastera navedeni su u Tablici I. Klasteri čiji članovi su orijentirani na vlastito poslovanje mogu kao e-rješenje uvesti samo e-nabavu ili e-prodaju. Klasteri orijentirani na kupce mogu koristiti jednostavna rješenja e-prodaje ili sustav upravljanja odnosima s kupcima (CRM), a klasteri orijentirani na kooperacijske projekte samo aplikaciju za upravljanje projektima (PM) koja podržava elektroničke transakcije između članova klastera

TABLICA I. E-RJEŠENJA

R. br.	Aktivnosti/Funkcija klastera	e-rješenje							
		e-nabava	SCM	ERP	PM	KM	BI	e-prodaja	CRM
1	Nabava	+	+						
2	Logistika		+						
3	Ljudski potencijali			+	+				
4	Kooperacijski projekti			+	+				
5	Outsourcing			+	+				
6	Proizvodnja			+	+				
7	Istraživanje i razvoj				+	+	+		
8	Prodaja		+					+	+
9	Istraživanje tržišta						+		+
10	Marketing								+
11	Služba za korisnike								+
12	Upravljanje znanjem					+			

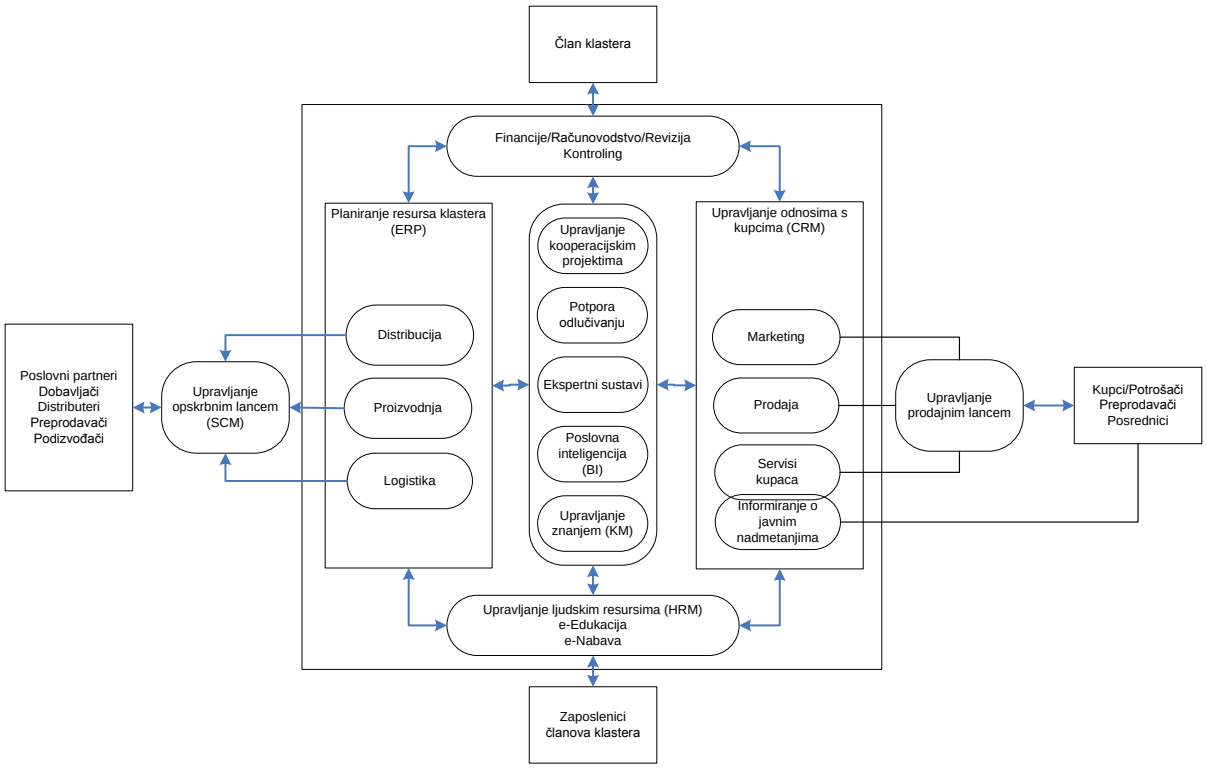
SCM-Supply Chain Management (upravljanje opskrbnim lanvem); ERP – Enterprise Resource Planning (planiranje resursa); PM – Project Management (upravljanje projektima); KM – Knowledge Management (upravljanje znanjem); BI – Business Intelligence (sustav poslovne inteligencije); CRM – Customer Relationship Management (upravljanje odnosima s kupcima):

i između klastera i njegovih partnera. Vertikalni klasteri mogu uvesti sustav za upravljanje opskrbnim lancem (SCM) koji se može povezati sa sustavom za planiranje resursa (ERP) ili ERP koji ima i SCM modul [2][3].

Potpuni **aplikacijski okvir e-poslovanja klastera** (eng. *e-business application framework*) obuhvaća sve funkcije klastera i podržava sve dionike (kao kod složenog poduzeća, v. sliku 1). **Dionici klastera** (eng. *stakeholders*) su poslovni partneri (kupci i dobavljači), zaposlenici i članovi klastera (za razliku od sustava e-poslovanja poduzeća gdje to mjesto pripada dioničarima).

Sustav e-poslovanja klastera (eng. *cluster e-business system*) treba integrirati aktivnosti, zajedničke funkcije i usluge klastera i informacijske sustave članova klastera te omogućiti sigurnu razmjenu informacija, dobara i usluga i djelotvornu i učinkovitu kooperaciju članova klastera. To mora biti fleksibilan i otvoren sustav koji podržava glavne kompetencije članova klastera i premještanje proizvodnje i usluga kompetentnijim dobavljačima (izmještanje kompetencija ili *outsourcing*). Fleksibilan, jer ima infrastrukturu koja se brzo prilagođava redizajniranim procesima, proizvodima i uslugama, te omogućava horizontalnu i vertikalnu integraciju procesa, članova klastera i konvencionalnih i e-kanala prema potrebama kupaca, a otvoren za uvođenje novih i redizajn postojećih e-rješenja i uključivanje novih članova i partnera radi stvaranja nove dodane vrijednosti i unaprjeđenje lanca vrijednosti klastera.

Interoperabilnost sustava e-poslovanja klastera (eng. *interoperability*) treba postojati na svim razinama tj. elektroničke transakcije obavljaju se u skladu s normama e-poslovanja i omogućavaju interakcije klastera i poslovnih partnera, između članova klastera te klastera i članova (*B2B – business to business*), između klastera i kupaca (*B2C – business to customer/client/ consumer*), između klastera i njegovih članova i državne uprave (npr. plaćanje poreza i carine, *B2G – business to government*) te između klastera i članova klastera i javnih službi (*B2S – business to service*). B2B interoperabilnost preduvjet je za smanjenje transakcijskih troškova klastera i njegovih partnera u realizaciji kooperacijskih projekata i zajedničkom nastupu na tržištu [2].



Slika 1 Aplikacijski okvir klastera [3][4]

SCM-Supply Chain Management (upravljanje opskrbnim lanvem); ERP – Enterprise Resource Planning (planiranje resursa); PM – Project Management (upravljanje projektima); KM – Knowledge Management (upravljanje znanjem); BI – Business Intelligence (sustav poslovne inteligencije); CRM – Customer Relationship Management (upravljanje odnosima s kupcima):

III. UTJECAJ KOOPERACIJSKIH ODNOSA NA RAZVOJ SUSTAVA E-POSLOVANJA KLASTERA

A. Kooperacijski odnosi i razvoj e-poslovanja klastera

Organizacijom klastera određuje se uloga članova klastera i kooperacijski odnosi u realizaciji zajedničkih proizvoda i kooperacijskih projekata klastera, te u nastupu klastera na tržištu.

Prema ulozi i interakciji članova razlikujemo četiri osnovna modela organizacije klastera (v. sliku 2) [5]:

a. Postrojbeni klaster,
b. Klaster vodećih članova,
c. Ekspoziturni klaster i
d. Državno centrirani klaster.

Razvoj i implementaciju sustava e-poslovanja u postrojbenim klasterima je vrlo složen pothvat i nije izvediv bez sinergijskog djelovanja menadžmenta i članova klastera.

Glavnu ulogu u planiranju i provedbi razvoja i implementacije sustava e-poslovanja u ostalim tipovima klastera imaju vodeći članovi, ekspoziture velikih kompanija odnosno javna poduzeća koji imaju kontrolu nad klasterom, Ako pri uvođenju e-poslovanja u klasteru primjenjuju svoje obrasce i model e-poslovanja, oni će smanjiti svoje transakcijske troškove, ali je pitanje da li na taj način mogu iskoristiti ostale prednosti klastera (znanje, inovativni potencijal, sinergija kooperacije i dr) i eliminirati svoje slabosti u odnosu na konkurentske opskrbne lance.

Strategiju razvoja e-poslovanja klastera treba zasnivati na prednostima članova klastera i uspostavi kooperacijskih odnosa s onim tvrtkama i ustanovama koje imaju resurse koji nedostaju klasteru za izgradnju konkurentnijeg opskrbnog lanca i stvaranje nove dodane vrijednosti. Članovi klastera se tako mogu fokusirati na svoje prednosti, a slabosti klastera se eliminiraju kroz kooperaciju, digitalizaciju procesa i elektroničke transakcije koje smanjuju transakcijske troškove.

B. Postrojbeni klasteri

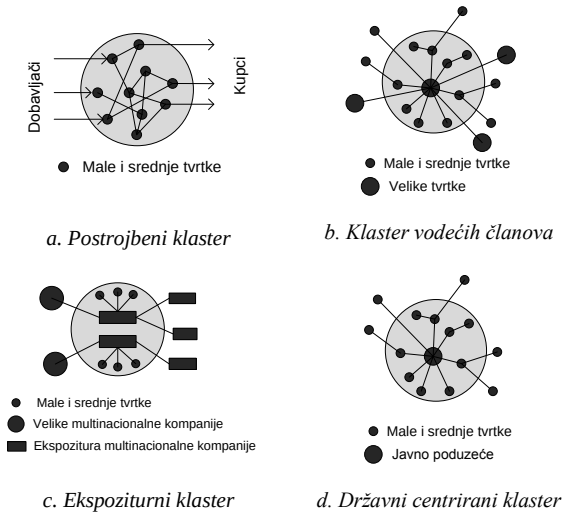
Postrojbeni klaster (eng. *marshalian cluster*) je model organizacije karakterističan za klastere u koje su uključene male tvrtke radi zajedničkog nastupa prema dobavljačima i/ili kupcima. U takvom klasteru članovi klastera kooperiraju s drugim članovima u realizaciji zajedničkih proizvoda i kooperacijskih projekata klastera tj. između članova klastera vladaju odnosi dobavljač-proizvođač ili odnosi izravne konkurencije (v. sliku 2a). Nema člana klastera koji bi zbog svoje veličine i snage mogao preuzeti kontrolu nad klasterom, pa su dinamika razvoja i stanje u klasteru određeni zajedničkim nastupom članova klastera na tržištu [5].

Postrojbeni klasteri mogu biti partneri ili članovi ostalih tipova klastera i ući u njihov opskrbni lanac, koristeći njihove distribucijske kanale ili ustupajući svoje distribucijske kanale, te stvarajući nove (zajedničke) kanale. Distribucijski kanali mogu biti konvencionalni, e-kanali ili njihova kombinacija [2].

Za uvođenje e-poslovanja u postrojbene klastere je neophodan sposoban menadžment klastera, zajednička informatička funkcija i angažman svih članova klastera u pripremi i implementaciji odabranog e-rješenja. Za integraciju opskrbnog lanca i primjenu složenijih modela e-poslovanja kakve koriste velike korporacije, postrojbenim klasterima u pravilu nedostaju potrebni resursi, pogotovu na početku njihovog djelovanja, tako da je jasno zašto većina postrojbenih klastera još uvijek koriste samo Internet za informiranje kupaca te obradu i izvršavanje narudžbi kupaca (informacijski posrednici, posrednici u transakcijama, marketinški e-kanali i samouslužni servisi [3]).

Implementacija sustava e-poslovanja postrojbenih klastera je složen razvojni projekt, jer se istovremeno s digitalizacijom poslovnih procesa moraju mijenjati kooperacijski odnosi ako se želi učinkovito i djelotvorno iskoristiti IKT potpora, pa su upravljanje promjenama (eng. *change management*) i upravljanja znanjem (eng. *knowledge management*) ključne funkcije menadžmenta postrojbenih klastera. Na dinamiku i opseg implementacije e-poslovanja utječu autonomija članova klastera, ograničene ovlasti menadžmenta klastera i velik broj sudionika u procesima učenja i stjecanja novih znanja, te se e-poslovanje postrojbenih klastera uvodi i razvija postupno (kao u velikim korporacijama koje postupno uvode složena e-rješenja koja su rizična zbog složenosti preustroja procesa, promjene odnosa u korporaciji i problema u implementaciji i integraciji [4]).

Kroz implementaciju e-rješenja i razvoj sustava e-poslovanja mijenjaju se kooperacijski odnosi i obratno, bez odgovarajućih kooperacijskih odnosa i sinergije djelovanja članova klastera nije moguće uspješno implementirati nova e-rješenja i nove e-poslovne slučajeve. Štoviše, ako su kooperacijski odnosi neadekvatni, sustav e-poslovanja ili nije moguće uvesti u



a. Postrojbeni klaster *b. Klaster vodećih članova*

c. Ekspoziturni klaster *d. Državno centrirani klaster*

Slika 2. Modeli organizacije klastera [5]

potrebnom opsegu ili će postati nedjelotvoran i neučinkovit, čime se dovode u pitanje izvedivost kooperacijskih projekata i zajednički nastup članova klastera na tržištu ili čak održivost postrojbenih klastera.

C. *Klasteri vodećih članova*

Klaster vodećih članova (eng. *hub-and-spoke cluster*) ima nekoliko velikih članova oko kojih se grupiraju male tvrtke koje izravno trguju s njima i ovise o njihovoj strategiji prema kupcima (v. sliku 2b). Male tvrtke su u pravilu dobavljači (proizvođači dijelova i specijalizirane tvrtke za pojedine faze proizvodnog procesa ili poslije-prodajne i druge usluge), a mogu biti i posrednici ili kupci (na primjer, mali trgovci). Vodeći članovi (eng. *hub firm*) određuju odnose unutar klastera i dinamiku njegovog razvoja. Tipičan primjer za taj model su klasteri koji se okupljaju oko vodećih američkih i evropskih proizvođača automobila [5].

Vodeći članovi razvijaju svoj opskrbni lanac u koji uključuju male tvrtke da bi postali konkurentniji. Oni se međusobno povezuju radi postizanja boljih uvjeta nabave i smanjivanja transakcijskih troškova, a na prodajnom tržištu ostaju i dalje konkurenti.

Male tvrtke ulaze u sustave e-poslovanja vodećih članova izravno ili preko postrojbenih klastera, tako što djeluju unutar e-dobavnih lanaca ili distribucijskih e-kanala vodećih članova. Klasterom upravljaju vodeći članovi u skladu sa zajednički utvrđenom strategijom.

Korištenjem e-rješenja vodećih članova u čiji su opskrbni lanac ušli, male tvrtke neće riješiti svoje e-poslovanje ako im sustav vodećeg člana ne omogućava punu interoperabilnost prema drugim kupcima i dobavljačima, državnoj upravi i javnim službama.

Ako vodeći članovi integriraju svoje dobavne lance u koje ulaze i ostali članovi klastera i dobavljači, onda je upravljanje zajedničkim dobavnim lancem funkcija klastera, a na razini klastera se uspostavlja sustav upravljanja dobavnim lancima koji je povezan sa sustavima e-poslovanja vodećih članova. To može znatno smanjiti troškove razvoja e-poslovanja vodećih članova i njihove transakcijske troškove u odnosu na odvojene sustave e-poslovanja.

D. *Ekspoziturni klasteri*

Ekspoziturni klaster (eng. *satellite platform cluster*) je model regionalnog povezivanja podružnica i proizvodnih pogona velikih (multinacionalnih) poduzeća s malim tvrtkama na određenom geografskom području radi smanjenja troškova nabave i radne snage i korištenja poticajnih mjera vlade (v. sliku 2c). Pri tom nema veza među industrijskim granama, a svi članovi klastera, koji su dijelovi velikih poduzeća sa sjedištem izvan regije, pod izravnom su kontrolom matične tvrtke [5].

Sustavi e-poslovanja ekspoziturnih klastera su usmjereni na potporu aktivnostima klastera, uspostavljanju i održavanju kooperacijskih odnosa i suradnju s državnom i lokalnom administracijom, dok se sve poslovne aktivnosti odvijaju unutar sustava e-poslovanja ekspozitura ili velikih multinacionalnih kompanija. Male tvrtke u potpunosti slijede obrasce e-poslovanja ekspozitura preko kojih ulaze u opskrbne lance njihovih matičnih kompanija.

E. *Državni centrirani klasteri*

Državni centrirani klaster (eng. *state anchored or state centered cluster model*) je model klastera u kojem se okupljaju mala i srednja poduzeća oko javnog poduzeća, državne ili neke neprofitne organizacije koja dominira u nekoj regiji i ekonomskim odnosima među članovima klastera kroz ugovore o javno-privatnom partnerstvu (v. sliku 2d). Taj model se može usporediti s klasterom vodećih članova u kojemu postoji samo jedan vodeći član klastera koji nije u privatnom vlasništvu (primjer takvih klastera su klasteri koje osnivaju američke vojne baze) [5].

Okosnicu sustava e-poslovanja državnog centriranog klastera čini sustav e-poslovanja javnog poduzeća preko kojeg male tvrtke i vanjski dobavljači ulaze u dobavne e-lance javnog poduzeća ili pružaju usluge u njegovim distribucijskim e-lancima.

IV. E-KLASTERI I E-CLUSTERING

A. *Uloga i značaj e-klastera*

Dosadašnji razvoj sustava e-poslovanja klastera bio je usredotočen na uvođenje i primjenu e-trgovine, razvoj sustava upravljanja opskrbnim lancima (SCM) i primjenu drugih e-rješenja te postupno uvođenje potpunog aplikacijskog okvira e-poslovanja (v. sliku 1). Pri tom klasteri koriste infrastrukturu vodećih članova ili uvode vlastiti sustav e-poslovanja u sklopu logistike, informatizacije zajedničkih funkcija i potpore menadžmentu klastera, uglavnom radi veće djelotvornosti i učinkovitosti te smanjenja transakcijskih troškova.

Na području tzv. TIMES tehnologija (eng. akr. *Telecommunication, Information technology, Multimedia, Entertainment, Security*), u koje ulaze telekomunikacije, informacijska tehnologija, multimedija, industrija zabave i računalna sigurnost, osnivaju se times-klasteri ili e-klasteri čiji su proizvodi i procesi potpuno digitalizirani. S makroekonomskog stajališta stvaranje e-klastera (eng. *e-clustering*, *cluster-building approach*) je inovativni pristup ekonomskoj politici, a s mikroekonomskog stanovišta je to strategija razvoja e-poslovanja klastera (eng. *e-cluster strategy*) [6]. **e-klasteri** (eng. *e-cluster*) obuhvaćaju procese kooperacije (eng. *processes of cooperation*) tj. tokove digitalnih proizvoda, sve (digitalizirane) procese, elektroničke transakcije i kooperativne odnose, te proces inoviranja proizvoda i procesa (eng. *processes of innovation*) i procesno orijentirano upravljanje znanjem (eng. *process oriented knowledge management*) – v. sl. 3 [7][8]. U tom smislu je e-clustering koncept jačanja starih klastera kojim se kroz stalno inoviranje lanca stvaranja vrijednosti pospješuje njihov rast, što je izuzetno važno za rast gospodarstva i

održivi razvoj na područjima opustošenim ratom, pogođenih ekonomskom krizom ili drugim nedaćama [9].

e-klaster kao oblik poslovne mreže može omogućiti brži rast postrojbenih klastera kroz bolje iskorištavanje znanja i inovativnog potencijala članova klastera i stalno unaprjeđivanje lanca stvaranja vrijednosti klastera (stvaranje nove dodane vrijednosti), što parcijalna e-rješenja ne omogućavaju jer ne obuhvaćaju procese inoviranja i upravljanja znanjem. Kod ostalih modela klastera je također ključno stvaranje dodane vrijednosti, tako da je za velike tvrtke formiranje e-klastera bolje rješenje od pukog integriranja malih tvrtki u njihov sustav e-poslovanja i stvaranja konvencionalnih kooperacija.

B. Koncept e-klastera

e-klaster je poslovna mreža koja omogućava horizontalno i vertikalno povezivanje članova klastera u e-lanac stvaranja vrijednosti i/ili stvaranje novog vrijednosnog lanca, upravljanje ovim lancima i procesom njihovog inoviranja, te upravljanje znanjem klastera.

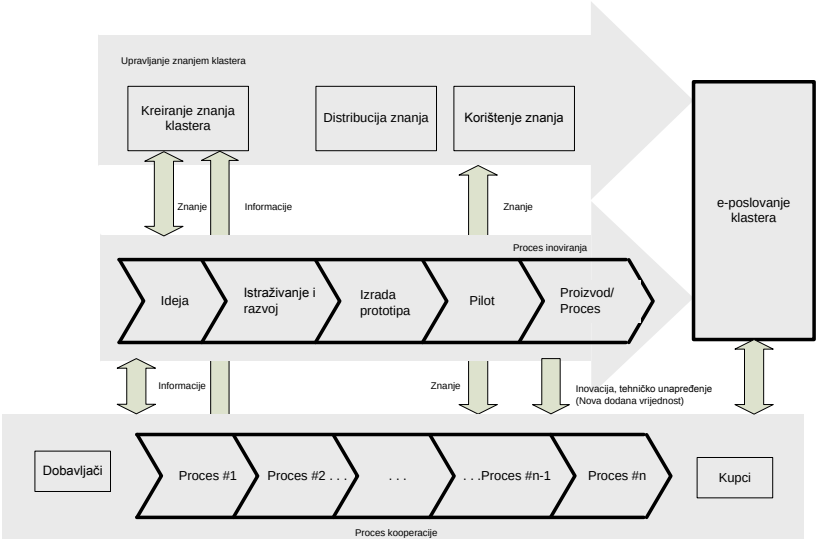
Lanac stvaranja vrijednosti klastera (eng. *value added chain*) su stupnjevi transformacije kojima protječe proizvod i usluga od dobivanja materijala i drugih potrebnih ulaza do konačne upotrebe. To je lanac aktivnosti članova klastera koji stvara dodanu vrijednost za kupca. U klasteru lanac stvaranja vrijednosti tvore članovi klastera i njihovi partneri, a svaki sudjelujući član ima svoj vrijednosni lanac. **Vrijednosni lanac klastera** (eng. *value chain*) je proces stvaranja vrijednosti u klasteru; to je postupak raščlanjivanja procesa kooperacije u strategijski relevantne aktivnosti koji se koristi za dijagnosticiranje konkurentske prednosti klastera.

Ako članovi klastera i njihovi partneri formiraju dobavne i distribucijske e-lance u kojima se obavlja razmjena digitalnih proizvoda, informacijskih usluga i informacija uz IKT potporu i u skladu s utvrđenim kooperacijskim odnosima i normama e-poslovanja, onda oni formiraju lanac stvaranja vrijednosti koji možemo, zbog isključive primjene elektroničkih transakcija u razmjeni dobara i usluga, nazvati elektroničkim (e-lanac stvaranja vrijednosti). Takav sustav e-poslovanja možemo nazvati e-klasterom ako obuhvaća i proces inoviranja digitalnih proizvoda i digitaliziranog procesa kooperacije čiji rezultat su inovacije koje povećavaju produktivnost članova klastera i podupiru budući rast produktivnosti i širenje poslovanja klastera, što je po Porteru [1] glavna odlika klastera. S obzirom da se u procesu inoviranja koristi postojeće i stvara novo znanje, te da je za uspješno uvođenje inovativnih proizvoda i procesa potrebno distribuirati potrebna znanja, e-klaster obuhvaća i upravljanje znanjem (e-knowledge management).

Proces kooperacije (v. sl. 3) određen je vrijednosnim lancem i lancem stvaranja vrijednosti klastera i čine ga proizvodni i logistički procesi u koje su uključene i IKT podržane aktivnosti razmjene dobara, usluga i informacija. Stupanj digitalizacije procesa kooperacije ovisi o opsegu implementacije e-poslovanja u klasteru i interoperabilnosti sustava e-poslovanja. To znači da bi e-klaster trebao podržavati sigurnu razmjenu dobara, usluga i informacija u procesu kooperacije i e-poslovanje članova klastera u skladu s njihovom ulogom u lancu stvaranja vrijednosti (uostalom, e-servisi su jeftiniji od vlastitih e-rješenja članova klastera koje treba integrirati u sustav e-poslovanja klastera).

Znanje je kritičan faktor poslovnog uspjeha klastera. **Organizacijsko znanje klastera** (eng. *cluster organisational knowledge*) obuhvaća znanja članova klastera i zaposlenika, znanje projektnih timova stvoreno u realizaciji kooperacijskih projekata i zajedničkih proizvoda klastera i znanje uključeno u proces realizacije proizvoda i projekata. **Upravljanje znanjem** je funkcija menadžmenta klastera koja obuhvaća kreiranje, distribuciju i korištenje organizacijskog znanja klastera.

Kreiranje znanja (eng. *knowledge production*) obuhvaća kreiranje novih ideja i pogleda, inovacije



Slika 3. e-klaster [7][8][10]

proizvoda, usluga i procesa i oblikovanje tržišnog nastupa kroz individualno i grupno učenje te prikupljanje znanja iz vanjskih izvora. Obuhvaća prikupljanje informacija iz procesa kooperacije i vanjskih izvora, pretvorbu znanja, izražavanje deklarativnog i proceduralnog znanja i validaciju znanja klastera [10]. **Proces inoviranja** obuhvaća razvoj inovacije od ideje, preko istraživanja i razvoja, izrade prototipa, pilot projekta do uvođenja proizvoda i novog procesa u proces kooperacije (v. sl. 3). U e-klasterima ga podržava podsustav kreiranja organizacijskog znanja klastera, koji se organizira kao kooperacijski projekt na razini članova klastera ili funkcije razvoja klastera, u suradnji sa znanstvenim i visokoškolskim institucijama i drugim partnerima ili u sklopu klastera znanja.

U e-klasteru posebnu važnost imaju procesi distribucije i korištenja znanja. **Distribucija ili podjela znanja** (eng. *knowledge difussion, knowledge share*) u e-klasteru obuhvaća deklarativno i proceduralno organizacijsko znanje, koje član klastera preuzima i ugrađuje u svoj proces i aktivnosti povezane s realizacijom zajedničkih proizvoda, kooperacijskih projekata i nastupom na tržištu. **Proces korištenja znanja** (eng. *knowledge apply*) obuhvaća primjenu deklarativnog i proceduralnog znanja kroz cijeli lanac stvaranja vrijednosti.

U **deklarativno znanje** (eng. *declarative knowledge*) spadaju kognitivna znanja (*know-what*), sustavno razumijevanje proizvoda (*know-why*) i sustavno razumijevanje tržišta (*know-what, know-why, know-where, know-when*). Osnovni izražaji deklarativnog znanja klastera su analize tržišta i studije, strategije, specifikacije zajedničkih proizvoda i usluga i organizacijski modeli ili strukture. U **proceduralno znanje** (eng. *procedural knowledge*) spadaju vještine i sposobnosti primjene kognitivnog znanja u dizajnu, konstrukciji i izradi proizvoda, prodaji i odnosima s kupcima, poslijeprodajnim uslugama i sl. (*know-how*), poznavanje tko ima i zna uspješno primijeniti takva znanja (*know-who*), gdje i kada (*know-where* i *know-when*). Ključni izražaji proceduralnog znanja su opisi logističkih i drugih procesa (poslovnici, procedure) [10].

Kroz stalno inoviranje proizvoda i kooperacijskog procesa i tehnička unaprjeđenja, što uključuje i postupni razvoj e-poslovanja, spiralno raste organizacijsko znanje klastera koje podržava budući rast klastera i povećanje produktivnosti njegovih članova uz relativno mala ulaganja, što je naročito pogodno za postrojbene klastere koji raspolažu s vrlo ograničenim resursima.

C. Povezivanje ICT tvrtki u e-klaster

e-klasteri omogućavaju jednostavno horizontalno i vertikalno povezivanje malih i srednjih ICT tvrtki radi zajedničkog nastupa na tržištu i realizacije zajedničkih proizvoda i kooperacijskih projekata. ICT tvrtke mogu imati sljedeće koristi od uključivanja u e-klaster: brže uvođenje složenih ICT proizvoda i usluga na ciljana tržišta, manji transakcijski troškovi, rast prihoda, manji troškovi uvođenja i primjene novih informacijskih tehnologija, rast kvalitete proizvoda i brži rast. Po uključivanju u e-klaster tvrtke mogu očekivati probleme povezane s kooperacijskim odnosima koji su i odnosi suradnje i nadmetanja među članovima klastera. Ipak, glavni izazov u stvaranju e-klastera je izgradnja sustava upravljanja znanjem koji će održavati spiralni rast znanja i konkurentske snage e-klastera kroz stalno inoviranje proizvoda, usluga i procesa kooperacije. Razvoj e-klastera bit će usporen bez potpore makroekonomske politike i poticajnih mjera za jačanje konkurentnosti ICT sektora.

V. ZAKLJUČAK

Sustavi e-poslovanja ne osiguravaju dugoročnu konkurentsku prednost klastera, jer izravno ne podupiru budući rast produktivnosti članova klastera i klastera u cjelini. U e-klasteru je to omogućeno kroz razvoj vrijednosnog lanca koji se zasniva na prepoznavanju aktivnosti koje donose strategijsku prednost klastera na ciljnom tržištu primjenom sustava upravljanja znanjem na što se nadovezuje unaprjeđenje lanca stvaranja vrijednosti klastera kroz proces inoviranja proizvoda i procesa i stvaranja nove dodane vrijednosti za kupca, u što spada i razvoj e-poslovanja. Ključnu ulogu u e-klasteru ima podsustav upravljanja znanjem, jer obuhvaća proces inoviranja proizvoda, procesa (uključivši i e-poslovanje) i nastupa klastera na tržištu, dijagnosticiranje konkurentnskih vrijednosti na kojima počiva vrijednosni lanac klastera i prioritete u razvoju procesa kooperacije kao lanca stvaranja vrijednosti u klasteru.

LITERATURA

[1] M.E. Porter, "Clusters and the New Economics of Competition" , Harvard Business Review, pp 77-90, November-December 1998.

[2] M. Davidović,"Distribucijski e-kanali klastera",XI znanstveni skup "Poslovna logistika u suvremenom menadžmentu",Osijek, 2011.

[3] M. Davidović, „Poslovni klasteri i e-poslovanje", 11. Konferencija o elektroničkom poslovanju e-biz2011, 2011, Zagreb

[4] R. Kalakota, M. Robinson,"e-Poslovanje 2.0",MATE,2002, Zagreb

[5] C. Boja,"Cluster Models, Factors and Characteristics", International Journal of Economic Practices and Theories", Vol.1, July, 2011

[6] U. Hansen, "e-clustering – an innovative approach for economic policy", Refereed Paper No. 650. ERSA – European Regional Science Assotiation Congress, Porto, 2004

[7] U. Hansen, "Innovation through process oriented knowledge management", 45th Congress of the Regional Science Assotiation, Amsterdam, August, 2005

[8] U. Hansen, "Innovation durch prozessorientiertes Wissensmanagement in einem e-cluster", Bovenau Research Publications No. 2, July, 2004.

[9] A. K. Bangura,"Using e-clustering to reinvigorate the Mano River Union", GLOBELICS 2009, 7th International Conference, 6-8 October, Dakar, Senegal.

[10] M. Davidović, "Uloga upravljanja znanjem u nastupu klastera na domaćem i stranom tržištu", XII znanstveni skup "Poslovna logistika u suvremenom menadžmentu", Osijek, 2012.

Electronic commerce in Romania features and dynamics

D. Enachescu, M. Zaharia

"Petrol-Gaze" University/Modeling, Economic Analysis and Statistics Department, Ploiesti, Romania
denachescu22@yahoo.com
marianzaharia53@gmail.com

Abstract: Companies have adapted to the needs and desires of increasingly demanding customer which resulted in the exceeded limits of the traditional business. Small or big, these have adapted to technological progress and understood the major advantages of using Internet in business development, electronic commerce enabling companies to sell products worldwide, increasing the number of potential customers by removing geographic barriers.

This paper presents the development and perspectives of e-commerce in Romania, related to Europe situation and the former socialist countries, evolution which is closely related to the outstanding increase in the number of Internet users and the accessibility of the medium, surprising also the particularities of this phenomenon. In a Europe in crisis, expansion and high growth of electronic commerce in Romania may seem unexplainable. The paper will also try to present the causes of these features.

I. INTRODUCTION

Electronic commerce in the Information Society is a basic element of the new economy and Internet is the main vector but also an informational and technological environment in which it makes its presence felt. Electronic commerce, a component of e-business, involves the use of networks and especially the Internet, digital communications and software, in electronic transaction based economic activities and associated infrastructure in the process of buying and selling.

With only a third of the world population that uses the Internet in 2012, growth prospects based e-business activities in the years to come are more optimistic and have obviously a real trend developing more powerful than other traditional economic activities.

The evolution of e-commerce is closely related to increasing Internet users and accessibility of this medium of communication, trends captured in numerous studies and statistics. Were used in this paper data from Eurostat databases[1] and Internetworldstats[4].

II. WORLD INTERNET USAGE AND GROWTH BETWEEN 2000-2012

A global analysis of Internet users in 2012 shows notable differences between regions. Thus, we see that the highest percentage of Internet users is found evident in most developed areas of the world: North America, Australia and Europe (63.2 to 78.6%) and the lowest in Africa (15.6%). Global rate of 34.3% (one third) of the world population uses the Internet (fig. 1).

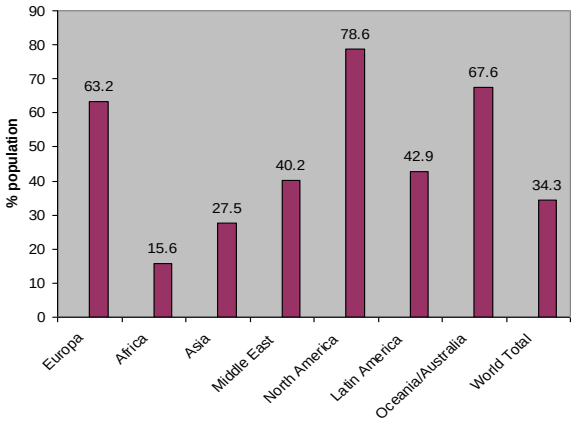


Figure 1. World Internet Usage and World Population Statistics , June 30, 2012
Data source: http://www.internetworldstats.com/stats.htm[4]

Very interesting is the dynamics of the number of Internet users worldwide in the period 2000-2012. Obviously they were absolutely spectacular growth in less developed areas of the world (over 3600% and 2600% in Africa in Middle East) and only 153% in North America. Global increase in the number of Internet users was 566% (fig.2).

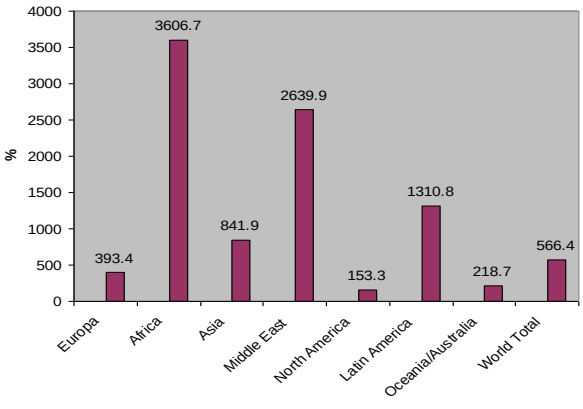


Figure 2. World Internet Usage Growth between 2000-2012
Data source: http://www.internetworldstats.com/stats.htm[4]

From this point of view, Europe, with a penetration rate of 63.2%, is evident in most developed areas echelon but do not forget that the former socialist bloc countries had not practical an IT&C infrastructure by 1990 and have recovered dramatically (for countries like Czech Republic, Poland, etc.) and even Romania and Bulgaria

are striving to achieve a high level (over 50%).

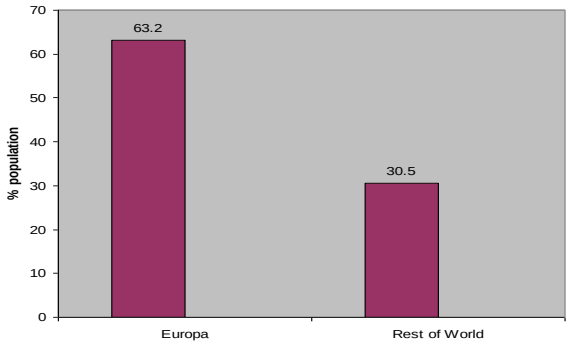


Figure 3. Internet penetration (% population) June 30, 2012
Data source: http://www.internetworldstats.com/stats.htm[4]

III. INTERNET ACCESS IN EUROPA AND EU COUNTRIES (icluding ROMANIA)

According to studies made by Eurostat, a statement of the level of Internet access for households in Europe and EU countries, in the period 2004-2012 and a population aged 16-74 years, is presented below (tabel 1). Countries were grouped so as to observe major differences in the dynamics growth of the number of households using the Internet between the EU developed countries and the countries of the former socialist bloc. But, there are also notable differences between countries in northern Europe and other countries in the EU. Thus, if the Nordic countries have a share of over 90% of the level of internet access, for the rest of the EU countries the percentage varies between 54% for Greece and 85% for Germany. Former socialist countries are differences between the 75% level of internet access for Estonia, Slovakia and only 51% for Bulgaria and 54% in Romania. But it should be noted that these countries had not an IT&C infrastructure until 2000 and were able to recover fast, mainly because private companies in the IT & C area started a nationwide systemic offensive.

TABEL 1. Level of Internet access - households (%)

	2004	2006	2007	2009	2010	2011	2012
EU (27 countries)	**41**	**49**	**55**	**66**	**70**	**73**	**76**
Norway	60	69	78	86	90	92	93
Sweden	:	77	79	86	88	91	92
Iceland	81	83	84	90	92	93	95
Denmark	69	79	78	83	86	90	92
Finland	51	65	69	78	81	84	87
United Kingdom	56	63	67	77	80	83	:
Belgium	:	54	60	67	73	77	78
Germany	60	67	71	79	82	83	85
France	34	41	55	69	74	76	80
Ireland	40	50	57	67	72	78	81
Spain	34	39	45	54	59	64	68
Italy	34	40	43	53	59	62	63
Portugal	26	35	40	48	54	58	61
Greece	17	23	25	38	46	50	54
Poland	26	36	41	59	63	67	70
Czech Republic	19	29	35	54	61	67	71
Latvia	15	42	51	58	60	64	69
Lithuania	12	35	44	60	61	62	62
Estonia	31	46	53	63	68	71	75
Slovenia	47	54	58	64	68	73	74
Slovakia	23	27	46	62	67	71	75
Hungary	14	32	38	55	60	65	69
Croatia	:	:	41	50	56	61	66
Romania	6	14	22	38	42	47	54
Bulgaria	10	17	19	30	33	45	51

Note: Percentage of households who have Internet access at home (all forms of Internet use).
Data source: http://epp.eurostat.ec.europa.eu/[1]

An analysis of the Romania situation show a dynamic growth (9 times) for the number of households connected to the Internet, from 2004 to 2012, due to the direct contribution in infrastructure development through internet and telecommunication private companies and also in conditions of nonexistence virtually an information infrastructure up in 1990 and timid steps in its development until 2000 (fig.4). Basically, in Romania one can speak of electronic commerce by the year 2000. In fact, the legislation in the field emerged in 2003 (Electronic Commerce Act).

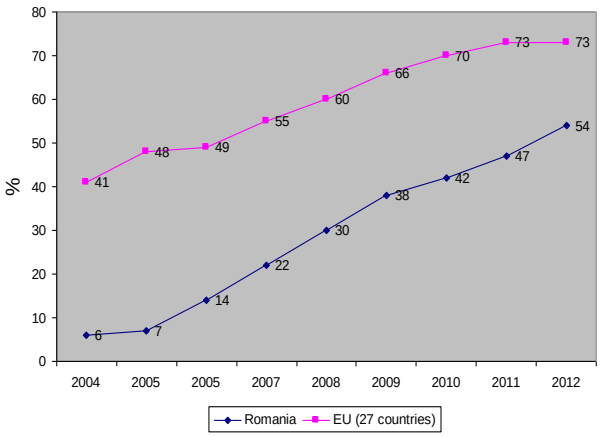


Figure 4. Level of Internet access - households %.
Percentage of households who have Internet access at home. All forms of Internet use are included. The population considered is aged 16 to 74.
Data Source: http://epp.eurostat.ec.europa.eu/[1]

Regarding the dynamics of Internet users in Romania, the graph below (fig.5) shows a dramatic increase of over 12 times only in 12 years but with a very pronounced jump between 2000 and 2004. Obviously, you can see the huge growth potential for companies involved in the development of ICT infrastructure as Romania, with a rate of only 54% connectivity and only 44% of the population is lagging behind other developed countries in Europe. These countries have levels above 70 .. 80%, even 90% connectivity (Sweden (92%), Norway (93%), France (80%)) and the average in European countries is 63.2% of the population.

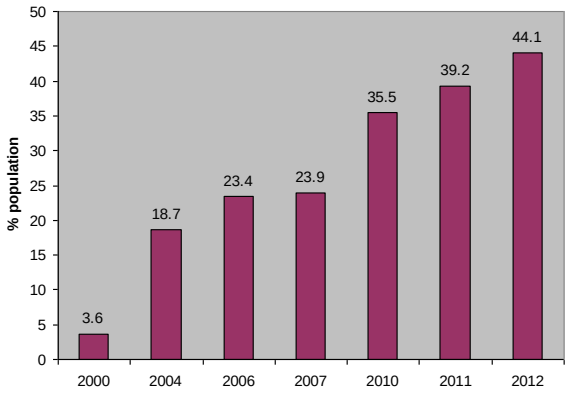


Figure 5. Romania, Internet users penetration (% of population)
Datasource: http://www.internetworldstats.com/stats.htm[4]

IV. E-COMMERCE IN ROMANIA

Electronic commerce has many advantages but also some disadvantages that can be analyzed from three points of view: the company, the consumer and society.

E-commerce activities take place all over the world, e-commerce is essentially global in both as a concept and as an achievement, is very quickly and closely following the exponential growth of the Internet and World Wide Web.

Regarding the impact on firms, e-commerce offers unique opportunities for business reorganization, redefinition of markets or create new markets both locally and especially internationally. E-commerce initiatives can generate cost reductions, increases in revenues and operational efficiency for companies looking to gain an edge in today's competitive economic environment. Modern enterprises are characterized by applying increasingly larger and more diverse demand, the existence of global competition and even increasing customer expectations. To meet these requirements, companies across the globe are in the process of organizational change and adapting their functioning. Electronic commerce is a way on a global scale, which facilitates and supports these changes.

In terms of how the companies understand to develop their business through e-business facilities are presented in dynamic following indicators: Share of enterprises' turnover on e-commerce (fig.6), Enterprises having purchased online (fig.7), Enterprises having received orders online (fig.8), in period 2003-2012 for Romania and some of the former socialist bloc countries compared with the EU average .

This analysis is done for some of the former socialist bloc countries and there are major differences. Thus, if in the case of Czech Republic and Hungary were exceeded growth dynamics of the EU average and in all former socialist countries e-commerce is booming, Romania and Bulgaria are still in the early development of this sector with a dynamic upward but still shy (fig.6).

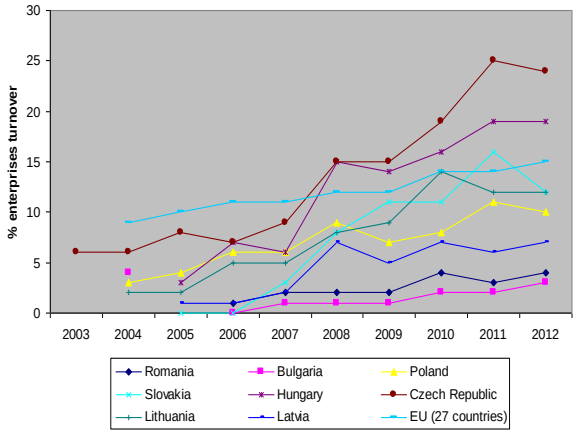


Figure 6. Share of enterprises' turnover on e-commerce (%). Enterprises' receipts from sales through electronic networks as percentage from total turnovera and with at least 10 persons employed
Data source: http://epp.eurostat.ec.europa.eu/ [1]

E-commerce has developed in Romania since the end of 1999 and has a very important growth dynamics so far, e-commerce is still near the beginning of the maturation phase. Even though now there is a huge increase in the number of companies but also in transactions in electronic commerce, fewer e-stores will survive this transition period and just over five years we can speak of a stable e-commerce market if economic crisis will not cause an economic collapse.

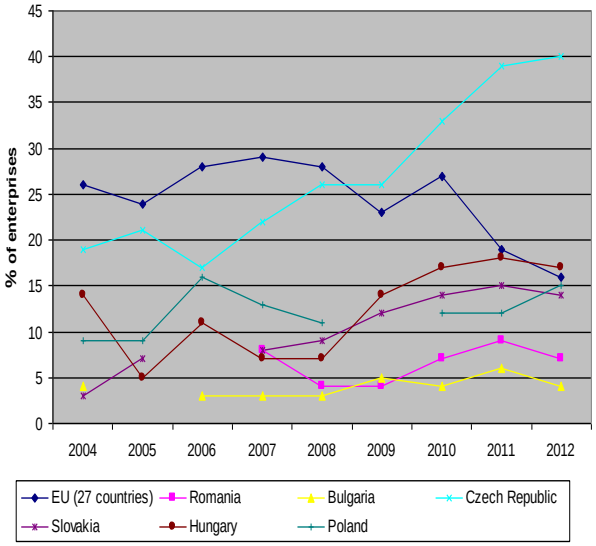


Figure 7. Enterprises having purchased online (% of enterprises with at least 10 persons employed)
Data source: http://epp.eurostat.ec.europa.eu/ [1]

If the economic crisis is seen in the number of EU average countries enterprises by reducing purchases and online orders since 2007 and more dramatically in the last two years, in the former socialist countries it was a generally increasing trend, even impressive for Czech Republic (a doubling of the number of enterprises having purchased and received orders online in 2007 compare to 2012 (fig. 7 and fig.8).

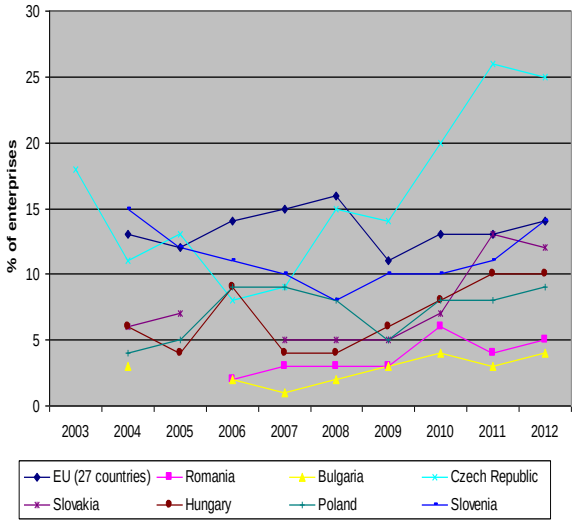


Figure 8. Enterprises having received orders online (% of enterprises with at least 10 persons employed)
Data source: http://epp.eurostat.ec.europa.eu/

In Romania, after a fall in orders received by enterprises having purchased and received orders online in 2007-2008, there was a significant increase in the number which is not related to the overall revitalization of the Romanian economy but due to increased appetite for buying for companies and individuals (fig.8). But we must not neglect another important aspect, namely that state companies were required to buy through the online auction SEAP.

There are multiple indicators that analyze consumer report from e-business activities. Thus, the graph below shows the dynamic number of individuals who used the Internet to order goods or services, aged 16-74, for socialist bloc countries (including Romania) and to the EU average (fig.9).

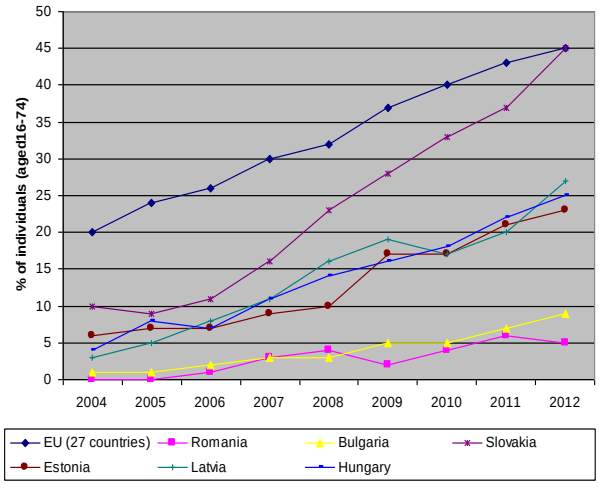


Figure 9. Individuals using the Internet for ordering goods or services (aged 16-74). Buy or order for private use.
Data source: http://epp.eurostat.ec.europa.eu/ [1]

Note that Romania and Bulgaria are both very much lagging behind the EU average and towards the countries of the former socialist bloc. Situation is determined both by the lack of consistent communication infrastructure, in the country such as rural areas are virtually no possibility of access to the Internet but also by a lack of culture and confidence of buyers use in the Internet and online payment methods. It can be said that the romanian people are both ignorant and traditionalists. Anyway, you can see a huge potential for development in this sector that has not been affected by the economic crisis and is in constant growth.

At the consumer level are multiple advantages of e-commerce, such as to trade 24 hours a day throughout all the year, from almost any location, with multiple choice, at the cheapest price for a product or service, rapid delivery of products and / or services (in some cases, consumers can interact with other buyers through electronic communities).

At the level of human society, the benefits of e-commerce company involves increasing the number of jobs, as EU representatives said „that internet-based economy creates 2.6 jobs for every job "offline" and online commerce could generate gains 204 billion". Moreover, e-activities (which includes e-business or e-commerce) is a real form for the sustainable development of society by developing and using systems that consume less resources and less polluting[2].

V. CONCLUSION

In Romania, the volume of e-commerce online payment recorded in 2011 an increase of 24% compared to 2010, reaching a value of 158.9 million euros, according Romcard. In 2011, the Romanian people have made 1.4 million online payment more than in 2010, growth is higher than 2009-2010 when progress was of 1.1 million transactions[2].

In 2012, e-commerce online payment Romcard, the main payment gateway in Romania, increased by 38% compared to 2011, reaching the value of 220 million and a total of 4.1 million transactions[3].

According to statements of Andreia Stavarache, Sales & Marketing Manager PayU, "in nearly seven years, online payments volume in Romania increased from 10 million to over 165 million, keeping constant the rate of double-digit growth even in times of crisis"[5].

At the event GPEC (eCommerce Awards) Romania 2012 was presented a comprehensive report on e-commerce in Romania, from which we extract these statistics[7]:

- Internet penetration rate is about 43% at national level and over 1 million Internet users in Romania are online shoppers (a total of over 9 million Internet users in Romania) of which 83.5% online shoppers live in urban;
- over 3,500 online stores are available in the e-market and the average age of buyers is between 25 and 34 years old, 54.8% of buyers are men and 45.2% are women and average salary of online shoppers is between 1,400 and 1,800 lei (about 300-500 EUR);
- e-commerce share of Total Retail is about 1.3% (very small) and the total amount is about 600

million EUR total e-tail (except tourism, services, utilities etc.). Over 90% payment on delivery is in e-tail (except services, tourism and so on, where the card payment sector is high compared to average).

But the forecast for 2013 are over 750 million EUR total e-tail so the growth prospects over 25%[7].

Electronic business and electronic commerce has many advantages but also have a number of limitations. The main disadvantages of electronic business transactions are related to security and consumer confidence in electronic business and Internet access is still expensive and / or inappropriate for a large part of the population.

In recent years there has been an increase of electronic business in our country and the trend is growing, e-commerce in Romania has not been touched by the crisis. Trend will be with big steps since Romania has to recover huge difference from EU countries. This will cause a new Eldorado for both companies involved in the development of ICT infrastructure as well as direct participants in the process, individuals or companies.

If in the rest of the developed countries of Europe the economic crisis has had an great impact in slowing rhythms of growth, Romania has a huge potential for the development of electronic business in general and e-commerce in particular and is a real business opportunity in this part of Europe.

REFERENCES

[1] http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database

[2] http://www.business24.ro/articole/comert+electronic+rom.

[3] http://www.business24.ro/internet/cumparaturi-online/comertul-electronic-cu-plata-online-220-milioane-de-euro-in-2012

[4] http://www.internetworldstats.com/stats.htm

[5] http://www.business24.ro/internet/plati-online/comertul-electronic-din-ce-in-ce-mai-profitabil-volumul-platilor-a-crescut-cu-38-la-suta-in-2012-1519953

[6] IE. Apavaloaie, *Evoluţia Comerţului Electronic in România*, rev. Oeconomica, Univ. Petru-Maior Tg Mures, vol IV, nr.4. dec.2010, http://www.academia.edu/1443298/EVOLUTIA_COMERTULUI_ELECTRONIC_IN_ROMANIA

[7] http://www.gpec.ro/blog/e-commerce-romania-2012

Web 2.0 tools support for SME start-up projects

Ivana Ogrizek Biškupić *, Igor Vrečko**, Krešimir Jurina ***

* Accredited College of Business and Management Baltazar Adam Krčelić, Zaprešić, Croatia
** Faculty of Economics and Business, Maribor, Slovenia
*** Polytechnic of Zagreb student, Specialist graduate professional study in IT Technologies programme, Zagreb, Croatia

ivana.ogrizek@vspu.hr , igor.vrecko@uni-mb.si , kjurina@tvz.hr

ABSTRACT - According to HTML development and Web 2.0 paradigm and considering business workshops in two year class project in higher education level, we have observed huge disproportion in awareness of free access and ICT support that web 2.0 provides. People who were participants in these workshops had minimum 5 years of work experience in business and management field (part time students of College of Business and Management „Baltazar Adam Krčelić"). Conclusions from these workshops had shown that most of the Croatian small medium enterprises SME spend huge amount of provided budget for unnecessarily part of ICT support. Considering these facts we have also noticed low level use of free open source Web 2.0 tools, partly, coming from not having an idea or not understanding Web 2.0 potentials and possibilities.
The main focus in this paper is on SME who need technology support for various purposes by conducting research with the aim of developing and implementing lifelong learning curriculum as support for SME business start-up projects by using free Web 2.0 tools.

Key words: web 2.0, SME, start-up projects, ICT business support.

I. INTRODUCTION

Small and medium-sized enterprises SME play a central role in the European economy. They are a major source of entrepreneurial skills, innovation and employment. In European Union 23 million SMEs provide around 75 million jobs and represent 99% of all enterprises [1]. However, they are often confronted with market imperfections. SMEs frequently have difficulties in obtaining capital or credit, particularly in the early start-up phase. Their restricted resources may also reduce access to new technologies or innovation.

In order to ensure that its enterprises remain competitive within the Single and global market, Croatian companies need to invest in new processes and new technologies including ICT development, development of new products, including testing and certification to ensure their compliance with international quality standards.

Accredited College of Business and Management Baltazar Adam Krčelić is private founded college with specificity of scientifically-based and practice-oriented studies, adapted to meet the requirements of the labor market and employers in the fields of economy, finances, communications and culture. There are 8 IT courses in 6 study programmes. We have developed curriculum in 3 IT courses to correlate with economy courses as support to entrepreneurship. The main focus of this curriculum development was on finding qualitatively better links within scientifically-based and practice-oriented knowledge and market needs. Participants of these workshops were our part-time students with minimum of 5 years working experience in business and management field. In these workshops we have observed huge disproportion in awareness of free access and ICT support that web 2.0 today provides which we find crucial for early start-up phase of obtaining capital for SMEs. These workshops had shown that most of the Croatian small medium-sized enterprises SME spend huge amount of provided budget for unnecessarily part of ICT support and we have also noticed low level use of free open source, Web 2.0 tools, partly, coming from not having an idea how to use or not understanding Web 2.0 potentials and possibilities at all.

These conclusions leaded us to the project for lifelong learning curriculum development intended to founders, employments or future entrepreneur of micro, small and medium-sized enterprise SME with purpose to inform, teach and provide IT support information of low cost solutions and how to use all the potentials of today's web-based 2.0 and 3.0 possibilities.

II. INFORMATION SUPPORT NEEDS IN SME BUSINESS START-UP PROJECTS

Information gathering and master over information needs proved to be very important for the success of SMEs. Firms that use more information had better business results and they are also expecting better performance in the future [2].

Significance of the information and information technology in SMEs is multi-layered and has attracted a lot of research interest in the past. Chien-Liang Kuo et al [3] showed how SMEs adjust business processes (for example Supply Chain Management) in response to the introduction of Information Systems. Alquier and Tignol [4] analyzed management issues regarding risks and they propose a knowledge base supported by appropriate IS. The risk factors are further researched by Daim [5]. Since multiple risk management tools are available, Leopoulos et al. [6] evaluate sixteen software tools and demonstrate that the use of appropriate software tools may enhance business and project-related operations and reduce costs. Another decision factor in SMEs is the competitive intelligence (CI). Salles [7] proposes a definition of CI and analyses the use of CI by the SMEs. The implementation of reporting systems in SMEs is also researched. Fernandes et al. [8] present a methodology for the balanced scorecard implementation in a SME. Formal knowledge management structures are rarely used in SMEs state Hutchinson & Quintas [9]. Similar reports come from Zhang et al. [10], who researched learning in SMEs and discovered two types of SMEs: the firms, focused on the single individuals and innovative firms, encouraging development of tacit knowledge.

The research efforts in the IS in project-oriented SMEs are largely focusing in supporting communication processes, management processes and knowledge management. Okamuro [11] for instance links the success of cooperative R&D with the quality, quantity and costs of external resources. The SMEs cooperation in R&D is further elaborated by Canavesio & Martinez [12], presenting a model of project oriented fractal company where all the project - and business components can be delegated among project partners. They (ibid.) imply that the enterprise model can be used to specify the requirements of an integrated information and management system for virtual SMEs networking. A case supported information model for the communication support is presented by Paech et al. The research in communication technologies is largely focused on internet based services, to provide, as explained by Redoli et al. [13], combined viewpoint of business and technology.

The investigation in cooperation is upgraded by researching virtual teams [14] and virtual manufacturing [15]. The formation and existence of virtual teams in SME environment is discussed by Matlay & Westhead [16]. Rezgui [14] analyses virtual team-working effectiveness and he identifies issues related to technology adoption, team identification, trust, and motivation and promotes software application hosting, as opposed to the software-licensing model. The virtual manufacturing scheduling problems are explored by Banaszak & Zaremba [15] who propose a heuristic method for addressing scheduling problems in virtual manufacturing.

To support the information needs in a multi-project business environment of SMEs, the information system should deal not only with the internal processes of a SME, but rather support information needs of all involved subjects. To find out which attributes need attention when developing a general IS model for a multi-project business environment, Perko and Vrečko [17] examined the business issues of project members. The main focus should be on the project communication and information sharing; on connecting planning and execution, and on providing diverse views on the project portfolios.

Some of the main issues worth focusing on when looking for IS in SMEs are:

- supporting the management of limited resources, such as a budget, human resources and access to project partners with specific competences;
- connecting business goals and business results with project performance information;
- supporting intra and inter-project communication as well as connecting the project members and partners from within the community;
- supporting multi-task, multi-project workflow planning, execution and management;
- supporting project evaluation and decision making in project portfolio management, upgrading project management tools to project portfolio management tools;
- providing an environment for innovative processes, reducing the costs and risks in using and reusing innovations.

Despite the importance of information and information technology, the process by which SMEs invest in and gain benefit from information systems (IS) is largely unexplored. Levy et al. [18] try to explain this issue, although the variances in their results exist. Due to lack in the research many of the issues concerning the IS structure, the costs and the potential benefits remain unresolved. This often forces the SMEs to use its valuable resources to focus on IS issues, rather than on key business issues.

Similarly and surprisingly, the way SMEs select and use IS did not received important attention in the past, regardless of their importance for SMEs strategic development as well as despite the fact, that information technology is one of the fastest moving field. Consequently, based on the rapid information technology development, information systems useful for business needs have been developing.

The demands for IS properties by project oriented SMEs are elaborated by Vrečko et al. [19], presenting an IS model that supports strategic and commercial multiple project operations. The model delivers information support for a project oriented business SMEs, whereas a model supporting the multi-project business environment is still missing.

III. WEB 2.0 TOOLS AS CONVENIENT SUPPORT FOR SME'S BUSINESS START-UP PROJECTS

Web 2.0 is built on the concept of reuse and relies on open, standards-based technology with required flexibility and scalability in dynamic environments. Web 2.0 technologies enable to create situational applications that rapidly unlock enterprise data and applications for business users and customers. Through aggregating and filtering corporate, individual and external information into customized and relevant business insight, Web 2.0 techniques help reduce information overload. To enable new dimensions of imagination and innovation across the business, IBM believes companies need to develop Web 2.0 approaches around three strategic components: economic impact, community value and enabling technology [20].

Enterprises increasingly rely on widespread teams and networks of people who need real-time access to one another and to common project resources. Ability to easily interact with customers, vendors and partners is critical to ongoing success. Business models based on Web 2.0 approaches create opportunities to adopt traditional sales and marketing strategies and reduce budget costs in future. Looking forward, web 2.0 and coming web 3.0 can potentially enable a much tighter connection between IT transformation and business transformation. Micro, small and medium-sized enterprises SMEs represent a great diversity of companies from high-tech businesses to start-ups tuned towards International markets, from more traditional companies to family businesses.

There's a whole aspect of Web 2.0 that can drive genuine business value and significant competitive advantage. The whole point of Enterprise Web 2.0 is to put best practices for creating Web experiences into the hands of business people, Web designers, and users so that we make the most of the systems, users, and information that we have [21]. Web 2.0 encourages us to put it all online, make it user organizable, findable, and to build a community around it.

Many SMEs in Croatia use commercial software solutions even they are in start-up phase or having financing problems. Even though they don't use open source and web 2.0 potentials enough but still not quitting paying for commercial ones. We have conducted a survey among our students of knowing the basics and potential of web based services that web 2.0 provides and how they can implement or just use them for free in their business or start-up projects and the results have showed (N:296) that 84 % doesn't know for what in business they could use Web 2.0 and couldn't define purpose.

If we analyze predictors of these results we can conclude that there is still big disproportion in understanding the meaning of Web 2.0 and using some of the services. To understand Web 2.0 potentials means putting emphasis on collective creativity of employees who create new contents and knowledge through collaboration and personal contribution.

The central element of enterprise 2.0 is the collective intelligence whereby employees who work on a joint project through the exchange of knowledge and experience contribute to increasing the value of knowledge which leads to easier and faster problem solving [22]. Development of Web 2.0 technologies in business replaces the cooperative form of cooperation (division of labor, where each person is responsible for a certain part of the job) with collaboration (working together in order to solve a problem).

There are several ways that Web 2.0 changed web possibilities, one of them is using concept of folksonomy which represent opposite of taxonomy and is a style of collaborative categorization of sites using freely chosen keywords, often referred to as tags. Tagging allows multiple overlapping associations that the brain itself uses rather than rigid categories which is example for new possibilities in organizing data (Examples: del.icio.us and Flickr). More than ever, Web 2.0 opened access and possibility for anyone to upload free open source software to the web and to comment quality of the product as well.

O'Reilly pointed [23] that infrastructure of the web relies on the peer production methods of open source and by that having an instance of collective, net-enabled intelligence (Linux, Apache, MySQL, Perl, PHP, Python code etc.). There are more than 100,000 open source software projects listed on SourceForge.net. Anyone can add a project, anyone can download and use the code and new projects migrate from the edges to the center as a result of users putting them to work. There are 80 now accessible free software's for Business and Enterprise with focus on Finance, scheduling, Office management, Project Management, e-Commerce, Reporting, Knowledge Management, Modeling etc.

For example cloud computing as form of Web 2.0 is more about evolving business models and how enterprises can pursue new opportunities through ICT. In Final Report for European Commission published in May 2012 about Cloud computing and it impact on e-Skills requirements [24] they analyze future development of cloud computing and role of SMEs. According to report from the 451 Market Monitor cloud computing marketplace will reach $16.7 billion in revenue by 2013 and according to Deloitte cloud-based applications will replace 2.34 pct. of enterprise IT spending in 2014 rising to 14.49 pct. in 2020. The overall conclusion is that

enterprise spending on on-premise solutions will fall as cloud computing reduces the need for licenses, hardware and software. Forrester forecasts that the global market for cloud computing will grow from $40.7 billion in 2011 to more than $241 billion in 2020. The total size of the public cloud market will grow from $25.5 billion in 2011 to $159.3 billion in 2020.

The group of SMEs covers more than 23 million companies in Europe. The costs of external services now are low and they can be tried out on a short-term basis, which means that companies have many alternatives and can gain economic advantages from a cloud solution compared to internal IT. They doesn't have to reinvest into the hardware infrastructure in the business and do not need to buy sophisticated backup systems. This will provide them with long term benefit. There is also greater licensing flexibility as they can pay monthly rather than other accountancy systems with annual fees. Generally SMEs have great potential in Web 2.0 environment by getting their business up to date and by that becoming more competitive and also not negligible in reducing their budget as well.

IV. AN OVERVIEW TO SOME WEB 2.0 TOOLS POTENTIALS

According to McAfee Enterprise 2.0 [24] is a concept of implementation and use of Web 2.0 tools in business. He defines six components of Enterprise 2.0 infrastructure and defines them as SLATES (Search, Links, Authoring, Tags, Extensions and Signals). Enterprise 2.0 and Web 2.0 tools are vital to the survival of today enterprises and it seems almost impossible to name some major company that does not use Facebook or Twitter as a tool to communicate with customers.

In this chapter we provide an overview to some free tools that we recommend and have implemented as workshops in our lifelong learning curriculum for SME employers or owners with aim of getting great benefits by Web 2.0 potentials.

SOCIAL NETWORKING TOOLS

Elgg is software that offers blog, networking, community, collection of news using feeds aggregation and file sharing features. In fact, it is one of the most versatile pieces of software that offers almost everything you need to start a social networking site.

Mahara is designed as a platform to allow its users to demonstrate their life-long learning, skills and development to selected audiences. With Mahara, you can create an e-Portfolio platform and set up an online learner community.

Also recommended: *Loved by Less, Xoops, Around Me*, etc.

WEB CMS (CONTENT MANAGEMENT SYSTEM)

Joomla is an open source content management system (CMS) for publishing content on the World Wide Web and intranets and a model view controller (MVC) Web application framework that can also be used independently.

WordPress is an open source blogging tool and a content management system (CMS) based on PHP and MySQL. It has many features including a plug-in architecture and a template system.

Wiki. Fundamental feature of Wiki system is that all employers of enterprise, regardless of the position, are participating equally in the process in creating new knowledge. As blogs, Wiki systems are also used in the project purposes. Project Wiki systems have more advanced structure then blogs, because they are not limited with chronological publication of information, and they are allowing new content publication to all members of the project team.

Also recommended: *Drupal, Typo3, e107* etc.

WEB DATABASES

MySQL is the world's most used open source relational database management system (RDBMS) as of 2008 that run as a server providing multi-user access to a number of databases.

Mongodb.It is an open source, high-performance, scalable, schema-free & document-oriented (JSON-like data schemas) database. There are ready to use drivers for most popular programming languages like PHP, Python, Perl, Ruby, JavaScript, C++ .

Also recommended: *Apache CouchDB, Neo4j* etc.

OFFICE

Apache OpenOffice (AOO) is an open-source office productivity software suite. It descends from OpenOffice.org (OOo), which was an open-sourced version of the earlier StarOffice. OpenOffice contains a word processor (Writer), a spreadsheet (Calc), a presentation application (Impress), a drawing application (Draw), a formula editor (Math), and a database management application (Base).

LibreOffice is a free office suite, developed by The Document Foundation. It is descended from OpenOffice.org, from which it was forked in 2010. The LibreOffice suite includes a word processor, spreadsheet, graphics editor, slideshow creator, database and math formula writer.

Also recommended: *Neo Office, Google Docs, Koffice* etc.

PROJECT MANAGEMENT

Codendi is an open-source collaborative development platform offered by Xerox. From only one interface, it gathers, all the needed tools for software development teams: management and versioning of code, bugs, requirements, documents, reporting, tests etc. It is mainly used for managing software project processes.

Redmine is a flexible project management web application. Written using Ruby on Rails framework, it is cross-platform and cross-database.

It includes calendar and gant charts to aid visual representation of projects and their deadlines.
Also recommended: *Project Pier, Trac, Project HQ* etc.

LEARNING MANAGEMENT SYSTEM (LMS)
*Docebo***.** In use in corporate and higher education settings. Offers support for a number of different learning models and is compatible with SCORM 1.2 and 2004. It offers interfaces to external systems such as video conferencing and HR systems.
eFront is designed to assist with the creation of online learning communities while offering various opportunities for collaboration and interaction through an icon-based user interface. The platform offers tools for content creation, tests building, assignments management, reporting, internal messaging, forum, chat, surveys, calendar and others.
Also recommended: *Moodle, Dokeos, Claroline* etc.

RSS (Really Simple Syndication) is a family of web feed formats used to publish frequently updated works such as blog entries, news headlines, audio, and video in a standardized format. An RSS document (which is called a "feed", "web feed", or "channel") includes full or summarized text, plus metadata such as publishing dates and authorship. Main feature of RSS is decartelization of content, way of connecting web content using feed of content that is consistent of XML document.

SOCIAL BOOKMARKING allows users to storage, classification, search and share websites that interest them. Tools allow registered users on a remote computer to save bookmarks, tagging whole or only parts of the website, add notes, share them with other users in the network and view bookmarks saved by other users.

PODCASTING. Basic purpose of podcasting is creating your own multimedia files, which can be easily published on the web. Using podcasting in business makes possible to convert business meeting, professional conference or job offer into multimedia format, which can be accessed by each employee or business partner.

DOCUMENT MANAGEMENT SYSTEM
Organizations are leaving a paper form manipulation of documents and transferring to a digital control system. In this way, it reduces the risk of data loss, reduce the need for space, facilitating the management of documents and finding them because system uses its mechanisms for navigation and research of indexed documents that are centrally available.

SOCIAL NETWORK. A social networking service is an online service, platform, or site that focuses on facilitating the building of social networks or social relations among people who, for example, share interests, activities, backgrounds, or real-life connections. SME can through these services create their own profile and present their business offer to potential clients or business partners.

MIND MAPS are diagrams created by the radial reflections that are used for the presentation of generated ideas. Mind maps can be used in all areas of business planning, organization, negotiation, project management etc. The basic advantages of these types of diagrams are the development of new knowledge, saving resources, and effective problem solving.

E-PORTFOLIO is a digital database on the web in which employees store their knowledge and experience in the form of multimedia data, with the goal of sharing this information with others. In this way, companies can better take advantage of a complete inventory of knowledge and experience of their employees and present themselves in a much better way to potential customers, business partners and investors.

CLOUD COMPUTING describes „on-demand“ IT, with infrastructure, applications and data distanced from users and held on remote internet servers. It represents a new consumption model, where software is accessed directly through web browser as if it were installed conventionally on the user's computer. Cloud computing has been discussed and hyped in recent years, but 2011 showed an increased use of cloud solutions in both large and small companies. The analysis [25] shows that SMEs can create business value in relation to business processes in four dimensions, namely business revenue and growth, agility, cost reduction and innovation. Cloud computing can potentially enable a much tighter connection between IT transformation and business transformation. Innovative ICT practitioners and managers think business transformation first, then how technology enables it.

V. CONCLUSIONS AND PROPOSITIONS FOR FURTHER RESEARCH

In the field of support and development of SMEs, Croatia, and similar Slovenia, is still lagging behind the most developed and also average developed EU countries Singer et al. [26]. In the paper we identified better management of information and use of state-of-the-art and relatively cheap or even free information technology and information systems as a promising tool among different possible ways to decrease this lag. While significance of the information and information technology in SMEs has attracted a lot multi-layered research interest, the process by which SMEs invest in and gain benefit from information systems (IS) is largely unexplored. Web 2.0 and forthcoming web

3.0 tools proved to have important role in helping SMEs to overcome their financial and other resources limitations which imped them to acquire necessary information systems. In the paper numerous web 2.0 tools have been shortly presented. The extent of presented web 2.0 tools speaks about a great potential this field have, while actual use of many of them till today was very limited. In the future for sure there is a need to develop a platform which would be used by SMEs as starting point to get to know with different open information systems (free or very cheap) assured through web 2.0 and web 3.0 technology. Further on such platform would need to help SMEs to quickly implement and start to use to specific situation adapted tool. Obstacles perceived by entrepreneurs in the form of lacking appropriate IS tools to start-up and to further develop their firm's would vanish and they would be able to focus themselves more on the making business and not so much on developing support for making business.

VI. REFERENCES

[1] THE NEW SME DEFINITION. User guide and model declaration. Enterprise and Industry publications, European Commission, 2003.

[2] Lybaert, Nadine. The association between information gathering and success in industrial SMEs: the case of Belgium. Entrepreneurship & Regional Development: An International Journal, 10(4), 1998., pp. 335-351.

[3] Chien-Liang Kuo, David; Chen, Wen-Hsien; Smits T. Martin. SME-based collaborative supply chain management: the impact of information technologies. International Journal of Management and Enterprise Development, 2(3-4), 2005., pp. 360-373.

[4] Alquier, A. M. B.; Tignol, M. H. L. Risk management in small- and medium-sized enterprises. Production Planning & Control, 17/3, 2006., 273-282.

[5] Daim, T. U., Ha, A., Reutiman, S., Hughes, B., Pathak, U., Bynum, W., et al. Exploring the communication breakdown in global virtual teams. International Journal of Project Management, 30(2), 2012., pp. 199-212.

[6] Leopoulos, V. N.; Kirytopoulos, K. A.; Malandrakis, C. Risk management for SMEs: Tools to use and how. Production Planning & Control, 17(3), 2006., pp. 322-332.

[7] Salles, M. (2006). Decision making in SMEs and information requirements for competitive intelligence. Production Planning & Control, 17(3), 229-237.

[8] Fernandes, K. J.; Raja, V.; Whalley, A. Lessons from implementing the balanced scorecard in a small and medium size manufacturing organization. Technovation, 26(5-6), 2005., pp. 623-634

[9] Hutchinson, V.; Quintas, P. Do SMEs do Knowledge Management?: Or Simply Manage what they Know? International Small Business Journal, 26(2), 2008., 131-154.

[10] Zhang, M.; Macpherson, A.; Jones, O. Conceptualizing the Learning Process in SMEs: Improving Innovation through External Orientation. International Small Business Journal, 24(3), 2006., pp. 299-323.

[11] Okamuro, H. Determinants of successful R&D cooperation in Japanese small businesses: The impact of organizational and contractual characteristics. Research Policy, 36(10), 2007., pp.1529-1544.

[12] Canavesio, M. M.; Martinez, E. Enterprise modeling of a project-oriented fractal company for SMEs networking. Computers in Industry, 58(8-9), 2007., pp. 794-813.

[13] Redoli, J.; Mompo, R.; Garcia-Diez, J.; Lopez-Coronado, M. A model for the assessment and development of Internet-based information and communication services in small and medium enterprises. Technovation, 28(7), 2008., pp. 424-435.

[14] Rezgui, Y. Exploring virtual team-working effectiveness in the construction sector. Interacting with Computers, 19(1), 2007., pp. 96-112.

[15] Banaszak, Z.A.; Zaremba, M.B. Project-Driven Planning and Scheduling Support for Virtual Manufacturing. Journal of Intelligent Manufacturing, 17/6, 2006, 641-652.

[16] Matlay, H.; Westhead, P. Virtual Teams and the Rise of e-Entrepreneurship in Europe. International Small Business Journal, 23(3), 2005, pp. 279–302.

[17] Perko, I., Vrečko, I. The information requirements and resources in dynamic multi-project business environment. Journal of communications and information sciences, 3(1), 2012., pp. 60-70.

[18] Levy, M.; Powell, P.; Yetton, P. The dynamics of SME information systems. Small Business Economics, 19(4), 2002., pp. 341-354.

[19] Vrečko, I.; Hauc, A.; Cancer, V.; Perko, I. Conducting multi-project business operations in SMEs and IS support Enterprise Information Systems for Business Integration in SMEs: Technological, Organizational and Social Dimensions. Hershey ZDA: IGI Global, 2010.

[20] IBM . The business value of Web 2.0 technology. ftp://public.dhe.ibm.com/pub/lotusweb/web20/10709800_Web_2.0_brochure.pdf (21.01.2013.)

[21] Hinchcliffe, Dion. Creating real business value with Web 2.0. ZD Net, 22.02.2006. http://www.zdnet.com/blog/hinchcliffe/creating-real-business-value-with-web-2-0/14 (17.12.2012.)

[22] Lee, M.R. From Web 2.0 to Conversational Knowledge Management: Towards Collaborative Intelligence, Journal of Enterpreneurship Research, Vol. 2, No. 2, 2007. (pp. 47 – 62)

[23] O'Reilly, T. What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software. 2005. http://oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=1 (11/2012)

[24] McAfee, A. P. Enterprise 2.0: The Dawn of Emergent Collaboration, MIT Sloan Management Review, Vol. 47, No. 3, 2006.

[25] Søndergaard Laugesen, N.; Lauritzen, J. R.; Carpenter, G.; Ellegaard, C. E.; Bucher, M.; Stabe, M. CLOUD COMPUTING, Cyber Security and Green IT (The impact on e-Skills requirements) Final Report, Prepared for European Commission. Danish Technological Institute & Fraunhofer, May 2012. http://ec.europa.eu/enterprise/sectors/ict/files/eskills/e-skills_and_cloud_computing_final_report_en.pdf (21.01.2013.)

[26] Singer, S.; Šarlija, N.; Pfeifer, S.; Oberman Peterka, S. Što čini Hrvatsku (ne)poduzetničkom zemljom? GEM Hrvatska 2002-2011. Osijek: Centar za politiku razvoja malih i srdnjih poduzeće i poduzetništva. 2012.

[27] Financing a Start-up Company in the Czech Republic, Comprehensive guide for entrepreneurs. EXPATS, 2006. http://www.expats.cz/prague/article/prague-business/start-up-companies/ (20.01.2013.)

[28] Orehovački, T.; Konecki, M.; Stapić, Z.. Primjena Web 2.0 tehnologija u poslovanju // CASE 20 - metode i alati za razvoj poslovnih i informatičkih sustava / Polonijo, Mislav (ur.). Rijeka : CASE d.o.o., 2008. (pp. 197-202) http://bib.irb.hr/datoteka/358251.Orehovacki_Konecki_Stapic.pdf (23.11.2012.)

[29] Panian, Ž. Poslovni informacijski sustavi, Mogući doprinos Web 2.0 tehnologija unaprjeđenju elektroničkog poslovanja. Element, Zagreb, 2010. (pp. 262 – 264)

e-Maintenance for ship electrical propulsion plants

Jasmin Ćelić, Aleksandar Cuculić
Faculty of Maritime Studies, Rijeka, Croatia
cele@pfri.hr

Abstract - Modern and efficient maintenance systems use advanced information and communication technologies in order to identify and minimize the causes of failures and eliminate the possibility of unplanned and costly breakdowns. With e-maintenance concept, productivity and effectiveness can be improved because the maintenance procedures are applied to the results obtained by the constant real time monitoring of system parameters. Computerized maintenance management systems allow fast, safe and relatively cheap gathering of the information necessary for the maintenance operations and sharing of knowledge between the parties involved in the maintenance process. By using an architecture oriented services such as e-monitoring, e-diagnostics, e-forecasting, e-logistics, etc., substantial savings can be obtained as a result of improved decision making at all management levels. Real time integration of key technologies that are part of the condition based maintenance and a predictive maintenance strategy with existing operational requirements and resources is a precondition for optimisation. In this paper, authors present an e-maintenance concept for ship electrical propulsion plants.

I. INTRODUCTION

Large industrial power plants like ship's electrical propulsion consist of various hardware (mechanical, electrical and electronics) and software systems, subsystems and components of different type and with different mean time before failure (MTBF). Maintenance strategies based on simple combinations of appropriate corrective and planned preventive maintenance are not particularly suitable for modern plants, therefore more advanced methods and processes such as Reliability Centred Maintenance (RCM), Condition Based Maintenance (CBM) and Predictive Maintenance (PdM) are used.

The need for applying modern concepts and strategies is especially evident in autonomous systems (e.g. ships and aircrafts) where the minimum maintenance downtime and continuous operation during the exploitation phase is expected. Maintenance of such systems must be performed in specific environments where the availability of any support is usually low or non-existent, while at the same time any operational downtime results in very high costs.

One of the solutions for above mentioned specific requirements is e-maintenance concept which ensures the cooperation between all key factors of maintenance process and enables timely and strong support to both technical and management personnel in maintenance related decision making process. This concept is based on the system maintenance according to its condition, and with the help of modern IT and communication technologies provides real time information about the level of degradation and malfunctioning at any time and in any geographical location.

E-maintenance consolidates existing remote maintenance principles with a modern Web services and e-collaboration principles which allow the exchange and sharing of information and knowledge [1].

Information about the system's state are collected by intelligent sensors and transmitted via the appropriated communication channel (wired or wireless) to a computer system equipped with special software solution (CMMS - Computerized Maintenance Management System) for further processing and analysis. Software solution for maintenance management can be a part of the local computer system, company level computer system or a system based on cloud computing (IaaS, Paas, SaaS ...). Processing and analysing of the collected data allows a wide range of different operations, such as e-forecasting, e-procurement, e-planning, etc., and support decision-making processes at different levels of management and expert support.

A computer system can automatically alert an authorized maintenance technical staff about the resulting degradation or malfunction and create a work order with instructions for handling and specified safety precautions. Technical staff can identify degraded or defective components or subsystem by using mobile devices (PDA, Smartphone, Tablet) and receive all the necessary information (technical specifications, schematic diagrams, manuals ...) and the help from an expert or expert system. All these characteristics make e-maintenance concept inevitable in today's maintenance strategies.

Maintenance systems in shipping industry are also changing very quickly in a way to accept the risks arising from the adoption of new maintenance strategies and begin to use e-maintenance concept.

For domestic companies, maintenance costs make about 20% to 30% of total costs [2] while at the global scale this number reaches up to 40% [3], so the implementation of new strategies and concepts for the optimization of maintenance process is desirable. Effectiveness and efficiency of e-maintenance concept is reflected in the fact that implementation cost today are

only about 10% of the total savings that can be achieved in maintenance.

One of the first successful examples is the U.S. Navy which included over 100 of their ships and submarines into ICAS (Integrated Condition Assessment System) project [4]. Effectiveness of implemented strategy is best shown by the fact that the savings of around half a million dollars per ship are achieved during the three years testing period.

Effective monitoring of system status can be performed on various marine facilities (HVAC, power generation, etc...), while this paper primarily points to the multiple reasons for the application of the e-maintenance concept in the electrical propulsion plants for marine vessels.

II. Ship Electrical Propulsion Plant

Fully integrated electrical propulsion systems are generally applied on board vessels that must meet high technological and environmental demands such as large cruise ships, ferries, DP drilling vessels, thruster assisted moored floating production facilities, shuttle tankers, cable layers, pipe layers, icebreakers, supply vessels, navy ships and LNG carriers [5].

Like all mission critical systems, electrical propulsion should operate without interruptions in order to avoid costs and equipment damage related to loss of vessel's manoeuvrability due to propulsion system malfunctions. Proper design and optimization of each propulsion system component during the project phase and choice of appropriate maintenance strategy during the vessel's exploitation phase certainly have a biggest influence on achieving this goal.

Due to its complexity, successful implementation of maintenance strategy for electric propulsion system is not an easy task. It requires a good knowledge base from various fields of competence such as electrical engineering, power engineering, power electronics, control systems, mechanical engineering, hydrodynamics etc. CDM and PdM strategies are impossible to implement without the support from the experts that cover specific fields through remote diagnostics services [6].

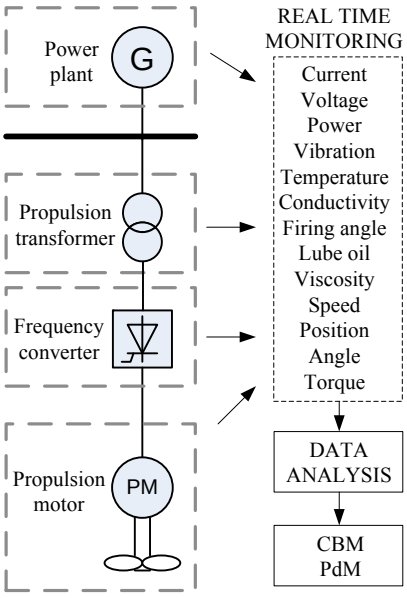


Figure 1. General concept of ship electric propulsion plant with monitored parameters

General concept of ship electric propulsion plant with the most important parameters required to be monitored for CBM implementation is shown in Figure 1.

Monitoring systems must detect all deviations from the component's normal working condition in an early stage, before it can affect overall system's performance in any way (Figure 2). This is especially important for a propulsion motor bearings, swivel bearings, steering gear hydraulics, azimuth propulsor seals and all other components that can be fixed and replaced only when ship is in dry dock, or its repair requires significant time that affect vessels operability and produce high costs for the ship owner.

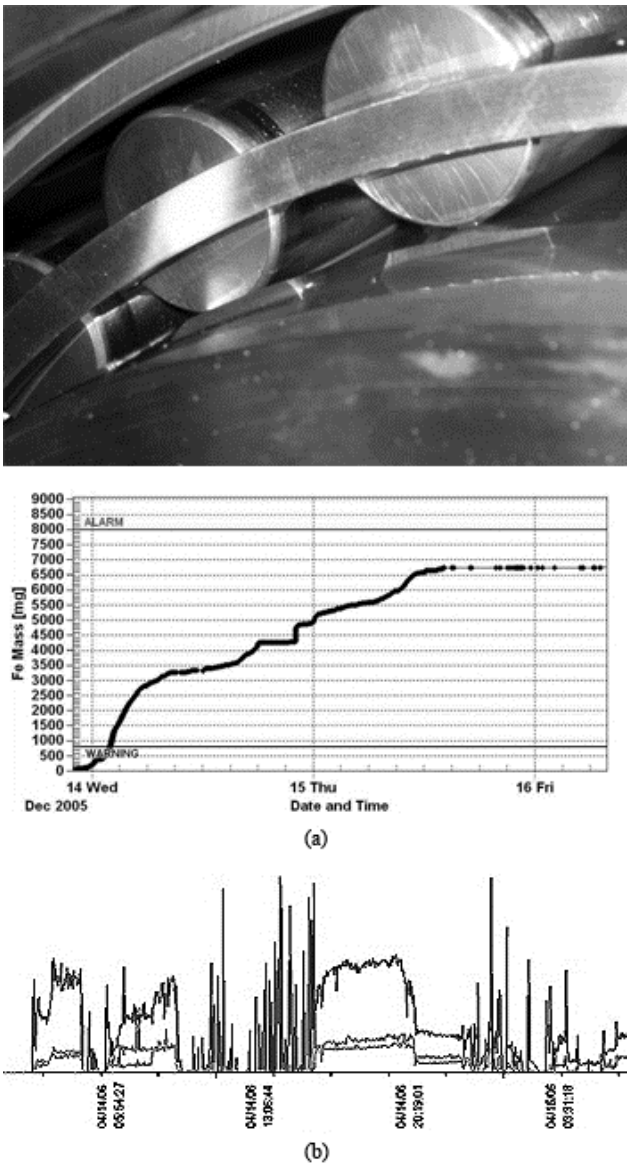


Figure 2. Early detection of propulsion motor bearing damage with combination metal particles scan in lube oil (a) and vibration monitoring (b)

Very often, this requires the use of more than one type of sensor for each parameter, or even the measuring of different system parameters that needs to be combined in order to obtain the condition of individual component (e.g. vibration monitoring and analysis of metal particles in lube oil for propulsion motor bearings).

The analysis of measured data is not only important purely for maintenance purpose but can be also used for analysing the influence of various vessel's operating conditions and manoeuvring procedures on electrical propulsion plant. The results can be used to identify the week spots, to train the crew to operate the vessel in such a way that the function of propulsion system is not compromised and for improving the future design of electrical propulsion systems. On a large time scale, this can also reduce the maintenance and operational costs significantly.

III. E-MAINTENANCE ORIENTED STRATEGY IN ELECTRICAL PROPULSION PLANT

Computerized maintenance management systems on ships usually process data related to the maintenance operations, entered manually by the appropriate staff. Based on the given information about the faults or failures characteristics, measurement results, corrective and preventive actions taken, state of the stack, procurement of spare parts, etc., the system plans appropriate maintenance strategy, generates work orders and allocates resources. These systems, usually called CMMS or EAM (Enterprise asset management), display safety considerations, parts descriptions, system specifications and maintenance procedures, all in accordance with the maintenance strategy chosen by the ship owner.

All systems present on the market have to meet the demands for improved quality and higher reliability, which contributes to the growing costs of preventive maintenance [7]. This problem requires a more effective approach to maintenance, therefore condition based maintenance system which present a base of e-maintenance concept is used. The implementation of this maintenance strategy which should result in the required optimization of maintenance is not possible without the development of appropriate technological solutions. Therefore, a key factor that enables the use of e-maintenance as a powerful technological framework for modern maintenance and asset management of the ship electrical propulsion plants is the development of wireless and mobile technology with Web-oriented software solutions [8].

The arrival of new e-technologies (Web technologies, smart identification tags, smart sensors, wireless communications, mobile devices) provides more efficient maintenance that is no longer based on the sometimes unnecessary preventive maintenance and performed only if it is truly necessary, given the state of the system.

In the Internet of things era, SOA (Service Oriented Architecture) based applications implemented through web services or RESTful services hosted on a cloud are supposed to be an adequate solution. Devices that are part

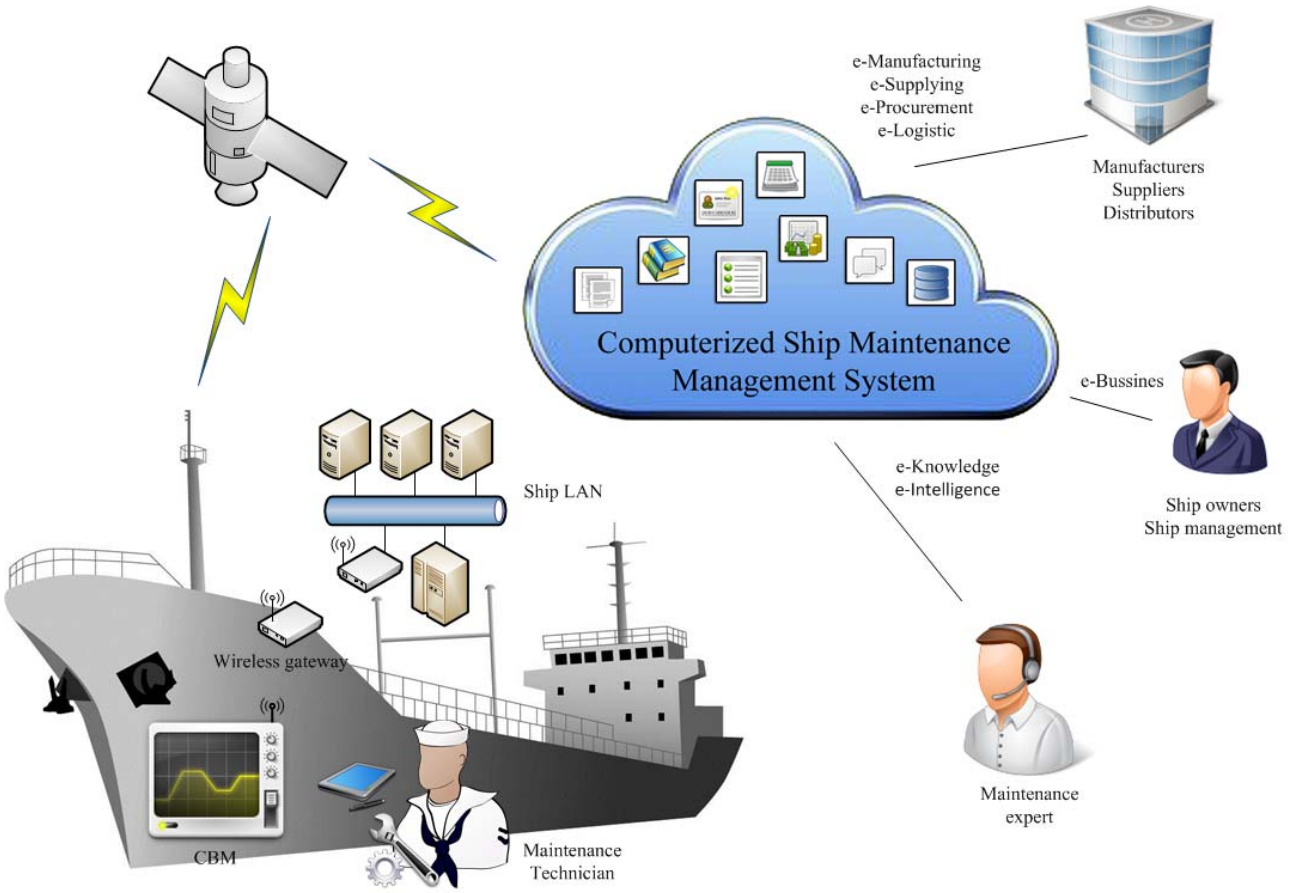


Figure 3. e-Maintenance concept applied on existing Computerized Ship Maintenance Management System

of the collaborative Wireless Sensor Network (WSN) should provide real-time data via SOAP or other protocol corresponding to the cloud-based enterprise application software, thus connecting the real and the virtual world.

The implementation of latest revision of the Internet Protocol (IPv6) with increased addressing ability into WSN has enabled the Integration of embedded devices and systems into the Internet. IPv6 over Low power Wireless Personal Area Networks (6LoWPAN) provides communication over IEEE 802.15.4 based networks (ZigBee, ISA100.11a, WirelessHART), building the basis for a wireless embedded Internet. Wireless sensing components can be now exploited and integrated as a part of online maintenance management systems which are using all advantages of Semantic Web and applications of ontologies in their multi-agent subsystems.

Smart sensors used in CBM systems are usually composed of one or more Microelectromechanical sensors (MEMS), memory and processor, smart tag and wireless communication unit equipped with a long life battery or energy harvesting system. MEMS are highly integrated devices that combine electrical and mechanical components and detect the value of a physical quantity or its change and convert it into a signal for an indicating or recording device [9]. Smart sensors measure key performance indicators of the equipment such as temperature, pressure, and/or vibration and relay the CBM data into the ship's wired communication system and CMMS via the host hub (Figure 3.).

Management of identity and exploitation data through propulsion plant lifecycle, tracking and synchronisation of data between sensors and online databases cannot be achieved without unique identification (ID) technology - smart tags [10]. In CBM system, the Radiofrequency Identification (RFID) technology is typically used.

Usage of mobile devices (PDA, smartphones, tablets) provides flexibility in the daily maintenance operations and helps maintenance personnel (technicians, engineers, maintenance managers) to have real time information about the monitored propulsion plant from different sources. Mobile devices identify monitored plant with RFID reader or QR code scanner and get all relevant data such as the current state of the plant, spare parts availability, maintenance and scheduling actions history, technical specifications, maintenance instructions and so on. Also, video, voice and other communication capabilities of mobile devices allow direct contact with the experts for a particular component or part of the propulsion plant, or access to specific e-applications, expert systems and artificial intelligence systems.

At the same time, these operations and the results of actions taken can be traced at all managerial levels, depending on the authorization and be used as a valuable help in decision making and improving the maintenance system, and thus the business.

IV. CONCLUSION

The development of communication and information technologies has enabled the practical application of e-maintenance concept in maintenance strategies on board marine vessels, when the implementation costs of such approach become justified. Specifically, the cost of electronic equipment necessary for the CBM implementation on which the e-maintenance concept is based has been relatively high, and its widespread use in maintenance systems far beyond the limits of viability.

The main advantage of using the e-maintenance concept in ship electrical propulsion systems is certainly the ensuring of preconditions for quality decision making process at management level through the reports and analytical tools based on real time data. This enables significant savings by prediction of unexpected breakdowns and the propulsion system downtimes, optimized maintenance operations, reduce of the maintenance staff working hours and relatively high education costs through remote diagnostics and support, optimized inventory and spare parts management, reduce of administrative costs and insurance premiums, etc.

All listed indicators are sufficient reason for the systematic application of e-maintenance concept not only in ship electrical propulsion plants but also in the other marine systems.

REFERENCES

[1] Muller, A., Marquez, A.C., Iung, B.: "On the concept of e-maintenance: Review and current research", Reliability Engineering & System Safety 93, 2008., pp. 1165-1187.

[2] Tomas, V., Šegulja, I., Čišić, D.: "Possibilities and problems in applying contemporary maintenance strategies in maritime services", Scientific Journal of Maritime Research, 19, 2005., pp. 29-41.

[3] Alhouli, Y., Ling, D., Kirkham, R., Elhag, T.: "On the Factors Afflicting Maintenance Planning in Mercantile Industry", COMADOM, San Sebastin Spain, 2009.

[4] Hogan, R., Cesarone, T., Dragun, D.: "Battle Group Automated Maintenance Environment", Proceedings of the 13th annual international ship control systems symposium explores automation in ship control, Philadelphia, USA, 2003.

[5] Adnanes, A. K.: "Maritime electrical installation and diesel electric propulsion", ABB Marine, Oslo, Norway, 2003.

[6] Sekaceva, N.: "Upgrading of the third generation of the Remote Diagnostic Systems for ABB Marine Propulsion Systems", Master's thesis, Lappeenranta, Finland, 2007.

[7] Jardine, A.K.S., Lin, D., Banjevic, D.: "A review on machinery diagnostics and prognostics implementing condition-based maintenance", Mechanical Systems and Signal Processing 20 (7), 2006., pp. 1483–1510.

[8] Emmanouilidis, C.; Jantunen, E., Gilabert, E., Arnaiz, A., Starr, A.: "e-Maintenance update: the road to success for modern industry", Stavanger, Norway, 2011.

[9] Holmberg, K., Adgar, A., Arnaiz, A., Jantunen, E., Mascolo , J., Mekid, S.: "E-maintenance", Springer, 2010., pp. 125.-171.

[10] Pátkai, B., Theodorou, L., McFarlane, D., Schmidt, K.: "Requirements for RFID based Sensor Integration in Landing Gear Monitoring – A Case Study", Auto-ID Lab, University of Cambridge, UK, 2007.

Implementing Responsive Web Design for Enhanced Web Presence

S. Mohorovičić
University of Rijeka, Faculty of Maritime Studies, Rijeka, Croatia
sanja@pfri.hr

Abstract - Modern companies, institutions, organizations, individuals, etc. have websites in order to extend their reach to audience or customers. However, it is not sufficient anymore just to have an appearance on web and to be recognized through various web search engines. People are increasingly using smartphones and tablets for accessing the Internet, not just desktop personal computers and notebooks, therefore websites need to be optimized for all these devices in order to provide the best user experience. Responsive web design provides a website with a flexibility to adapt to any of these devices, i.e. their resolutions. The paper presents statistics and predictions of market trends regarding the devices and user experiences in web browsing and m-commerce. Responsive web design is researched, along with its benefits and potential problems.

I. INTRODUCTION

A modern website is a must have tool for any company to increase its visibility towards potential customers. It is common for companies, institutions, organizations and individuals to have websites to reach audience or customers. However, it is not enough just to be present on web and available through web search engines anymore. People are spending increasing amount of time online and they are increasingly using smartphones and tablets for accessing the Internet, so websites need to be optimized for all these devices in order to provide the best user experience. Besides various screen sizes and resolutions, different web browsers and platforms, some differences also exist in the ways users interact with their devices: using a mouse, touching the screen or making movements.

Internet became accessible virtually to anyone and anywhere. However, a number of websites are still not optimized and viewable for all the devices that can be used for web browsing, mainly due to the technology used for website creation.

While some web technologies are near extinction (e.g. Flash), new technologies and web standards are in their rise (e.g. HTML5 and CSS3) to allow websites better adjustment and visibility.

In several articles and popular sites [1] [2] [3] that discuss web design trends, year 2013 has been declared as the year of responsive web design. Predictions are that such web design will be one of the biggest marketing trends in 2013. Responsive web design is a new web design approach that enables flexibility of a website to adapt to any device.

II. WHY DIFFERENT DESIGN IS NEEDED?

Following research results and predictions indicate that a change in traditional web design is needed to meet the market needs.

A. Device Market

According to Gartner's research report with shipments projections for smartphones, tablets, ultramobiles and PCs from 2012 to 2017 (Table I) [4], the combined shipments of PCs, tablets and mobile phones is expected to grow 9% to total 2.4 billion units in 2013 from 2012. In 2012, more tablets and smartphones were sold than PCs. Sale of desktop PCs and notebooks is expected to decline 7.6 % in 2013, while the shipments of tablets and smartphones are expected to grow. The most significant growth is projected for tablets: worldwide shipments are predicted to 197 million units in 2013, which is a 69.8% increase from 2012 shipments of 116 million units.

Device shipments are predicted to continue their growth, reaching more than 2.9 billion units in 2017.

With time, besides having the greater number of total devices, the proportion of particular devices will significantly change: there will be less PCs and more tablets and smartphones. There is a significant shift from PCs to mobile devices in these projections.

Similar research was conducted by International Data Corporation (IDC) [5]. Table II shows their results and predictions. They expect that tablet shipments will surpass desktop PC shipments in 2013 and notebook shipments in 2014.

TABLE I. WORLDWIDE DEVICES SHIPMENTS BY SEGMENT, 2012-2017 (THOUSANDS OF UNITS) [4]

Device Type	2012	2013	2014	2017
PC (Desk-Based and Notebook)	341,263	315,229	302,315	271,612
Ultramobile	9,822	23,592	38,687	96,350
Tablet	116,113	197,202	265,731	467,951
Mobile Phone	1,746,176	1,875,774	1,949,722	2,128,871
Total	**2,213,373**	**2,411,796**	**2,556,455**	**2,964,783**

TABLE II. SMART CONNECTED DEVICE MARKET BY PRODUCT CATEGORY, SHIPMENTS, MARKET SHARE, 2012-1017 (UNITS IN MILLIONS) [5]

Product Category	2012 Unit Shipments	2012 Market Share	2017 Unit Shipments*	2017 Market Share*	2012—2017 Growth*
Desktop PC	148.4	12.4%	141.0	6.0%	-5.0%
Portable PC	202.0	16.8%	240.9	11.0%	19.3%
Tablet	128.3	10.7%	352.3	16%	174.5%
Smartphone	722.4	60.1%	1,516	67%	109.9%
Total	1,201.1	100.0%	2,250.3	100.0%	87.3%

* Forecast estimates

IDC's and Gartner's results show small differences in their projected numbers but the trends of device market changes are common in both their projections.

B. Web Browsing, M-commerce and User Experiences

Web browsing through mobile devices has covered 10.11% of website page views in May 2012 [6]. Mobile internet usage continues to grow. According to TechCrunch [7] and Gartner [8], more people globally will have access to the Internet through mobile devices rather than through desktop computers in 2013. IDC predicts that by 2015, more Internet users will be accessing the Internet through mobile devices than through PCs in the U.S. [9].

Consumers use a variety of devices to reach different goals on the Internet (e.g. browsing the Internet, shopping online, managing finances). Google's research shows that 90% of people use multiple screens sequentially [10], i.e. move between devices.

Since the increasing number of people own mobile devices, more and more companies adapt to mobile Internet usage. By 2016, mobile commerce is expected to grow to 5 times of its current size [8].

According to Google's research of users' expectations and reactions towards their mobile website experiences, 67% of users are more likely to buy a product from a mobile-friendly site so mobile-friendly sites can turn users into customers [11]. Mobile users are five times more likely to abandon the task if the site isn't optimized for mobile usage and 79% of mobile users will search for another site to complete the task [8]. If a company doesn't have a mobile-friendly site, both company's reputation and potential profit are affected, and users will turn to competitors' site.

More statistics about mobile commerce, mobile web, mobile marketing, consumers, mobile payment, etc. can be found in [12] and [13].

III. MOBILE-FRIENDLY WEBSITE CREATION APPROACHES

Content on the web should be accessible everywhere, anytime and with any device: personal computers (desktops and netbooks), smartphones, tablets and televisions. How can it be achieved?

One approach to create mobile-friendly website is to create separate (two or more) versions of the same website, e.g. one for desktop computers, one for mobile phones and one for tablets. For each type of site, different URL is given. If a website, which has a mobile and a desktop version, is accessed on tablet, a mobile version of the website will most likely be loaded. Tablets are bigger than smartphones and that version of website is not optimized for them. When having multiple versions of websites, updates need to be executed on more than one place, thus a greater possibility for errors exists.

In most cases, a better approach would be to create one website based on responsive web design principles that works correctly on all devices and resolutions that exist today, and is also ready for the future devices. Once the code is written, it can be run on any device, which is a better option than building separate websites for each device. "A huge trend in 2013 will be that most e-commerce retailers will be using responsive web design." [14]

IV. RESPONSIVE WEB DESIGN – KEY FEATURES

The term "responsive web design" was first used and explained by web designer Ethan Marcotte in 2010 [15]. "Responsive web design is the approach that suggests that design and development should respond to the user's behavior and environment based on screen size, platform and orientation." [16]

Responsive web design implies a different way of thinking, i.e. a slight change in web design philosophy. A responsive web page has one URL, one HTML code and its content is shown according to the defined CSS3 media queries on multiple devices (desktop PCs, notebooks, smartphones, tablets and televisions). It will automatically scale and adjust content to various screen sizes. The aim is to achieve readability and navigation on any device with minimum of resizing and scrolling.

Key features of responsive web design are:

- a flexible (fluid) grid,
- flexible images and
- CSS3 media queries [17].

Flexible grid consists of columns expressed in relative widths (e.g. percentages, ems), as proportions of their containing element, rather than in fixed, inflexible pixels. The grid is resizing as the viewport (the area available for viewing the web page) is changing.

Flexible images move and scale proportionally (shrinking or enlarging) as their flexible container resizes. Besides the approach of resizing images according to the screen size, there is also an option to create multiple

versions of an image for different resolution ranges or an option to crop images.

Media queries are a module from the CSS3 specification that allow building multiple layouts using the same HTML documents. They are conditional statements that can identify not only media types (screen), but can also inspect the physical characteristics of the device and the browser that render targeted web page, e.g. the browser width, orientation. Style sheets are selectively served based on the device and browser features. Each media query has two components: a media type and the query. The query consists of the name of the feature and a corresponding value. All web browsers support CSS3 media queries. Some older web browsers lack media query support, but there are some alternative solutions that could be used for those particular browsers (JavaScript).

The most commonly used media feature is width. Figure 1 shows an example of a simple media query. Using max-width and min-width for checking resolution ranges below or above the certain breakpoints, enables conditional usage of parts of CSS designed for those ranges. Breakpoints are moments, i.e. certain pixel width ranges on which one responsive size range with corresponding layout changes to another [18]. Usually, browser's viewport dimension is taken into account, rather than device's screen size.

As the web browser width is changing, style, layout and proportions of the website content change as well so users can enjoy a seamless experience on any device [19]. Figure 2 shows a responsive website on several devices.

A. Mobile First and Desktop First Responsive Web Design

There are two approaches for implementation of responsive web design. In mobile first and desktop first responsive web design approaches, designing starts at reference resolutions and with use of media queries adapting to other resolutions. Main resolution breakpoints and a content that will be shown on a web page in various ranges of resolutions should be defined.

In mobile first approach, by default, a layout appropriate for smaller screens is defined and then the design is progressively enhanced as the resolution increases. On the other hand, desktop first methodology takes desktop resolution as starting point and gracefully degrades design as the resolution decreases.

Modern websites are being designed for more devices and more resolutions than ever before. Responsive web design is "the only durable, flexible and future-proof approach to building websites for today's multi-screen world" [20]. The process of responsive website creation is less linear than traditional website creation. Designing and building a website based on responsive design involves more collaboration between designers and other team members (e.g. front-end developers) than before. Unlike traditional website designs, workflow of responsive web design is a somewhat different. Main phases (planning, design, development and delivery) are interlacing and iteratively repeating.

```
@media screen and (max-width: 768px) {
    body {
     font-size: 100%;
    }

}
```

Figure 1. An example of a simple media query

V. BENEFITS OF RESPONSIVE WEB DESIGN

Benefits of the implementation of responsive web design can be divided in three categories regarding to whom they are referred: benefits for webmasters, benefits for developers and benefits for end users.

Key benefits of implementation of responsive web design are one, content focused, device-independent website, long-term money and time savings, easy maintenance, better Search Engine Optimization (SEO) managing [21], more consistent user experience and usability.

Responsive websites have a single URL (one web) serving the same HTML for all devices so maintenance and updating content is easy. Changes and content editing can be made in one place (e.g. through a single CMS), unlike updating a separate desktop and mobile site. Having one URL is very important for sharing on various social networks, especially since social networks are often accessed through mobile devices. In that way, visitors can go directly to a website regardless of the device they are using.

Google recommends the industry best practice of using responsive web design [22], which is another reason for implementation.

With increasing number of devices and platforms for web browsing, building the site based on responsive design will maintain flexibility and better user experience. Responsive site enables consistency, keeping the same look and feel throughout all devices. Having a better user experience as a competitive advantage, more customers will become loyal, improving company's market share [23].

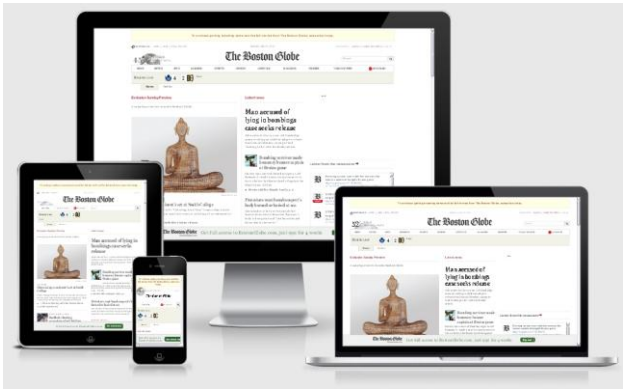


Figure 2. The Boston Globe's website loaded on several devices; an example of responsive web design

VI. Downsides And Potential Problems

Responsive web design projects are more time consuming to create and will cost about 10-20% more upfront than regular websites [24]. However, there is no need for a separate mobile site.

Some mobile devices and web browsers are not compatible with CSS media queries so alternative solutions should be taken into account for those devices. Web browsers are not uniformly supporting new web standards. Some sites have complex modules that won't function or would be hard to use on smaller devices. Responsive images are an unsolved problem of responsive web design and there have been many attempts to solve it [25]. One solution could be serving lower size images to smaller screens.

Most users expect a web page to load in less than 4 seconds [26]. Slow loading of a web page on mobile devices is the biggest performance problem for many responsive sites, mainly because of over-downloading (e.g. download and hide or shrink images, extra CSS download). Larger or unnecessary images can impact loading time. Page size and HTTP requests should be reduced for faster page load and better performance.

Guy Podjarny has conducted performance test on 347 responsive websites in March 2012 and 2013 [27]. Despite a fact that website changes its look across different screen sizes, the weight and load time of the website hardly changes. Figure 3 shows differences between page size on smallest and largest screens in the test conducted in 2013. Despite a fact that mobile version of responsive site was loaded on smallest screens, majority of the sites have roughly the same size (72%) on the smallest and on the largest screens. Comparing the results from 2012 and 2013 tests, the average page in 2012 weighted only 6% less on a small screen than on a large screen, compared to 9% in 2013 (Figure 4). The results are not good, but a slight improvement is evident.

During the designing and deployment process, the performance should be kept in mind, and website should be kept as lightweight as possible to ensure a good user experience [28].

Responsive web design is not so suitable for the

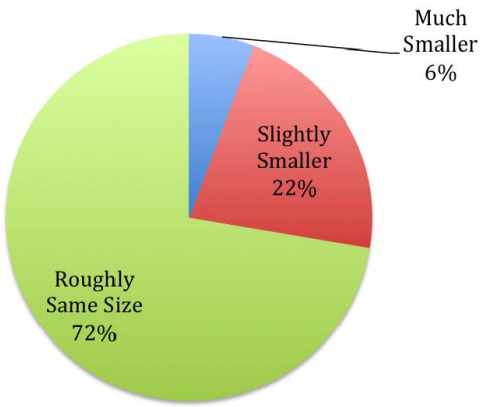


Figure 3. Page size, smallest resolution vs. largest resolution [27]

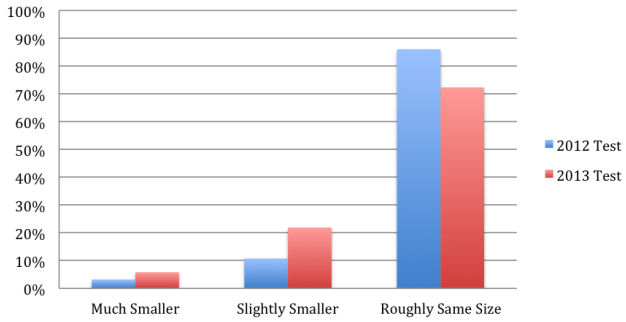


Figure 4. Page size, results from 2012 and 2013 [27]

advertisers. It is harder to place banner advertisements within a responsive design [29] than in fixed width design.

In some cases, it is better to have a separate desktop and mobile site than having one responsive site. For most organizations, implementation of responsive web design will yield improved results and long-term savings and with other benefits will outweigh any negativity [30].

VII. Conclusion

It is hard to keep up with the different devices and resolutions on the market. Responsive web design adapts the web page to different screen sizes and it is also prepared for the future-devices that haven't been released yet. Along with greater number of mobile devices, the importance of responsive web design is also increased.

Mobile devices are changing the way that commerce works. It is important for a business or a commerce website to be optimized for optimal viewing experience, reading and navigation with minimum of resizing. Implementation of responsive web design can result with greater number of visitors, increased sales and customer satisfaction.

More work should be done in defining standards and best practices of responsive web design and finding the ways for improvements (e.g. better responsive web design performance and responsive images).

There are many examples of successful implementation of responsive web design, and its implementation increases on daily basis. Responsive web design is becoming the standard of present web design. It can be concluded that responsive web design has a promising future, and will continue to develop.

References

[1] P. Cashmore, Why 2013 Is the Year of Responsive Web Design, 2012, http://mashable.com/2012/12/11/responsive-web-design

[2] J. A. Sanchez, Infographic: 2013 The Year of Responsive Design, 2012, http://www.uberflip.com/blog/infographic-2013-the-year-of-responsive-design

[3] J. Krenz-Kurowska, Web design trends for 2013, 2013, http://99designs.com/designer-blog/2013/02/21/web-design-trends-for-2013/

[4] Gartner Says Worldwide PC, Tablet and Mobile Phone Combined Shipments to Reach 2.4 Billion Units in 2013, 2013, http://www.gartner.com/newsroom/id/2408515

[5] Worldwide Smart Connected Device Market Crossed 1 Billion Shipments in 2012, Apple Pulls Near Samsung in Fourth Quarter, According to IDC, 2013, http://www.idc.com/getdoc.jsp?containerId=prUS24037713#.UWG9CZNTB8E

[6] Mobile Internet and The Death of Desktop, http://shoutoutstudio.com/mobile-internet-takeover/

[7] Infographic: Responsive Web Design, http://www.artofdeveloping.com/2012/09/infographic-responsive-web-design.html

[8] M. Bishop, The Post-PC Revolution is Here – Don't Panic!, 2013, http://blog.moovweb.com/2013/02/the-post-pc-revolution-is-here-dont-panic/

[9] IDC: More Mobile Internet Users Than Wireline Users in the U.S. by 2015, 2012, http://www.businesswire.com/news/home/20110912005213/en/IDC-Mobile-Internet-Users-Wireline-Users-U.S.

[10] The New Multi-Screen World, 2012, http://services.google.com/fh/files/misc/multiscreenworld_final.pdf

[11] M. Fisch, Mobile-friendly sites turn visitors into customers, 2012, http://googlemobileads.blogspot.com/2012/09/mobile-friendly-sites-turn-visitors.html

[12] Industry Statistics, http://www.digby.com/mobile-statistics/

[13] Global mobile statistics 2012 Home: all the latest stats on mobile Web, apps, marketing, advertising, subscribers, and trends, http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats

[14] M. Chowdhary, Responsive Web Design Will Be Expected On All Devices, 2012, http://multichannelmerchant.com/crosschannel/responsive-web-design-will-be-expected-on-all-devices-15092012/

[15] E. Marcotte, Responsive Web Design, 2010, http://alistapart.com/article/responsive-web-design

[16] K. Knight, Responsive Web Design: What It Is and How To Use It, 2011, http://coding.smashingmagazine.com/2011/01/12/guidelines-for-responsive-web-design

[17] E. Marcotte, Responsive Web Design, A Book Apart, New York, 2011

[18] The Savvy Marketer's Guide to Responsive Web Design, 2012, http://www.enablingideas.com/wp-content/uploads/SavvyMarketersGuide_ResponsiveWebDesign_Aug-7.12.pdf

[19] S. Pastore, "The Role of Open Web Standards for Website Development Adhering to the One Web Vision", International Journal of Engineering and Technology, vol. 2, no. 11, 2012, pp. 1824-1834.

[20] E. Savitz, The Future Of The Web: The Case For Responsive Design, 2012, http://www.forbes.com/sites/ericsavitz/2012/12/27/the-future-of-the-web-the-case-for-responsive-design/

[21] J. Taylor, 3 Reasons Why Responsive Web Design is the Best Option For Your Mobile SEO Strategy, 2013, http://searchenginewatch.com/article/2253965/3-Reasons-Why-Responsive-Web-Design-is-the-Best-Option-For-Your-Mobile-SEO-Strategy

[22] Building Mobile-Optimized Websites, 2012, https://developers.google.com/webmasters/smartphone-sites/

[23] S. Gamble, How Responsive Web Design Improves Customer Loyalty, 2013, http://www.sweettoothrewards.com/blog/2013/04/16/how-responsive-web-design-improves-customer-loyalty/

[24] J. Mazzei, "Responsive Web Design", Business NH magazine, 10/2012, pp.22-23.

[25] T. Kadec, Implementing Responsive design, New riders, Berkley, 2013

[26] 2012 Mobile User Survey, 2012, http://www.keynote.com/docs/reports/Keynote-2012-Mobile-User-Survey.pdf

[27] G. Podjarny, Real World RWD Performance – Take 2, 2013, http://www.guypo.com/uncategorized/real-world-rwd-performance-take-2/

[28] J. Wisniewski, "Responsive Design", Online searcher, jan/feb 2013, pp.74-76.

[29] D. Olson, Choosing between responsive web design and a separate mobile site to improve mobile visitors' experience, 2013, http://www.foraker.com/choosing-between-responsive-web-design-and-a-separate-mobile-site-to-improve-mobile-visitors%E2%80%99-experience/

[30] M. Crist, A. Francesconi, A. Paradise, and G. White, 2012, Responsive Design Guide, http://cantina.co/wp-content/uploads/2012/01/ResponsiveDesignGuideFnll0110121.pdf

GLGPS

International Conference on

GOVERNMENT, LOCAL GOVERNMENT, PUBLIC SERVICES

Steering Committee

Chairs: Mladen Mauher, Polytechnic of Zagreb, Croatia
Vanja Smokvina, City of Rijeka, Croatia

PAPERS

Problemi prijenosa podataka iz prostornih planova u geoinformacijski sustav prostornog uređenja u županiji

Adam Butigan*, Mile Pavlić**
*Zavod za prostorno uređenje Primorsko-goranske županije, Rijeka, Croatia
**University of Rijeka/Department of Informatics, Rijeka, Croatia
adam.butigan@pgz.hr , mile.pavlic@ris.hr

Sažetak - Izgradnja Informacijskog sustava prostornog uređenja RH je nužan je uvjet gospodarskog razvoja Republike Hrvatske. Razvoj ovog informacijskog sustava formalno je definiran Zakonom o prostornom uređenju i gradnji, a u suglasju je s europskom direktivom o korištenju i razmjeni digitalnih prostornih podataka, INSPIRE Direktivom.

U procesu izgradnje informacijskog sustava prostornog uređenja potrebno kreirati bazu prostornih podataka generiranu iz podataka sadržanih u prostornim planovima uređenja i u provedbenim urbanističkim planovima. To je ogroman organizacijski i tehnološki zahvat. Nesporno je da kvaliteta i brzina izgradnje Informacijskog sustava ovisi o kvaliteti izrade prostorno planske dokumentacije i njenom transferu u informacijski sustav.

Na temelju dosadašnjih iskustava na razvoju geografskih informacijskih sustava koji su se bazirali na podacima iz prostornih planova uređenja opisan je jedan od temeljnih problema izgradnje informacijskog sustava prostornog uređenja. U radu su opisani izvori pogrešaka u prijenosu podataka iz CAD sustava (u kojima se uobičajeno kreiraju prostorni planovi) u GIS sustave (u kojima se ti isti podaci mogu prostorno analizirati i integrirati s drugim informacijskim sustavima) koji bitno definiraju vrijeme izgradnje, kao i sadržaj samog informacijskog sustava prostornog uređenja.

Uvod

Hrvatska, ulazi u red razvijenijih društava u svijetu. Tome u prilog govori činjenica da se postupci prostornog planiranja provode desetljećima, postoji prostorno planerska infrastruktura, a na temelju prostornih i urbanističkih planova donose akti za gradnju i kontrolira se provedba planova.

Organizacija izrade prostornih i urbanističkih planova, odgovornosti i relacije između različitih razina planiranja prikazana je u tablici br. 1.

Nositelj izrade dokumenta/plana	Vrsta dokumenta/plana	Razina dokumenta/plana
Država	Strategija prostornog razvoja Republike Hrvatske Program prostornog uređenja Republike Hrvatske	Strateški dokumenti
Županija (regionalna samouprava)	Prostorni plan uređenja županije Prostorni plan područja posebnih obilježja	Razvojni planovi
Općina ili Grad (lokalna samouprava)	Prostorni plan uređenja Općine ili Grada	Razvojno provedbeni planovi Općine ili Grada
	Urbanistički plan uređenja Detaljni plan uređenja	Provedbeni planovi

Tablica br. 1

Kako je vidljivo iz tablice br. 1, izrada i usvajanje prostornih i urbanističkih planova u odgovornosti su regionalne i lokalne uprave, odnosno na županijama, općinama i gradovima. Riječ je o velikom broju planova.

Na dan 1. prosinac 2012. g. u Primorsko-goranskoj županiji bilo je važećih 540 prostornih i urbanističkih planova.

Podaci kreirani u planovima trebali bi se moći koristiti nebrojeno puta i za različite namjene. Evidentno je da postoji potreba za izgradnjom informacijskog sustava kojim će se omogućiti prikupljanje, obrada, sistematizacija i distribucija podataka nastalih izradom prostornih i urbanističkih planova. Izgradnja informacijskog sustava prostornog uređenja ovisna o jedinicama lokalne i regionalne samouprave.

U razvoju informacijskog sustava bitno je osigurati lakoću korištenja, mogućnost integracije odnosno osigurati interoperabilnost podataka nastalih izradom prostornih i urbanističkih planova, što je u skladu s Europskom INSPIRE Direktivom (Official Journal of the European Union, 2007. g.) [1], kojom se definira interoperabilnost prostornih podataka "kao mogućnost objedinjavanja, preklapanja i korištenja prostornih podataka, interakciju usluga, bez intervencija od strane korisnika tako da je rezultat objedinjenih prostornih podataka koherentan, skupovi podataka imaju dodanu vrijednost, uz poboljšanje kvalitete usluge".

Zbog svojeg integrirajućeg i kontrolnog karaktera u procesima izrade prostornih i urbanističkih planova, županijski zavodi za prostorno uređenje su prepoznati kao centri izgradnje informacijskog sustava prostornog uređenja, što je potvrđeno Zakonom o prostornom uređenju i gradnji iz 2008. g. (Narodne novine, 2007. g.) [2], gdje su formalno definirane formalne osnove za izgradnju informacijskog sustava prostornog uređenja u RH.

Sadašnje stanje

U Županiji se s razvojem geoinformacijskog sustava kao svojevrsnim prethodnikom informacijskog sustava prostornog uređenja započelo još 1995. godine. Na temelju zaključka županijske Skupštine je uspostavljen županijski geografsko informacijski sustav (GIS) za potrebe izrade i usvajanja Prostornog plana uređenja Primorsko-goranske županije, te je uspostavljena temeljna baza podataka o prostoru županije (Službene novine Primorsko-goranske županije, 1995. g.) [3]. Županija je svojim formalnim odlukama podržala zajedničku suradnju s nadležnim ministarstvom na objedinjavanju raspoloživih podataka nastalih pri izradi prostornih planova županija, odnosno kako će se u budućnosti pokazati, suradnju na razvoju informacijskog sustava prostornog uređenja RH (Poglavarstvo Primorsko-goranske županije, 2002. g.) [4].

Na osnovu navedenih formalnih odluka ustrojila se baza podataka geoinformacijskog sustava koja se zasniva na podacima koji su nastali tijekom izrade županijskog prostornog plana.

Podaci u bazi su podijeljeni u četiri logičke cjeline, sukladno strukturi samog županijskog prostornog plana: 1) Opći podaci o prostoru; 2) Infrastrukturni sustavi; 3) Uvjeti korištenja, uređenja i zaštite prostora; 4) Podaci iz tekstualnog dijela prostornog plana. Baza je ustrojena na podacima čiji se raspon kreće od županije kao cjeline, do pojedine općine ili grada. Sukladno tome mjerilo prikupljanja podataka se kreće od mjerila 1:25000 do 1:100000, a atributni podaci su određeni entitetom mjerila. Vođenje, ažuriranje i održavanje baze podataka vodi se u Zavodu.

Na osnovu iskustava stečenih pri izradi i korištenju ovog geoinformacijskog sustava utvrđeno je da su podaci iz ove baze dovoljni za izvođenja prostornih analiza na regionalnoj razini, ali je za potrebe izgradnje informacijskog sustava prostornog uređenja potrebno kreirati bazu podataka koja će osim podataka iz županijskog prostornog plana sadržavati podatke iz svih važećih prostornih i urbanističkih planova, odnosno podatke sadržane u prostornim planovima općina i gradova, te podatke iz urbanističkih i detaljnih planova uređenja.

Kod povećanja izgradnje informacijskog sustava koji se sadržavati baze podataka nastale iz podataka prostornih planova nižeg reda, dakle prostornih planova općina i gradova potrebno je organizirati dvije grupe aktivnosti:

1. Izrada grafičkog dijela prostornih planova uređenja općina ili gradova radi se u krupnijim mjerilima, odnosno na mjerilima 1:25000, te dio kartografskih prikaza u mjerilu 1:5000. Za usporedbu županijski prostorni plan uređenja se radi na mjerilu 1:100000. Sadržaj podataka između mjerila 1:25000 i 1:100000 je 16:1, a između mjerila 1:100000 i 1:5000 400:1. Iz navedenih odnosa vidljivo je povećanje sadržaja baze podataka planiranog informacijskog sustava. Za potrebe izgradnje informacijskog sustava potrebno je utvrditi model kojim će se podaci iz važećih prostornih planova uređenja općina ili gradova automatizirano unositi u informacijski sustav.
2. Odgovornost izrade i usvajanja prostornih planova uređenja općina ili gradova u nadležnosti je općina i gradova, dok su izrađivači navedenih planova tvrtke s ovlaštenjima za izradu prostornih planova. U županiji je ustanovljeno 36 jedinica lokalne samouprave, a sve su općine i gradovi jednom ili više puta usvajali prostorne planove uređenja za svoje područje nadležnosti. Za potrebe izgradnje informacijskog sustava potrebno je utvrditi model kojim će se omogućiti izgradnja sustava za praćenje izrade i usvajanja prostornih planova, te za nove usvojene ili mjenjane planove provesti postupak iz točke 1.

Daljnje proširenje baze podataka na razinu provedbenih urbanističkih planova, urbanističkih i detaljnih planova uređenja, proporcionalno povećava problem izgradnje informacijskog sustava, kako zbog dodatnog broja planova tako i dodatnog sadržaja podataka u planovima, te povećanja kompleksnosti praćenja usvajanja novih planova ili stvaljanja van snage postojećih planova. U Županiji je na dan 1. prosinac 2012. g. važećih, usvojenih 178 urbanistička plana, te 257 detaljna plana uređenja. Podaci koji se trebaju obraditi u ovim planovima prikazani su u najkrupnijim mjerilima (mjerilo 1:500 do 1:5000) (Izvor „Registar prostornih

planova“ Zavod za prostorno uređenje Primorsko-goranske županije, 2012. g.) [5].

Izrađivači prostornih planova su se u protekla dva desetljeća adaptirala na informatičku tehnologiju. Za grafički (kartografski) dio prostornih planova uobičajeno koriste CAD programske alate, a za izradu tekstualnog dijela planova MS Office programske alate. Tijekom razdoblja adaptacije nastajali su planovi koji su bili različiti po formatu, obliku i sadržaju.

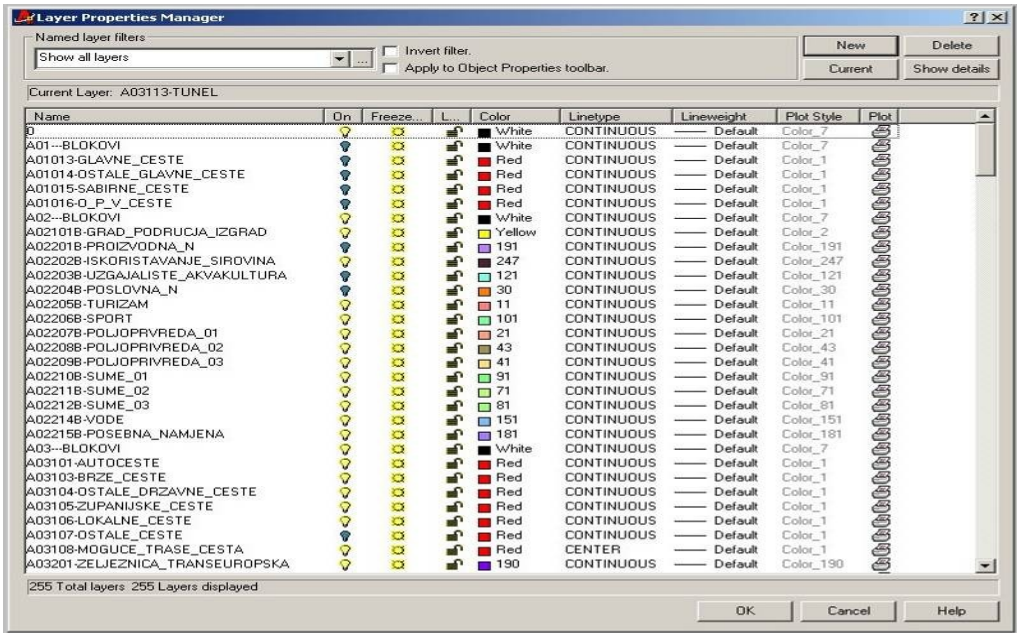
S ciljem ujednačenja sadržaja, oblika i formata prostornih planova, nadležno je Ministarstvo donijelo „Pravilnik o sadržaju, mjerilima kartografskih prikaza, obveznim prostornim pokazateljima i standaru elaborata prostornih planova“, (Narodne novine, 1998. g.) [6]. Pravilnik je definirao izgled obaveznih grafičkih priloga prostornog plana, te na taj način dijelom standardizirao izradu grafičkog dijela prostornih planova pomoću CAD programskih alata.

Međutim u Republici Hrvatskoj Zakonom o prostornom uređenju i gradnji je i dalje definirano da je samo onaj plan koji je iscrtan na papiru i potpisan od strane odgovorne osobe službeni prostorni plan.

Situacija sa stanjem izrade prostorno planskih dokumenata u drugim državama Europe je slična situaciji u Hrvatskoj. Praktično sve europske države i dalje drže da je samo onaj plan koji je iscrtan na papiru i potpisan od strane službene osobe jedini službeni prostorni plan. Jedini izuzetak od ovog pravila je Nizozemska, gdje po zakonu o prostornom planiranju iz 2008. g., svi novi prostorni planovi moraju biti digitalni i kao takvi imaju službenu valjanost u poslovima izdavanja akata za građenje. (M. Schrenk, J. Neuschmid, W. Wasserburger, Plan4allProjekt, 2011. g.) [7].

Sadašnje stanje izrade prostornih planova je svojevrsna prijelazna faza iz analognog sustava prostornog planiranja u digitalni informacijski sustav prostornog planiranja.

Za potpuni prelazak u digitalni informacijski sustav prostornog planiranja potrebno je utvrditi nedostatke i slabosti izrade postojećeg načina izrade prostornih planova pri izradi grafičkog dijela prostornih planova pomoću CAD programskih alata, otkriti probleme u prijenosu podataka iz prostornih planova u geoinformacijski sustav prostornog uređenja, a nakon toga utvrditi metodiku kojom će se utvrđeni problemi u prijenosu ukloniti.



Slika br. 1

I. Slabosti i nedostaci kod izrade prostornih planova

Izrada grafičkog dijela prostornog plana u CAD programskim alatima provodi se u jednoj CAD datoteci.

Podaci koji su podloga za izradu grafičkog dijela prostornog plana importiraju se u CAD datoteku bilo fizički ili virtualno kao referentna datoteka. Sam proces izrade grafičkog dijela plana provodi se crtanjem crteža, a crtanje dijelova plana u crtežu je razdijeljeno na slojeve.

CAD (Computer Aided Design/ Računalom podržano crtanje) se definira kao korištenje računala za stvaranje i uređivanje crteža. AutoCAD je najzastupljeniji CAD programski alat u izradi grafičkog dijela prostornih planiva. AutoCAD je alat koji je od početka bio razvijan kao programska aplikacija koje radi s datotekama, a AutoCAD crteži su uobičajno organizirani unutar jedne datoteke u slojevima. Entiteti crteža preuzimaju svoje atribute (boja, tip linije, tip objekta) iz sloja na kojem su

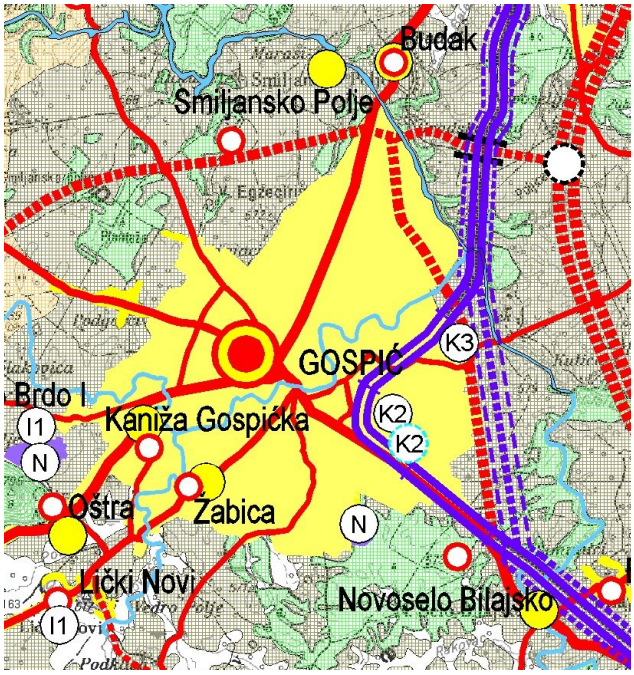
kreirani. Ovo je jednostavan i učinkovit način za organiziranje podataka, ali je nužna pažljiva kontrola kvalitete kako bi se osigurala dosljednost u izradi crteža.

Zbog toga što razvoj CAD programskih alata treba zadovoljiti potrebe inžinjerskog svijeta gdje su ponekad dozvoljene pogreške milimetarske vrijednosti, u CAD programskim alatima je puna pažnja posvećena upravljanju podatcima bez gubitka preciznosti. CAD alate karakterizira omgućavanje crtanja ili projektiranja realnih objekata precizne geometrije, te mjenjanje ili uređivanje geometrije bez gubitka preciznosti u stvarnom prostoru (Removing Obstacles Between CAD and GIS Professionals, S Curry, Autodesk Strategic White Paper, 2004. g.) [8].

Tako npr. u jedan sloj se mogu smjestiti rubovi građevinskih područja, u drugi sloj koridori vodovodne mreže, a u trećem sloju mogu biti prikazani objekti na vodovodnoj mreži. Organizacija slojeva nije propisana niti definirana, te je moguće da je za svaki prostorni plan različita. Na slici br. 1 (na prethodnoj stranici) prikazana je lista CAD slojeva iz jednog od prostornih planova.

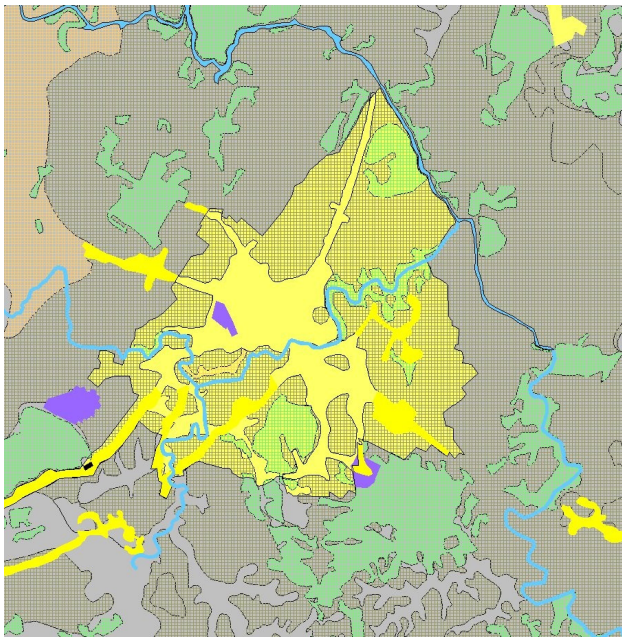
Drugi nedostatak izrade prostornih planova CAD programskim alatima je nemogućnost uspostavljanja relacije između slojeva. Ovaj nedostatak rezultira nekoherentnim podacima, povećanjem broja slojeva u crtežu, te mogućem crtanju istog objekta u više slojeva

Kao primjer na slici br. 2 prikazana namjena površina, građevinsko područje naselja, infrastrukturni koridori i uvjeti korištenja kao pozadinski sloj.



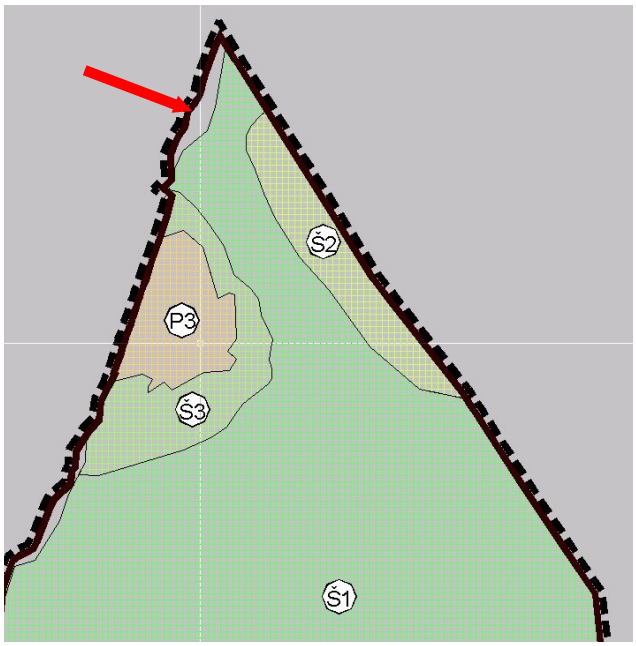
Slika br. 2

Na slici br. 3 za isto područje, namjena površina je preklopljena s uvjetima izgradnje, ali su uvjeti izgradnje naglašeni i prikazani preko namjene površina, a sloj kojim se prikazuje infrastruktura je isključen. Uočava se da različit način prikazivanja dvaju tema rezultira različitim čitanjem istog prostora. Budući da postoji nekoherentnost između dvaju sloja nije moguće utvrditi one površine koje imaju istu namjenu površina, a različite uvjete gradnje. Odnosno nije moguće dati odgovor na pitanja u kojemu su postavljana dva ili više uvjeta.



Slika br. 3

Treći nedostatak izrade prostornih planova CAD programskim alatima nastaje također zbog nekoherentnosti između grafičkih elemenata prostornog plana. Kod crtanja grafičkih elemenata događaju se tehničke greške koje nije moguće provjeriti. Ove se greške, kada su male veličine, ne vide se kod iscrtavanja plana. Na slici br. 4 strelicom je naznačen je primjer takve pogreške, gdje za označenu površinu nije definirana namjena.



Slika br. 4

Isti tip pogreške je prikazan na slici br. 5, na kojoj je uvećan sloj granica namjene zemljišta. Kako je vidljivo na slici, linije susjednih površina nisu identične i na tom mjestu generiraju pogreške

Slika br. 5

Iz istog razloga, događa se razlika između vrijednosti očitanih iz grafičkog dijela prostornih planova i tekstualnog dijela plana.

Pvršine u koje su nacrtane u planovima kao što su građevinska područja trebale bi biti identične u iskazu površina u tekstualnom dijelu plana. Na slici br. 6 prikazane su razlike između navedenih površina u jednom od prostornih planova.

Na temelju gore navedenih primjera može se zaključiti da postojeći način izrade prostornih planova pati od slijedećih nedostataka:

1. Podaci u prostornim planovima su međusobno nepovezani
2. Prostorni planovi organizirani su kao zasebni crteži
3. Prostorni planovi imaju različit sadržaj i organizaciju podataka
4. Model podataka i sadržaj podataka nije dokumentiran.
5. Različiti su formati podataka
6. Ne postoji uspostavljen sustav kontrole kvalitete izrade prostornih planova
7. Ne može se uspostaviti jedinstven pristup sustavu, u svrhu održavanja podataka, pretrage i analize, te distribucije

Naziv grada GRAD		Oznaka	Ukupno iz tekstualnog dijela (ha)	Ukupno iz GIS baze (ha)
ISKAZ PROSTORNIH POKAZATELJA ZA NAMJENU POVRŠINA				
Građevinska područja naselja izgrađeni dio GP naselja	ukupno ukupno	GP	**470,01** 281,92	**459,62** 271,34
Izgrađene strukture građevinskog područja izvan naselja		K R IS	**27,45** **4,08** **16,76**	**27,46** **4,08** **16,74**
Poljoprivredne površine -obradive -vrtače i manje obradive površine	ukupno	P $P3_1$ $P3_2$	**85,97** 59,19 26,78	**87,07** 63,69 23,38
Šumske površine - gospodarske - posebne namjene (šume u zaštićenim dijelovima prirode-prijedlog zaštite)	ukupno	Š Š1 Š3	**84,32** 13,02 71,30	**84,01** 13,02 70,99
Ostale poljoprivredne i šumske površine	ukupno	PŠ	**445.88**	**455.42**
Vodne površine - vodotoci - jezera - akumulacije - retencije - ribnjaci	ukupno	V	NEMA	NEMA
Ostale površine postojeće groblje planirano groblje	ukupno	G1 G2	**5,53** 1,33 4,20	**5,52** 1,33 4,19
Grad ukupno			**1140**	**1140**

Slika br. 6

II. Zaključak

Prostorni planovi su nosioci podataka i informacija za potrebe izgradnje informacijskog sustava prostornog uređenja, te kao takvi trebaju biti oblikovani na način da budu upotrebljivi za izgradnju informacijskog sustava prostornog uređenja.

Analizom usvojenih prostornih planova utvrđeno je da izrada grafičkog dijela prostornih planova namijenjena izradi kartografskih priloga na papiru, a izrađivači prostornih planova proizvode planove različitih formata, različite strukture.

U radu su prezentirani temeljni nedostaci prostornih podataka iz grafičkog dijela usvojenih prostornih planova nastalih korištenjem CAD programskih alata, kada se isti prostorni podaci koriste kao podaci za stvaranje i ažuriranje baze podataka postojećeg geoinformacijskog sustava u županiji. Utvrđeno je da se podaci koji se sada kreiraju tijekom izrade prostornih planova bez dodatne informatičke obrade ne mogu koristiti za automatiziran unos informacijski sustav prostornog uređenja.

U postupku izgradnje informacijskog sustava prostornog uređenja nužno je definirati sadržaj, oblik i format prostornih i urbanističkih planova sukladno strukturi baze podataka informacijskog sustava prostornog uređenja, te definirati metodiku kojom će se omogućiti da sadržaj prostornih i urbanističkih planova automatizirano puni baze podataka informacijskog sustava prostornog uređenja. Autori ovog rada upravo rade na jednoj takvoj metodici.

Literatura

[1] European Commission: Directive 2007/2/EC of The European Parliament and of The Council: Establishing an infrastructure for spatial information in the Community (INSPIRE), Brussels, Official Journal of the European Union, 2007. g.

[2] Zakon o prostornom uređenju i gradnji, Narodne novine, 2007. g.

[3] "Informacije o izradi Izvješća o stanju u prostoru, programu mjera i Prostornog plana Županije primorsko-goranske", Službene novine Primorsko-goranske županije, 1995. g.

[4] "Informacija o upravljanju prostorom u regionalnoj i lokalnoj samoupravi", Zaključak Poglavarstva Primorsko-goranske županije, 2002. g.

[5] "Registar prostornih planova" Zavod za prostorno uređenje Primorsko-goranske županije, 2012. g.

[6] "Pravilnik o sadržaju, mjerilima kartografskih prikaza, obveznim prostornim pokazateljima i standardu elaborata prostornih planova" Narodne novine, 1998. god.

[7] Planning system in Europe and SDI, Manfred Schrenk, Julia Neuschmid, Wolfgang Wasserburger, Plan4allProjekt, 2011. g.

[8] Removing Obstacles Between CAD and GIS Professionals, S Curry, Autodesk Strategic White Paper, 2004. g.

Measuring success of implementation County development strategy by using Balanced Scorecard Method

Robertina Zdjelar
Koprivnica-Krizevci County
Nemciceva 5, 48000 Koprivnica, Croatia
robertina.zdjelar@kckzz.hr

Abstract - County development strategy is the main strategic document of a county (regional government) in Croatia regulated by the Law on Croatian Regional Development [10]. It is the framework for Activity plan of implementation County development strategy and planning projects and activities, financial plans during a period of three or in the future of seven years. In this article author will present the procedure from phase of collecting propositions and preparing projects and activities, creating a documentation of them, creating a pool of projects and activities, deciding about criteria for acceptance by creating measurement system (strategic objective, priority, measure, action, performance driver, performance indicator), acceptance of propositions, implementation and recording relevant information about implementation activities or projects, evaluation of them. The evaluation of achieved results gives knowledge necessary for next cycle of planning and decision making. As an example of measuring efficiency of project implementation by Balanced Scorecard Method will be presented project of implementation e-Government system. The assumptions for implementing procedure mentioned above and Balanced Scorecard Method in Croatian regional government also will be shown in this article.

I. INTRODUCTION

In last fifteen years many changes were happened in context of strategic planning in public organizations. The beginnings of strategic planning in public organizations are described by J.M. Bryson [1]. It is some kind of evolution of public sector and it happens slowly but continually. The first generation of counties strategic document was Regional Operative Programme accepted to period of three years. The next generation of counties strategic document is County development strategy which also covers period of three years until 2013. and seven years in the nearest future.

Defining County development strategy is obligated by law. The structure of it is also defined by law. So, the County development strategy contains strategic goals, SWOT analysis, consideration of available resources, first of all natural resources, time, financial and human resources for translating the strategy into the action. This is obligation of County assembly, County Prefect, county offices, stakeholders and publicity. Translating strategy into action is obligation of County Prefect, heads of county offices who have to manage all processes, resources planning, to do activity plan, to analyze problems, prepare planning documentation for project that will be the best for achieving strategic goals.

From the other side, the process of budget planning for the period of three years (in next cycle it will be seven years), which is also regulated by law, defines what is necessary for precise and objective budget planning. The Budget is a financial tool for implementing strategy. Strategic goals could be successfully achieved only if activities and projects are precisely planned, implemented and checked. Quality of project planning documentation is relevant for successful implementation and control. These documents are used by operative officers in their everyday job who take actions oriented toward strategic goals defined in County development strategy. Reporting about implementing County development strategy and measuring success doesn't need to be additional job with a lot of papers; it is just presentation of results of everyone's everyday job. An interesting, simple and very practical thought was given by P. Drucker, who said: "Everyone need to learn to ask two questions: What information do I need to do my job – from whom, when, how? and: What information do I owe others so they can do their job, in what form, and when?" [2].

Measuring success of implementing strategy, in the context of spending public money, is the top theme of all governments and publicity all over the World. It is a matter of social responsibility at all instances of public governance. The European Commission has articulated four challenges for The Republic of Croatia in the process of preparation for membership in EU and some of them are inefficient public governance at central/local level and weak involvement of partners[1].

These two reasons, between others, are the most important for taking more care to organizing,

[1] European Commission - Position of the Commission Services on the development of the Partnership Agreement and Programs in Croatia for the period 2014-2020, Launch Event in Zagreb, 31 January 2013, http://www.mrrfeu.hr/default.aspx?id=1315

analyzing and reorganizing business processes in public governance at all levels. They are also objective reasons for articulating necessity of measuring success of implementation of county development strategy.

The example of applying Balanced Scorecard Method (BSC) for translating strategy into action will be presented in this article. It will be, also, presented which assumptions have to be fulfilled for implementing BSC in County business system for measuring strategic achievements. There will be mentioned the gaps between what such kind of business system has to but it hasn't fulfilled for implementing Balanced Scorecard Method.

II. IMPLEMENTATION OF THE COUNTY DEVELOPEMENT STRATEGY

County strategy contains mission, vision, strategic goals, measures which means the way some goal could be achieved, metrics (performance drivers) for measuring and objective performance indicators of goal achievement. Implementation of the County development strategy is continual process which includes County as initiator of strategy, counties regional agency as a coordinator, other stakeholders and because of that it is very complex issue.

Managing at top level of county offices in the process of implementation county strategy is obligation of County Prefect and his/her deputies. These meetings result with general reference for implementing strategy. They should be undertaken at least once a month or more often. This is a mechanism of continuous process improvement and a kind of control. It provides delegation of rights and obligations to head of county offices.

The next level is middle management. Translating strategy into action is the main task of heads of county offices. Heads of county offices should have project management skills for planning projects, assessing risks, planning and scheduling resources, implementing projects, take control over all phases of project management and disseminate results of projects. The appropriate tools and techniques are recommended for all of these phases. For effective planning and implementation of projects each head of office has to be familiar with them if he wants to do his job well. Special topic for research is a question – How many heads of office in Croatian counties have project management skills in terms of Project Management Institute? This is one of more important assumption for better planning, implementing, control of every project and disseminating the results of project. So the precondition is to educate heads of office and other lead officers for project management, techniques and tools for managing project. It should be applied to all kind of project – no matter of financial sources (EU contribution, state budget, county budget or other sources).

In the accordance with top management head of office communicate the strategic decisions translated to project and project goals to operative staff. Continuous top down communication and dissemination of important information is the most important for project success. If it doesn't work there is a problem at operational level.

Each project that is defined at middle level consists of activities. Each of them is oriented to results which have to be achieved at the time and as it is planned otherwise there is a risk of project failure. The operative officers should be competent to respond to all tasks which are delegated to them from head of office [5]. Competence of officers is a special subject for research, especially human resource planning, scheduling specialists between projects and development of human resources. It is also an important part of County strategy. If county has no competent staff for some task it is a potential risk for project implementation. The engagement of outsourced staff makes additional expenses that have to be planned. There is also a risk of dependency about outsourced human resource.

III. ASSUMPTIONS FOR APPLYING THE BALANCED SCORECARD METHOD

The Balanced Scorecard method is theoretically described and explained in many editions by its founders Norton and Kaplan [3],[4],[6]. It is renowned by many business systems which have used it for measuring business success [4]. In this article I will not describe the BSC method in details but I will concentrate on actual situation in Croatian public management and its ability for implementing BSC in Croatian regional government practice. The authors of [7] and [8] were also considered BSC in Croatian government institutions.

The main concept of BSC is based on four perspectives: finance, relationship with stakeholders, internal processes and learning, growth and development. Today's law basis in The Republic of Croatia which has to be implemented in Croatian counties regulates many important issues that could be related with the concept of BSC. They will be listed hereafter.

In the field of **finance perspective** head of county finance and budget office has to manage and control all program documents that consists of projects and activities included in The Budget of County. Every program, project and activity which is accepted and included in The Budget has to be planned in accordance with the Regulation of budgetary classifications [11] and Annual guidelines of budget planning [12]. It determines which attributes for

program, project and activity has to be defined, monitored, controlled and reported. The main attributes specified in program document are connected with strategy goal and indicators which give the answer if the goal is or isn't achieved, the financial amount and sources of financial resources. Project has more details that have to be specified [12].

The **stakeholder perspective** is the complex part of measuring success because of many reasons. For example there are many stakeholders who have different influence to planning projects. County government has to balance financing many different social needs (economy, promotion of the region, health, education, social program, cultural program). Stakeholders are directly known from the main strategy goals so they have to be included in implementation of project directly.

A stakeholder's satisfaction with fulfilment of their needs is checked on elections. The publicity has opportunity to say is it or is it not satisfied with implemented programs, projects or activities and does it suppose that public money is invested well. This perspective is directly connected with County mission and Budget.

The **internal processes** in county are more and more in focus of observing the functionality and efficiency of the business system but there is still a gap between what is really necessary to analyze and what it is done because of law regulation. The emphasis is on formalization of processes so recent activities are oriented to defining procedures without taking care enough about efficiency and effectiveness of business process. It will be the next phase of implementing control system or it could be said that it will be the next level of maturity of public government business system.

Learning, growth and development is the component of BSC which is recognized and also defined by law in Croatian public government. The strategy goals directly determine needs for capacities for example improvement of human resources skills or ICT resources. The law regulation determines that County has to define annual plan for education and improvement human resource skills which are recognized as necessary in accordance with programs, projects and activities which are planned in determined period. The department of human resource development is organized in almost all Croatian counties and solving the human resource issues has to be the main mission of it. The emphasis shouldn't be only on the planning development of human resources for the skills that are necessary for the system; for example educating risk manager, internal auditor, information security manager. The focus has to be oriented to developing capacities for making possibilities for implementing more projects oriented to stakeholders and for achieving strategic goals.

Developing all mentioned elements of BSC is continual process which is related with the law regulation but it doesn't depend only on law, it depends on vision of county top management and ability to predict need for improvement in area of county's weaknesses. The strength, opportunities, weakness and treats are defined in county's development strategy. More attention was oriented to recognizing and defining strength, opportunities, weakness and treats of county's area and all these important issues were not considered in details and were just mentioned in context of county business system. Considering available capacities, finding possibilities and choose the best strategy to fulfil preconditions for providing services (projects and activities) oriented to goals of county development strategy are the duties of county's top management.

IV. CONTROLING, MONITORING AND MEASURING OF SUCCESS

The model of public government functioning is presented at Figure 1. The authors of BSC method suggested five principles of a strategy focused organizations that are the most important for success [4]:

- translate the strategy to operational terms
- align the organization to strategy
- make strategy everyone's everyday job
- make strategy a continual process
- mobilize change through executive leadership.

If we analyse the model presented at Figure 1 in accordance with these principles we can notice that translating strategy to operational terms is happening in the process of budget planning – in the part of "PLANNING". There is the procedure for proposing project ideas which are connected with strategy objectives and included in County development strategy.

County as a business system, as an organization, consists of few offices (departments). Each of them has its own scope, role, knowledge base, internal culture. Head of office is a member of middle management. This level of staff has the most important role because from the one side they directly participate in the process of strategy planning and on the other side they participate in the process of strategy implementing. They are responsible for collaboration and communication between different levels of processes (up to top management and down to operative staff). Any change in the organization has to be initiated and mobilized by head of office and approved by top management.

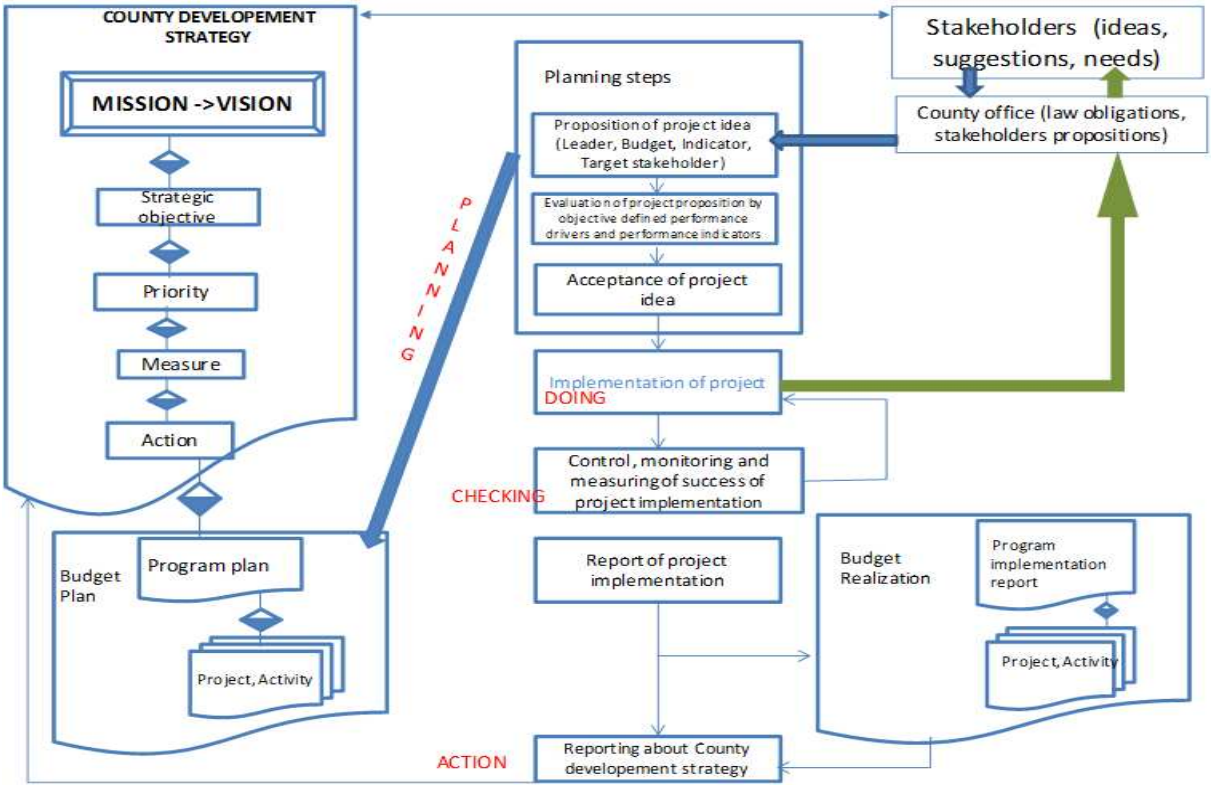


Figure 1 Model of public governance functioning

Strategy has to become everyone's everyday job. It could be seen at Figure 1. in the part of "DOING". It means that it has to be achieved through projects at operational level. Many possible risks could obstruct the implementation of project but they have to be considered through risk assessment in the project planning. It is the obligation of project manager. He is responsible for predicting possible treats and for finding solutions for proactive problem solving. About skills and human resource capacity was told in the chapter II.

The top management is the most interested for making strategy a continual process. If the top management checks the implementation of strategy often and organize coordination meetings about that issue then the annual reporting wouldn't require more time spending.

If the strategy is everyone's everyday job (phase of "DOING" at Figure 1.) and everyone does his job in accordance with project task schedule or project activity plan, then all important information for reporting are available. Collecting information for reporting shouldn't be additional job only after the project is finished. If it is so, then there is a gap between plan of implementation and the process of implementation of projects. Collecting information has to be a continual process during project implementation and has to be monitored by someone (project manager, head of office). The information about project implementation has to be written because of documenting of implementation. It is a part of operational processes. So who is responsible for omissions and failures? The answer is all of employees – everyone for his or hers tasks.

Measuring the success of implementation County development strategy is a part of top and middle level business processes. It should be continual and be a part of "CHECKING". Checking without measuring the indicators of each project which has to be implemented is not a correct done job. How do we know does the result is achieved if we don't know where we start, where we are and where we would like to be? The conclusions about success of implementation are not objective if there are no measured indicators. The "ACTION", that should be undertaken as a feedback to make changes in project implementation and to the next cycle of strategy making, depends about that.

The law regulation didn't emphasize the continuity in measuring success, because it is a matter of management. Measuring success is obligation of management by default. There was just some recommendation from Ministry of finance for using BSC method. The emphasize in auditing and controlling in financial function is put at the reasonable spending public money and isn't put at providing effective and efficient public services with main purpose of achieving strategic objectives. So, this is another gap between what it is done and what has to be done.

There are all elements of BSC in the counties practice. An example of using BSC for Strategy goal: Development of e-regional and local government, to establish and connect government systems which can provide an Internet services to citizens and other users of government services is presented in Table 1.

Table 1. Balanced Scorecard for Strategy goal [9] Development of e-regional and local government (1-4-1), to establish and connect government systems which can provide an Internet services to citizens and other users of government services [13]

Analyzing and planning financial resources – FINANCE PERSPECTIVE		
Action	Performance driver	Performance indicator
Increase own financial resource	Financial independence	• current incomes in Kunas (Kn) • available budget for project in Kn
Improve the efficiency of financial resource	- Well done procedure of ICT equipment procurement based on detailed specification of procurement object - Well done education of ICT and administrative staff	• number of satisfied specified request of quality of procurement object and amount of procurement in Kn
Analyzing incomes from administrative taxes	Analyze Taxes paid to State Budget and to County Budget	• incomes in Kn
Public relations – STAKEHOLDERS PERSPECTIVE		
Publishing information about implemented counties e-services	Informing citizens of County about new e- services	• Number of publication • Number of radio and TV news casting in local electronic media
Implementation of possibilities for check in which phase is the case through Internet	Informing citizens of County about new possibilities	
Increase number of e-services	Analyze of possibly to implement new e-services	• Number of services • Number of e-services
Analyzing customs complaints before and after implementation of e-services	Quality control – number and types of complaints	• Number of complaints before implementation of e-services/100 services in a period of year • Number of complaints after implementation of e-services/100 services in a period of year • Number of reasons for complaints
Analyzing cases and law regulative – INTERNAL PROCESSES PERSPECTIVE		
Analyzing number of customs request by purpose, number of their appearance and grouping by purpose in groups – content analyzing	Continuity of analyzing requests	• Total number of requests in last 3 years annually • Number of types of requests • Groups of requests and number of required templates
Analyzing number of errors in finalized documents	Continuity of Quality control	• Number of errors in finalized documents/ Total number of documents
Resources for providing services – INTERNAL PROCESSES PERSPECTIVE		
Analyzing resources (technical, human, experience)	Analyzing resources	• Number of available ICT resources • Capacities of ICT resources in accordance with possible number of request annually • Number of administrative and ICT staff
Optimize ICT and human resources	Optimization of resources, determining of human resource and specification of ICT resources	• Detailed specification of necessary resources that have to be obtained and installed • Number of administrative staff in accordance with optimized business processes
Optimizing internal processes – INTERNAL PROCESSES PERSPECTIVE		
Reorganizing internal processes from acceptance of customs request to phase of closing case with promptly responding to requests and providing documents	Reducing unnecessary steps in procedures of solving cases and measuring the time from request to closing the case	• Number of steps in procedures of solving cases • Number of days from request to closing the case
Implementation of e-services	Putting e-services in use	• Number of days for implementation • Evidence of unplanned situations
Improving resources – LEARN, GROWTH AND DEVELOPEMENT		
Improve ICT infrastructure – development	Investment in ICT equipment and software; procurement procedure	• Number of necessary equipment • Hours out of services • Guarantee period • Amount of procurement
Improve ICT staff skills - learning	Continuity of investment in ICT staff skills	• Number of educated and certificated ICT personal • Number of ICT specialists for development and maintenance of the system
Improve administrative county staff ICT skills – learning	Continuity of investment in administrative county staff ICT skills	• Number of educated and certificated administrative county staff

V. CONCLUSION

Development of society and process of Croatian integration to Europe Union bring a lot of challenges for public government in Croatia. There are a lot of issues that have impact to County development strategy in 2014. – 2020. which have to be adopted in 2013. Setting up strategy and implementation of it is actual topic for most Croatian counties because of new planning cycle, because of revision of success of implementation of actual strategy document.

There are many possible ways to solve strategic issues but the right way has to include analyzing objective indicators which were measured during actual strategic period. It is a part of "ACTION" at Figure 1. Which method for measuring success is used is very important thing. The BSC method gives more because it observes not just financial indicators; it observes success of business system as a result of effective business processes, as a result of competitive human resources. The success of government business system has to be measured by stakeholder's satisfaction. That's the mission of all levels of government.

REFERENCES

[1] J.M. Bryson, Strategic Planning for Public and Non profit Organizations, Revised edition, Jossey-Bass Publishers, San Francisco, USA 1995.

[2] F.P. Drucker, Managing the Non-Profit Organization, HarperCollins Publisher, New York, USA, 1992.

[3] R.S. Kaplan and D.P. Norton, Translating strategy into action The Balanced Scorecard, Harward College, USA, 1996.

[4] R.S. Kaplan and D.P. Norton, The strategy focused organization – How Balanced Scorecard companies thrive in the new business environment, Harvard Business School Press, Boston, Massachusetts, USA, 2001.

[5] H. Mintzberg, Structure in fives Designing effective organizations, Prentice Hall, A Simon & Schuster Company, Englewood Cliffs, New Jersey, 1993.

[6] N-G. Olve, J. Roy, M. Wetter, A practical guide to using the Balanced Scorecard, Wiley & Sons, Chichester, West Sussex, England, 2001.

[7] M. Mauher and V. Obersnel, Intelligent City Balanced Scorecards for the City of Rijeka, Mipro 2007: 15-23, Opatija, Croatia, 2007.

[8] Ž. Dobrović, M. Tomičić, N. Vrček, Towards an effective e-Government: Implementation of a Balanced Scorecard in the public sector, http://www.egov.lt/uploads/file/Dobrovic, Tomimic, Vrcek.pdf, downloaded: March 18th 2012

[9] Koprivnica-Krizevci County, County development strategy 2011.-2013., Koprivnica, 2011., www.kckzz.hr , downloaded: April 12th 2012.

[10] Law on Croatian Regional Development, Official Gazette No. 153/09

[11] Rules for budgetary classification Official Gazette No. 26/10

[12] Ministry of finance, Annual guidelines for budget planning, http://www.mfin.hr/hr/lokalni-proracuni, downloaded: July, 2012.

[13] R. Zdjelar, The Balanced Scorecard Method as Background for e-Government in Regional Government, http://www.ceciis.foi.hr/app/index.php/ceciis/index/pages/view/ProceedingsArchive2012

Effects of e-governmment in Croatia

Klasinc, Anton-Jan
* Vertex projekt d.o.o.
ajklasinc@yahoo.com; ajklasinc@gmail.com

Abstract – This paper aims to asses the transformational impact of EU integrations and e-government on the structure of government and its effectiveness and efficiency. Two examples that are used for the purpose of analysis are the Ministry of Science, Education and Sports and the government agencies that cooperate with it, and the Ministry of Finance, Tax Administration. The first is an example of the agencification model that was stimulated by European integrations in the period from 2001-2009, mainly in order to decentralize decision making and governing of the EU funds. The other example is the cash fiscalization process of Tax Administration, that was stimulated by the processes that started from 2010-2012 and are connected with the need to react to the economic crisis with the onset in 2008 that continued to the present. This process is the opposite to the first one, as it is not the process of decentralization of fund distribution, but of centralization of collection of funds, that is based on cash accounting, not accrual accounting. It can be shown that both processes were stimulated by European integrations and EU fund management, i.e. by the need to increase the collection of funds by governmnet taxing in order to meet the requirements for EU fund programmes and in order to implement European values and reduce corruption. This process is blocked by the lack of trust of citizens in government and the unequal redistribution of collected funds.

I. INTRODUCTION

The process of decentralization in Croatia started with the onset of European integrations and increased form 2001-2009, with the pro-EU governments of prime minister Mr. Račan and two such governments of the prime minister Mr. Sanader. This was done in parallel with the application of DIS – decentralized implementation system of distribution of EU funds, that is adminsitered in connection with the creation of National Fund system, in correlation with state budget. This system of decentralized fund distribution was based on accrual accounting and budgeting and was largely dependent on agencification model and e-government in order to create a rather complex government structure of agencies and fund management that was able to connect to the local government. The central agency in this respect was the CFCA – Central Funding and Contracting Agency (SAFU), previously CFCU – Central Funding and Contracting Unit within the Minstry of Finance (MoF). Its processes of tenders and contracting are largerly dependent of e-government structures, just as public procurement has become in the years 2001-2009. The problem with this model was the rather pervasive culture of corruption that led to the blocking of EU fund administration by the EU in 2009, in order to improve the DIS and National Fund system in Croatia.

TABLE 1 AGENCIES CREATED PER YEAR, 1992-2010

Year	Number of agencies established	Total	Year	Number of agencies established	Total
1992	1	1	2002	5	31
1993	3	4	2003	7	38
1994	2	6	2004	5	43
1995	2	8	2005	5	48
1996	2	10	2006	6	54
1997	6	16	2007	8	63
1998	3	19	2008	9	71
1999	3	22	2009	4	75
2000	0	22	2010	2	77 (-15)
2001	4	26	2011	1	63

The other process that led to the cash fiscalization act that entered in vigour on 1.1.2013. was the opposite, the process of centralization of tax collection and the reduction of the role of local government and decentralized tax units.The reason for cash fiscalization was mainly to curb fraud and corruption by citizen and tax authorities, and to control the gray economy in cash transactions, in order to increase tax collection and finance the budget and bloated government structures that were put in place during Sanader era. Those structures were mostly clientelist structures connected with interest groups, that were in just in one part justified by the need of EU integrations and functions associated with this process and distribution of EU funds. In fact, it would almost seem that their function didn't influence their size and structure, but rather other way around, which may, perhaps humourously, be compared to certain tumorous outgrowths in human bodies.

Both processes were largely dependent on the introduction of e-government. The first one because it would have been hardly possible to maintain such a complex decentralized structure of agencies, EU fund management and tender management that went all the way down to the end users without e-government assistance, and the second one because it would have been impossible to centralize tax collection from the smallest tax payer directly to the central Tax Administration without the

assistance of e-government in the form of fiscalized cash registers and permanent on-line connection to the MoF.

Decentralization of funding is a common way of improving the quality of funding and it is designed so that it supports the projects of best quality. On the other hand, the centralization of tax collection is a way of increasing the tax collection and making it more equally distributed to all tax payers, in order to eliminate disloyal competition from grey economy and enable more funds to be disbursed to the projects of best quality. The main problem with this in Croatia is the culture of corruption and distrust of state authority that was already present in before the creation of independent Croatia. European integrations introduced new values of transparency and impartiality and also specific rules governing conflict of interest, which were, in our opinion, reflected in these two divergent processes.

II. AGENCIFICATION IN CROATIA – CASE OF MOSES

The increase of the government agencies is clearly in corelation with the growing of IT usage by firms, citizens and government, and the progress of e-government project in Croatia. In 2009 there were in total 75 agencies.

Agencies are defined as organizations that perform functions at arm's length from government. Such organizations with a relative independence from government have had a long tradition in Croatia, e.g. Chamber of Commerce. The trend of intense agencification in the period 2001-2009 may be explaind by new political and administrative values that were in some degree of correlation with European integrations:

I Democratization and decentralization

II 3 E – effectiveness, efficiency and economy as part of NPM – New Public Management Agenda

III Europeanization

IV Rationalization of the public sector

It can be shown that the informatization of public services that was also present in this period, and that intensified in the period of 2005-2009 with the introduction of the project of e-government, was also motivated by those values.

The normative underpinning for agency creation in Croatia came from the formal obligations in the Stability and Association Agreement (SAA) signed with the EU (2001).[1] In fact two third of all agencies were founded in relation with EU accession. However, this may not be seen as only motivational factor in agency creation. Some other factors may also be present, as the pollitically elected functionaries in the government recognized that liberating some parts of civil service from legal obligations of general state administration could be useful to it for its own purposes.The agencies were founded directly by the government, by an Act or Regulation and were not a part of standard civil service, although they operated in coordination with ministries that served as a hub. As for type, there are executive, regulatory and expert agencies.The Ministry of Science, Education and Sports (MoSES) thus had several executive agencies, such as the Agency for Science and Higher Education, Agency for Vocational Education and Training and Adult Education, and as well as others. The idea was to liberate the agencies from the direct political pressure, as well as from legal requirements that were obligatory in general state administration, thus leaving the public managers to manage their „business“. This complex system of agencies depended largely on informatization, which was previously under the responsibility of CSO for e-Croatia, curently Ministry of Administration (former CSO for Administration). It has been argued that IT played an important role in realizing the NPM. The emphasis on performance indicators and performace targets (Pollitts element 2) which require the continuous generation, monitroing and analysis of organisational statistics, is proedicated on an organisational information infrastructure only made feasible bs electronic IT, as well as the increasing use of contractual arrangements (element 4) in which contractors are required to reularly communicate and work with government agencies and also to report their performance as a means by which to assess whether contractual obligations are being met, requires the deployment of a networked electronic information infrastructure. Managerialism (element 3) similary requires such infrastructure.[3] It is clear that such infrastructure is generally similar to the one used in EU project management and supervision, that is notably one of the most important functions of new agencies.The transformational effect of e-government is thus visible not directly, but indirectly as enabling the process of agencification as a means of introducing new administrative and political values of democratization, decentralization, efficiency and rationalization of the public sector in relation with european integrations and europeanization, but also underpinned by the values of NPM. On the other hand, although the claims that ICT technology provides for greater transparency of government activities and structure, in the period of intense agencification and e-government project (2005-2009) Croatia experienced an increase in corruption on all levels. That was apparently due to inability of public managers to manage because of their lack of training or incompetency, and mostly because the governing structure that made certain parts of administration exempt from legal obligations enabled them to behave as private companies that had a board of directors, which were in fact members of the ruling party (Croatian Democratic Union), with the prime minister as the president and CEO of this „company“. As „middle managers“ they had little independence from the big boss at the top and could be easily replaced, as there were no performance standards other than obeying the command from above, and legal safeguards were weakened, so that they could not resort to them as ordinary civil servants. Thus the process of agencification in Croatia was not only the process of decentralization, but also of concentration of power in the hands of certain political structures and interest groups, with the most powerful structure at the top that consisted of the prime minister and the ruling party leadership. The e-government structures aided in such development by concentrating the executive power in the agencies, liberated from the legal control typical for general state administration, and left to strong „management“ control of

the political leadership in the government and the ruling party. Therefore, it is not clear whether the agencification really improved the effectiveness and efficiency of the government in general, and what were the real effects of apparent democratization and decentralization, although it enabled the management of EU funds and IT structures and the introduction of e-government were largely responsible for this improvement in functioning of Croatian government.

III. CASH FISCALIZATION AS NEW CENTRALIZATION

The need for better financial control that resulted from the corruption of structures that were left uncontrolled and concetrated power in Sanader's government era, especially local administrations that were also controlled by central „top managers“ in Zagreb, led to the need of re-centralization of tax administration, by the practice of cash fiscalization. The legal underpinning for this new practice is rather simple and technical. This, in fact represents the effort of the state to maintain the balance between the new, principle-based and value oriented legal framework that was introduced by adjusting to European Acquis, and old, rule-based framework that was simple and cheaper to put in practice and execute. Employing ICTs in cash fiscalization as a means for better tax collection and curbing grey economy represents a shift from the values of democratization and decentralization towards the values of 3 E. However, the „management“ approach was abandoned and instead replaced by „technical“ aproach, which is based on purely technical solutions for carrying out public policies (such as taxation). In the case study of e-tax in Japan, two key factors were examined: the increase in government performance in tax collection and the responsibility of IT structures and e-government for this improvement. [5] It seems clear that the increas in performance was in fact really due to e-government and IT structures, but it has also raised many worries about the surveillance capacities of the government and the possibility of violating privacy rights of individuals. Although such initiatives are usually successful, as they reduce the burden of tax administration on individual and business taxpayers, while providing self-assessment possibility, it is also clear that this depends largely on a wider political culture regarding public administration, and the relationship between taxpayers and the state. [3] In Croatia, the relationship of taxpayers towards the state and the pervasive culture of corruption may be problematic and account for difficult application of such e-government tools for increasing government efficiency and performance. In fact, Croatian people have not had a state for a long time and the dominant culture of „cheating“ the state was present during that period as a means of opposition toward the foreign power expressed through the state. Nowadays, the notion of such power was replaced by „neoliberalism“ and its policies, and the state is no longer seen as an enemy, but rather its neoliberal policies that are regarded as being imposed by powers alien to the state.

IV. CONCLUSION

In this paper it was suggested that both the proces of decentralization of funding as well as the process of centralization of tax collection were spurred by EU integrations and enabled by the initroduction of e-government structures. The paper's aim was to asses the transformational impact of e-government on government structures, for which the empyrical data were used. However, it would seem that the actual transformational impact of e-government in the first case was limited because of the lack of change of organisational culture in public administration, that was polluted by pervasive corruption and can be characterised as state capture by malignant structures outside of government, and on although the second case has not yet provided sufficient data to make final conclusions about the transformational impact, it would appear that its transformational impact is mainly directed towards those tax paying firms that have already been in default or close to it, and that would have disappeared from the market anyway. The significant increase in tax collection is not to be expected, and therefore the positive development in terms of stimulating the economy through EU funds will probably not be achieved. The reason for this may be seen in the refusal of current government to reduce the size of the administration and increase its efficiency and effectivenes, that was recently recognized by the rating agencies, such as Moody's, that reduced the rating of Croatia from Baa3 or investment rating, to Ba1 or „junk“ as it is colloquially known.

BIBLIOGRAPHY

[1] Musa, Koprić: What Kind of Agencification in Croatia? Trends and Future Directions, Transylvanian Review of Administrative Sciences, Special Issue, pp 33-53

[2] Gasco, M.: Implementing e-government Projects: Organizational Impact and Resilience to Change, in

[3] Henman, P.: Governing Electronically, E-Government and the Reconfiguration of Public Administration, Policy and Power, Palgrave Macmillan

[4] Homburg, V: Understanding E-Government, Routledge, 2008

[5] Chatfield, T.A: Public Service Reform through E-Government , a Case Study of „E-tax“ in Japan

[6] Politt, C. and Bouckaert, G.: Public Management Reform: A Comparative Analysis, 2nd ed., Oxford: Oxxford Univestiy Press, 2004. York: Routledge, 2004, pp. 22-49.

[7] Allen, R., Tomasi, D: Managing Public Expenditure – A Reference Book for Transition Countries, OECD, Paris, 2003

[8] Pilat, D.: Economic Impact of ICT, OECD, Paris, 2004

[9] Šimurina, J., Hruška,D., Marković, M.: E-Government in Croatia

E-Hrvatska, jedinstven informacijski sustav

Prof.dr.sc Slavko Šimundić*, Josipa Šimić, mag.iur.**
* Sveučilište u Splitu, Pravni fakultet, Split, Hrvatska
slavkosimundic@pravst.hr
** Pučko Otvoreno učilište Algebra, Hrvatska
josipa.simic@predavaci.algebra.hr

Sažetak - E-Hrvatska projekt je informatizacije s kojim se započelo 2000.godine. Sama realizacija projekta započela je 2003. godine kada je definiran kao jedan od prioriteta Vlade RH. Danas, 12 godina poslje, pitamo se koliko je uspješno proveden taj projekt.
Projekt je zamišljen i prezentiran javnosti kao jedinstven informacijski sustav u cijeloj Hrvatskoj. Da li su poštovana pravila izgradnje informacijskog sustava, da li su provedeni svi koraci koje bi takav rad zahtjevao i koje su to vrijednosti koje možemo mjeriti da bismo projekt e-Hrvatska mogli proglastiti uspješnim ili neuspješnim.
Hipoteza ovog rada je da ne postoje jedinstveni parametri kojima se projekt E-hrvatska može svrstati pod nazivnik informacijskog sustava. Da bismo provjerili tu tvrdnju pokušat ćemo pronaći zajedničke parametre svih projekata unutar projekta E-Hrvatska te utvrditi kriterije za njihovu valorizaciju.
Cjelovita slika e-Hrvatske i njezinih podprojekata (e-Uprava, e-Sudstvo, zemljišne knjige, e-Zdravstvo i dr.) pruža nam mogućnost uočavanja zajedničkih značajki, prednosti i nedostataka, analiziranje koncepta provedbe te analizu prednosti i nedostataka rascjepkanosti provedbe projekta na više manjih projekata.

Ključne riječi: e-Government, e-Hrvatska, jedinstveni informacijski sustav.

I. UVOD

Danas, u 2013. godini, na pragu ulaska Hrvatske u Europsku Uniju postavlja se pitanje koliko je Republika Hrvatska, poučena dosadašnjim povijesnim iskustvima, jaka i spremna na očuvanje vlastitih interesa i zadovoljenje potreba vlastitih građana. Da li je to država koja je na temelju svojih povijesnih tekovina i razvojnopolitičkih dostignuća spremna u izazovima suvremenog svijeta, ispunjavati potrebe svojih građana ili je kao što je ponekad kroz povijest bila, spremna biti tek marioneta nekih vanjskih interesa?

Ustav iz 2001. godine navodi: „*Polazeći od iznesenih povijesnih činjenica, te općeprihvaćenih načela u suvremenu svijetu i neotuđivosti i nedjeljivosti, neprenosivosti i nepotrošivosti prava na samoodređenje i državnu suverenost hrvatskog naroda, uključujući i neokrnjeno pravo na odcjepljenje i udruživanje, kao osnovnih preduvjeta za mir i stabilnost međunarodnog poretka, Republika Hrvatska ustanovljuje se kao nacionalna država hrvatskog naroda i država pripadnika autohtonih nacionalnih manjina.*“[1]

U Hrvatskoj vlast se dijeli na zakonodavnu, izvršnu i sudsku, gdje je uprava samo dio izvršne vlasti, ali kako je to najniži oblik kojim država „pruža ruku svojim građanima“ u razmatranju informatizacije Republike Hrvatske, upravom ćemo se također pozabaviti.

U veljači 2003. Godine Hrvatska se kandidira za članstvo u Europskoj Uniji, na koji dobiva pozitivan odgovor sredinom 2004. te 3.listopada 2005. započinje s pristupnim pregovorima. Promjene koje se očekuju i koje je EU zahtijevala i još zahtjeva od RH opsežne su i dalekosežne, i nužno je da svi građani Hrvatske budu na transparentan način upoznati s procesima i promjenama koje nas očekuju. Transparentnost informacija je jedan od segmenata o kojem ćemo razmišljati te ga je izuzetno bitno navesti već u samom početku, zbog ogromnog utjecaja koji mediji danas imaju u formiranju javnog mijenja u svim područjima života, pa tako i u ovom za Hrvatsku iznimno bitnom pitanju pristupanja EU.

II. POTREBA ZA INFORMATIZACIJOM DRŽAVE

A. *Odgovor moderne države na potrebe modernog društva*

Poimanje države i njenih zadataka mijenjalo se tijekom vremena; do 18.st. glavni zadatci države su uključivali financije, pravosuđe, diplomaciju, vođenje ratova te održavanje unutarnjeg poretka. Nakon drugog svjetskog rata većina vlada preuzela je odgovornost za sigurnost, obrazovanje, ekonomski napredak i socijalno blagostanje svojih građana.[2]

Ako država služi ispunjenju potreba svojih građana, moramo se zapitati koje su to potrebe koje svakodnevni život stavlja pred pojedinca. To su svakako prvo osnovne fizičke potrebe. Razmatrajući pak i stanovišta iznesena u Općoj deklaraciji o ljudskim pravima UNa, kao što su pravo na život, slobodu i osobnu sigurnost, pravo na zaštitu privatnog i obiteljskog života, slobodu kretanja i izbora stanovanja, pravo na državljanstvo i posjedovanje imovine, pravo na i slobodu mišljenja i mnoga druga, te uzimajući u obzir sve životne situacije u kojima se sva navedena (ali i mnoga druga prava i potrebe) trebaju moći ostvariti na jednostavan, učinkovit i neupitan način, postavljamo pitanje koja to i kakva vlast može ostvariti?

[1] Preuzeto sa: http://narodne-novine.nn.hr/clanci/sluzbeni/232289.html
[2] Kostakoplou, 2002.

B. Da bi to ispunila kakva država mora biti?

Da bi ispunila svoju zadaću u modernom društvu vlast, a time mislimo prvenstveno na modernu državu, mora biti efikasna i učinkovita, jednaka prema svim svojim građanima u zaštiti svih njihovih prava i ostvarivanju svih njihovih sloboda na cijelom svom teritorijalnom području poštujući sve zakone i uredbe koji proizlaze iz Ustava kao konstituirajućeg akta te iste vlasti.

Ovdje se uputno osvrnuti na razmišljanja profesora Pusića o ulozi državne vlasti u svakodnevnom životu; „*Državna uprava zadužena je za neposrednu provedbu zakona, donošenje propisa za njihovu provedbu, obavljanje upravnog i inspekcijskog nadzora te druge upravne i stručne poslove. (...) Rad tijela državne uprave je javan. Tijela državne uprave su dužna upoznati javnost i izvještavati je o obavljanju poslova iz svog djelokruga, davati građanima podatke, obavijesti i upute te im pružati stručnu pomoć u poslovima zbog kojih im se obraćaju*."[3]

Temeljno obilježje suvremenog svijeta je dualizam između tradicijskog načina življenja (…) i novog, postmodernog stila života.[4] Globalizacija se u ovim ili onom aspektu uvukla u sve pore života pa nam danas pojam globalnog sela [5] ne predstavlja ništa novo i spektakularno. Više ne razmišljamo o televizijskom signalu, radio signalu, GMS signalu, a kamoli o Internetu kao fenomenu suvremenog svijeta. Ne samo da se ne pitamo kako smo samo živjeli bez svega toga, nego što je daleko bitnije, ne propitujemo kakve pozitivne ili negativne utjecaje radio, tv i moderna komunikacija ostavljaju na društvo i pojedinca u njemu.[6]

Prvotno obilježje današnjeg svijeta je složenost, a tek nakon nje neusklađenost pa shodno tome i potreba za organiziranošću. „*Osnovna je karakteristika današnje epohe, posebno druge polovine dvadesetog stoljeća, pojava i neslućeno narastanje složenosti u svim oblastima suvremenog života – od svakodnevnog života i rada u grandioznim mnogomilijunskim urbanim kompleksima, do planetarne mreže zračne komunikacije. Bez presedana u cjelokupnoj dosadašnjoj ljudskoj povijesti, složenost je sveprisutna realnost svijeta u kojem živimo. U središtu suvremenog teorijskog i praktičnog zanimanja, fenomen složenosti, ne samo da stalno proširuje postojeće tijekove intelektualne spoznaje, već i neposredno uzrokuje nastanak novih koncepcija, metodologija i "tehnologija" znanstvenog saznanja, s jedne, kao i konstrukciju složenih instrumenata i njihove praktične aplikacije, s druge strane. Fenomen složenosti – taj duh iz svjetiljke koji se više ne da u nju vratiti – narasta do neslućenih razmjera i zahvaća pojave, odnose unutar i odnose između gotovo svih oblasti od osobnog i društvenog značaja*".[7]

Postoje naravno mnogi znanstveni i stručni radovi iz mnogih društvenih područja koja se osvrću na navedene fenomene. Profesor Lozina na primjer smatra da je „*Suvremeno društvo karakterizirano je velikim, prije svega tehnologijskim promjenama. Globalizacija i internetizacija pretvorile su svijet u mali džepni mobitel…*"[8] Mnogi autori slažu se s ovim razmišljanjem držeći da je otuđenje i asocijalizacija jedan od najizraženijih aspekata komunikološke globalizacije, te da homo sapiens današnjice mada umrežen u čitav svijet putem Interneta ostaje sam i otuđen u jednom eminentno depresonaliziranom svijetu.[9]

Ipak postoji i veliki broj eminentnih znanstvenika koji drže da su dobrobiti i prosperitet ostvareni novim tehnologijama uvelike nadmašili nedostatke i manjkavosti tih istih pojava. Nedvojbeno je da se društvo razvija, da se tehnološki napredak, a time ni razvoj komunikacija više ne mogu spriječiti. Pažljivo uzimajući u obzir nedostatke i loše utjecaje informatizacije društva, ipak bih prepustili kompetentnijim stručnjacima iz oblasti sociologije, psihologije i drugih nauka da se pozabave negativnim uplivima modernih pojava na današnje društvo, a u daljnjem izlaganju u ovom radu pozabavili se pozitivnim aspektima navedenih pojava i mogućnostima da ih još bolje iskoristimo u razvoju vlastitog društva i države.

Time zapravo odgovaramo na pitanje: čemu informatizacija? Država mora početi promatrati moderne tehnologije i znanosti, kako tehničke tako i društvene, poglavito informatizaciju, kao alat u ostvarenju svojih ciljeva. A kao što smo već prikazali kroz povijesni razvoj države i politike, uvijek se provlači jedna nit koja veže tkivo naroda uz državu: ostvarenje fundametalnih životnih potreba.

C. Informatizacija države

Nedvojbeno je da nove tehnologije mogu omogućiti državama lakše i jednostavnije ispunjenje zadanih ciljeva, pojednostavniti komunikaciju kako između pojedinaca i struktura koje predstavljaju državu tako i između institucionalnih jedinica i građana. Nedvojbeno je u današnjem modernom društvu da pojedinac od države očekuje napredak i prosperitet koji se mora odraziti u nastojanjima države da koristi sva raspoloživa sredstva da osigura najbolje za svoje građane. Dvojbeno je u kolikoj mjeri je država spremna iskoristiti pozitivna moderna dostignuća u isto vrijeme ograničavajući njezine negativne pojave.

Postoje struje u znanosti koje smatraju da nove tehnologije potiču uključivanje zajednice ne više samo u globalni svijet i otvaranje prema istom, nego u lokalne akcije i odnose povezujući time globalno s lokalnim, učvršćujući narodni identitet, veze unutar manje zajednice i osjećaj pripadnosti jednom narodu. Za takvu vrstu zajedništva unutar informacijske teorije spominje se izraz „glokalizacija". Taj termin označava prožimanje globalnog i lokalnog, njihovu nužnu interferenciju, odnosno utjecaj procesa globalizacije na lokalno društvo.

[3] Pusić, 2002.

[4] Lozina, 2004.

[5] Izraz koji je 60ih godina skovao teoretičar Marshall McLuhan, koji je kazao da se zahvaljujući elektroničkim medijima vraćamo na seoski način života.

[6] Za detalje videti: Miss Representation; projekt o prikazu žena u medijima; http://www.missrepresentation.org/

[7] Šimundić, 2007.

[8] Lozina, 2004.

[9] Ibis

Dakle glokalnost odražava globalnost u lokalnom kontekstu.[10]

Ovdje je potrebno istaknuti i razmišljanja profesora Eugena Pusića[11] o državi kao organizacijskoj jedinici koja je ustrojena radi zadovoljenja potreba njenih građana. „*Politički sustav je skup institucionalnih struktura i elemenata orijentacije u svijesti upravljanih na utvrđivanje i primjenu općih kriterija za dodjeljivanje šansi zadovoljenja interesa u grupi ljudi obuhvaćenih tim sistemom*“[12]. Ovakvo promišljanje političkog sustava vrlo je blisko organizacijskom poimanju sustava uopće. „*U kibernetičkom stajalištu sustav se shvaća kao nešto prepoznatljivo, sređeno prema nečem što je nesređeno, ili skup elemenata povezanih najmanje jednom ili više prepoznatljivih korelativnih povratnih veza.*“[13]

Prema tome državu možemo promatrati i s političko-povijesnog i sa organizacijskog stajališta u koje moramo uključiti moderna saznanja kibernetike kao „*suvremene teorije upravljanja i odlučivanja produktivnim i neproduktivnim sustavima*.“ [14] Ali prvenstveno, državu moramo razmotriti s pravnog, teritorijalnog i upravnog aspekta, ali u svakom aspektu koristeći i saznanja moderne hrvatske znanosti čiji je zadatak uvijek se kritički osvrnuti na postojeće stanje te ponuditi bolja i efikasnija rješenja.

Budući da institucije države zapravo čine politički sustav, njihove karakteristike pa među njima i informatiziranost te stupanj i tip organizacije uvelike će ovisiti o organiziranosti izvora podataka o kojima ovise. Stoga organizacija a time i informatizacija, koja pruža optimalnu organiziranost podataka, treba biti u fokusu razmatranja šire akademske zajednice svake države koja želi postići visoku efektivnost i produktivnost.

2000. godine Europa i svijet nalazili su se u dobu koje je okarakterizirano kao informacijsko društvo, društvo znanja koje se gradi na razvoju i primjeni informacijskih tehnologija. Temeljne funkcije i procesi u informatizacijskom društvu uglavnom su organizirane oko mreže. Mreža presudno određuje društvenu aktivnost u zajednici, a implikacije mreže za državu postaju fundamentalne. Države se prilagođavaju novom informacijskom dobu, mijenjanjem svoje uloge iz suverenih sudionika svjetske politike u sukreatore iste, zajedno s ostalim akterima međunarodne zajednice.[15]

D. E-Government

Neki od načina preobrazbe države iz klasičnog modela upravljanja i klasične uprave, preko stvaranja Nove javne uprave je stvaranje e-Government-a.

E-Government predstavlja primjenu komunikacijske i informacijske tehnologije kako bi se osigurala efikasna, efektivna, transparentna i pouzdana interakcija između vlada na međunarodnoj razini, između vlade i vladinih agencija na federalnoj, regionalnoj i lokalnoj razini, kao i između vlade te građana i poslovnih ljudi. Ona mora omogućiti ljudima olakšan pristup i korištenje informacijama.[16]

Pojavom nove javne uprave, prvenstveno u Velikoj Britaniji i SAD-u pod vladavinom Margaret Thatcher i Ronalda Raegana, polažu se solidni temljeni razvoju današnjih e-Governmenta na čvrstim temeljim neoliberalizma u političkom poimanju države kao subjekta koji na minimalistički način se upleće u rad, individualnost i slobodu pojedinca.

Do početka 21.stoljeća koncept nove javne uprave se globalno razvio pod utjecajem razvoja mnogih znanosti, među inima informatike, komunikologije, menagementa i drugih. Pojave debirokratizacije, decentralizacije, menandžerskog upravljanja javnim servisima, poduzetništva, denacionalizacije i mnoge drugih činile su uzbudljiva razdoblja u razvoju mnogih zemalja.

Primjeri tih zemalja trebali su i trebaju biti, sada i ubuduće, primjeri na kojima će se temeljito i sustavno graditi informatizacija Hrvatske kako bi ista na pragu ulaska u novu državno-pravnu zajednicu imala odlučnu snagu u zaštiti nacionalnih interesa i mogla konkurirati velikim državama kako gospodarski, ekonomski, kulturno i u svakom drugom području.

Ključni aspekti kroz čiju prizmu treba promatrati E-Government i njegovu provedbu u svakodnevnom radu javnih tijela, a koja možemo naučiti iz iskustva drugih zemalja, su sljedeća:

Interna upotreba računala i računalnih mreža već se pokazala relevantnom prednošću za javnu upravu, zbog lakšeg upravljanja velikom količinom podataka i smanjivanjem njihova repliciranja. Ovoj nužnosti treba dodati i mogućnosti koje nam pružaju računalne mreže čime se osigurava elektroničko čuvanje podataka i njihova višestruka dostupnost na lak, jednostavan i učinkovit način.

Javnost kao aspekt uporabe podataka možemo promatrati samo kao nadležnost pojedinih agencija da upravljaju tim istim podatcima (najčešće osobnim podatcima građana), ali to pitanje možemo i proširiti na prava građana da raspolažu podatcima od važnosti za javni interes (statistički pokazatelji, financiranje javnog sektora i sl.

Dok je Javnost jednosmjerna komunikacija od javnih službi prema građanima Komunikacije bi trebale biti dvosmjerna komunikacija između javnih službi istog i različitog ranga međusobno, te javnih službi i građana. Takva komunikacija može se odvijati osobno ili što je preporučljivije telefonom, elektronskom poštom, raznim web portalima i slično.

Transakcije su ujedno i najkompleksniji aspekt jer podrazumijevaju protok ne samo vrijednosti kao što je novac, nego i protok pravno obvezujuće korispodencije putem Interneta. Cilj provedbe ovog aspekta trebao bi biti potpun protok informacija: građani koji elektronskim putem podnose obrasce za dobivanje dokumenata, relevantni podatci koji se šalju za to nadležnoj agenciji i tamo obrađuju automatski tamo gdje je to moguće,

[10] Ibis

[11] Eugen Pusić (1916.-2010.), hrvatski akademik i profesor, redovni profesor Pravnog fakulteta u Zagrebu.

[12] Pusić, Eugen, Razvedenost i povezanost, Biblioteka Moderen

[13] Žužul, 1996.

[14] Šimundić, 2007.

[15] Lozina, Klarić, 2003.

[16] Ibis

rješenja i odluke javnih tijela poslane u elektronskom, ali ipak pravno važećem obliku i slično.
One Stop Shop je servis putem kojeg građanin može jednostavno i brzo riješiti poslove za koje mu trebaju određenja ovlaštenja koja posjeduje samo javno tijelo: promjena prezimena, registracija vozila, promjena prebivališta, izrada osobnih dokumenata, pribavljanje raznih dozvola i rješenja su poslovi koji u klasičnoj upravi građaninu pojedincu oduzimaju puno vremena i financijski su zahtjevna. One Stop Shop rješenja, su projekti putem kojih bi građanin na jednom mjestu dobio relevantne informacije, digitalno ispunio potrebne obrasce, izvadio dokumente, dobio dozvole i rješenja. [17]
Svi ovi aspekti ponaosob, ali i projekt e-Governmenta u Hrvatskoj zahtjeva pažljivo razrađen i pomno promišljen legislativni okvir koji bi zadovoljio obe strane; građanina u nastojanjima da na jednostavan, transparentan i učinkovit način ostvari svoja prava i zadovolji svoje potrebe i javne službe u nastojanju da učinkovito i po pravilima struke rade svoj posao.

III. INFORMATIZACIJA REPUBLIKE HRVATSKE

Uvažavajući prethodno navedena razmišljanja, neupitno je da državu možemo promatrati kao organizaciju i da se Hrvatska nalazi u vremenu u kojem je e-Government bio logičan sljedeći korak u razvoju države.

Uspostavom Hrvatske države u vrijeme obrane od agresije, razumljivo je da je državna politika imala većih i važnijih prioriteta od informatizacije i unaprjeđenja uprave što bi trebao biti prioritetni zadatak svake efikasne države. U poslijeratnom razdoblju obnova ratom pogođenih područja, rješavanje pitanja socijalne sigurnosti građana i zadovoljenja njihovih primarnih potreba te uspostava centralizirane vlasti postale su primat ne samo vladajućim političkim opcijama nego i svim slojevima društva koje je težilo normalizaciji života, radu i sigurnosti, ali nažalost i vidanju rana, bilo fizičkih, psihičkih ili socijalnih u svim slojevima društva. Ipak pod utjecajem globalizacije prodor novih informatičkih tehnologija u sve pore svakodnevnog života nije se mogao zanemariti.
U vrijeme kada su druge europske države koje nam danas predstavljaju uzor prolazile kroz faze decentralizacije, jačanja socijalne države i uvođenja (tada) novih tehnologija u ustrojstvo i rad države, Republika Hrvatska se trudila svojim građanima osigurati osnovne uvjete mira, sigurnosti i rada.

Tek afirmacijom težnji za ulazak Hrvatske u Europsku Uniju postavilo se pitanje učinkovitosti i efikasnosti državnog aparata koji bi mogao „konkurirati“ državnim aparatima europskih država s kojima je Hrvatska željela ući u zajednicu.

Ipak unatoč svemu nedopustiva je situacija u kojoj se nedovoljno promišljalo o problemima koji su drugim državama osigurale napredak i razvoj a koji su zahvaljujući sustavnom zanemarivanju Hrvatsku doveli u neravnopravan položaj. Problemi su se rješavali stihijski i napola gdje je najčešće izvlačen „čarobni štapić“ centralizacije i hijerarhizacije, bilo da se radilo o sudstvu, upravi, zakonodavstvu ili teritorijalnoj podjeli.

Za uspješno funkcioniranje svake organizacije kao sustava potrebno je izgraditi njezin informacijski sustav. [18] Ako državu za potrebe ovog razmatranja shvatimo kao organizaciju u širem smislu riječi, moramo se složiti s navedenom definicijom. Također se slažemo da na kvalitetu svakog informacijskog sustava utječu mnogi čimbenici, da li se o svakom od njih podrobno razmišljalo i svaki korak postepeno provelo u procesu informatizacije Hrvatske, ostaje upitno.

Nažalost svjedoci smo da se ključne odluke donose bez dugotrajnog promišljanja, brzopleto i ovisno o trenutnom političkom interesu struje koja je na vlasti. Dugoročni planovi i programi su najčešće mrtvo slovo na papiru koje jedan politički splet okolnosti promjeni u jako kratkom vremenu. Smatramo da je tako velike odluke kao što su one vezane za proces informatizacije koje će imati dalekosežne posljedice zahtijevaju više rada i promišljanja a posebno uključenja akademske zajednice koja se dugo bavi tom problematikom.

Pitamo se, slijedom navedenog, da li rezultat (ili rezultate) procesa informatizacije Hrvatske možemo promatrati kao informacijski sustav? Što je to što podrazumijevamo pod pojmom informatizacije Hrvatske i koji su to elementi, planovi, projekti i segmenti koji obuhvaćaju tako širok pojam?

Podrazumijevajući sveobuhvatnost teme koju razmatramo, bez obzira na neostvarenje političkih planova ili njihovu neefikasnost moramo se osvrnuti na planove i programe koje su donosile institucije vlasti u RH.

A. Planovi i programi

1991. Republika Hrvatska prelazi iz socijalističko-mješovitog gospodarstva u funkcionalno tržišno gospodarstvo. Uočeno je da se na tom putu nalaze problemi koji zahtijevaju strateško planiranje i sveobuhvatne akcije da bi se dostigao cilj postizanja povoljnih stopa rasta i održivog razvoja prema uzorima zemalja EU. Problemi vanjske neravnoteže, socioekonomskih razlika kako među građanima tako i među pojedinim regijama, problemi prostornog planiranja i onih vezanih uz očuvanje i zaštitu okoliša, kao i rad na primanju Republike Hrvatske u punopravno članstvo EU potakli su razvoj klime, kako političke tako i općedruštvene koja je omogućila i RH da se upusti u projekt informatizacije društva.

Jedan od prvih većih koraka desio se 2000. godine kada formirana radna skupinu za izradu strategije informatizacije Hrvatske. Ovaj strategijski dokument postavljao je sva prava pitanja; Što treba učiniti u organima državne uprave da bi ih se osposobilo za strateško promišljanje i ostvarivanje informacijskog društva u Hrvatskoj? Što treba učiniti kako bi se ubrzao ulazak Hrvatske u europske integracije sa stajališta zahtjeva informatizacije? Što treba učiniti kako bi država i javne djelatnosti postali uzorna informatizirana infrastruktura cjelokupnog razvoja Hrvatske? Što treba

[17] Skrobotz, 2007.

[18] Šimović, 2010.

učiniti da informacijska tehnologija doprinese podizanju konkurentne sposobnosti hrvatskog gospodarstva te potakne gospodarske promjene i inovacije?

2003. godine predstavljen je projekt e-Hrvatska; skup aktivnosti i akcija koje za cilj imaju stvaranje uvjeta da se hrvatsko društvo transformira u informacijsko društvo poduzimanjem koordinirane, sveobuhvatne i dinamične akcije, kako bi se građanima i gospodarstvu Hrvatske omogućila najkvalitetnija i najširа moguća uporaba i razmjena informacija i time otvorio prostor aktivnog sudjelovanja u globalnim tokovima..

2006. godine Vlada RH donosi Strateški okvir za razvoj 2006.-2013., kojim je definiran smjer gospodarske politike u navedenom razdoblju. Njime se razrađuju modeli i putovi jačanja Hrvatske kao konkurentnije i bogatije zemlje te utvrđuje prioritete i akcije čije će provođenje osigurati stabilan gospodarski rast, zapošljavanje i bolji životni standard. [19] Glavni ciljevi utvrđeni ovim dokumentom su razvoj ljudi i znanja, razvoj znanosti i tehnologija, poglavito informacijsko-komunikacijske tehnologije, poboljšanje socijalne kohezije i socijalne pravde, unaprjeđenje područja prometa i energije, očuvanje prostora, prirode i okoliša, regionalni razvitak, makroekonomska stabilnost, gospodarska otvorenost, unaprjeđenje financija i kapitala, poboljšanje poduzetničke klime, provedba privatizacije i restrukturiranja te posebno nova uloga države koja se ogleda u poboljšanoj i unaprijeđenoj, modernoj i učinkovitoj upravi.

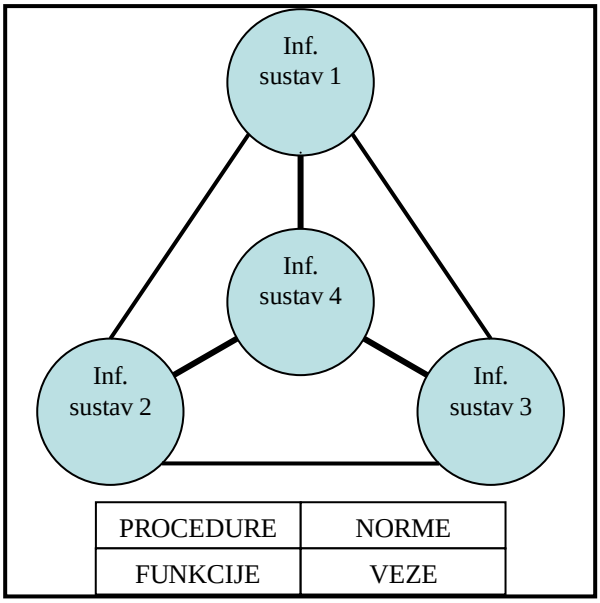
Tijekom godina mnoga su nastojanja da se pobliže definiraju pojmovi, ciljevi i planovi, kako od strane političke scene tako i od strane znanstvene zajednice. Iskazane nade i strahovi povezani s pojmom E-Government nisu ništa drugačiji u Hrvatskoj nego li su u drugim europskim zemljama. Ipak naglašava se da ovaj proces nije sam po sebi lijek za sve probleme. Također je postalo jasno da izrada zakona ili implementacija pravilnika za područje E-Governmenta je zadatak sam po sebi kompleksan i zahtjevan kao i područje kojim se bavi. Postalo je jasno da nikako nije preporučljivo jednostavno prevesti strani zakon o području E-Governmenta na hrvatski. Naprotiv, zakon se treba razviti i iznjedriti na temeljima hrvatskog zakonodavstva vezanog za područje javne uprave.[20]

Ipak, čini nam se da je područje našeg interesa dosta šturo i zatvoreno, do dokumenata se dolazi sporo i teško, državna administracija je još uvijek nedovoljno spremna pomoći znanstveno-akademskoj zajednici u pribavljanju dokumenata na kojima će temeljiti svoja istraživanja, komunikacija s državnim tijelima je štura i odgovori u većini slučajeva, nažalost, dolaze za zakašnjenjem.

IV. ODLIKE INFORMACIJSKOG SUSTAVA U PROJEKTU E-HRVATSKA

Informacijski sustav je skup dizajnerskih informacijskih objekata koji predočuju koncepte modela organizacijskog sustava, i to njihovu prošlost i sadašnjost, implementiranih na resurse koji su ili nositelji informacija ili izvode procese nad podatcima, a informacijski objekti razmjenom informacija komuniciraju (povezani su) međusobno i s okolinom.[21]

Jedinstven informacijski sustav je (nad)sustav koji na jedinstven i sebi specifičan način, procedurama, funkcijama, normama i vezama; ovisno o njegovoj ulozi, cilju i informacijskim (pod)sustavima koje obuhvaća, objedinjuje te informacijske sustave u logičku cjelinu.



Da bismo e-Hrvatska mogli nazvati takvim sustavom on bi u svojim projektima trebao pokazivati zajedničke značajke koje bi se mogle dalje uspoređivati i valorizirati, nažalost na takve zajedničke atribute u našem istraživanju nismo mogli apstrahirati. Razmatrajući svaki od pojedinih projekata unutar projekta e-Hrvatska kao što su e-Uprava, e-Zdravstvo, e-Katastar, Sudski registar, e-Spis, Sudačka mreža i mnogi drugi, nismo mogli uočiti zajednička svojstva prema kojima bi se (bar neki od projekata) mogli podvesti pod već navedenu definiciju informacijskog sustava. Poneki od njih su tek baza podataka podržana interaktivnim pretraživačkim sučeljem.

Nažalost moramo istaknuti da ni cijeli projekt ne sadrži obilježja jedinstvenog informacijskog sustava prema definiciji koju smo dali.

Ipak ne treba biti pretjerano kritičan i zanijekati uspješnost puta koji su mnogi od njih prevalili, od papirnatih knjižurina i prenošenja informacija usmenom predajom, do elektronskih baza podataka dostupnih građanima i sustava koji u svakodnevnom životu pokazuju značajna poboljšanja za građane.

Ono što ističemo jest činjenica da među sustavima projekta e-Hrvatska (ovdje ih više nećemo niti nazivati informacijskima) ne postoji jedinstvenost i konzistentnost.

[19] Izvor: http://www.mrrfeu.hr/userDocsImages/Publikacije/Strateski_okvir_za_razvoj_2006_2013.pdf

[20] Skrobotz, 2007.

[21] Pavlić, 2011.

V. OČEKIVANJA OD E-HRVATSKE

Što očekujemo u daljnjem razvoju projekta e-Hrvatska?.

Funkcionalnost države je opravdano jedan od temeljnih postulata koje njeni građani zahtijevaju. Danas kada je izvrsnost zahtjev svakog tržišta, kada su učinkovitost i isplativost ono što zadržava radna mjesta i stvara nova država mora biti funkcionalnija u svom svakodnevnom radu.

Jednostavnost. Ponekad, u državama kao što je Hrvatska, glavni i primarni problem ne predstavlja samo prilagodba na nova rješenja koja će informatizirana država pružiti svojim građanima. Veći problem je bipolarnost građana i općenito korisnika (tu ubrajamo i razne unutarnje entitete samog sustava) u njihovoj sposobnosti da na njima jednostavan način uopće koriste takve usluge. Veliki postotak građana uopće ne posjeduje potrebna znanja i vještine kao i sredstva (računalo, Internetsku vezu i sl.) da bi mogao se koristiti takvim uslugama. Također ni sve jedinice lokalne uprave i samouprave, nisu spremne u ovom trenutku, ni svojim kapacitetima financijskim, organizacijskim i ljudskim popratiti mogućnost pružanja svih usluga koje su građanima dostupne. Ovaj problem treba gledati kao prednost jer se državi daje razumno vrijeme da provede mnoge prilagodbe i poboljšanja. Ovakva rješenja trebala bi biti dodatna mogućnost u dužem prijelaznom razdoblju u kojem će se moći koristit i stari i novi načini.

Transparentnost. Pod transparentnošću podrazumijevamo jasno definiran zakonski okvir koji će precizno pojedine zakone i pravilnike koji će pak do u detalje pravno regulirati pojedina područja i pitanja. Pitanje procesa izrade zakonodavnog okvira kao i pitanje samog njegovog sadržaja trebali bi pri određivanju obujma i ciljeva takvog posla uzeti u obzir iskustva i znanja s terena. Praktična iskustva najnižih i srednjih razina administracije koja najbolje iz praktičnog života mogu potvrditi i definirati potrebe krajnjih korisnika u takvom procesu.

Financijska isplativost je stavka koju nije jednostavno postići, poglavito zbog činjenice da informatizacija sama po sebi donosi značajne financijske izdatke. Novi hardware, novi software, edukcija službenika i implementacija novih rješenja uvijek donose nove financijske izdatke. Potrebna je nova računalna oprema, podatkovna infrastruktura, skladišta podataka, razne baze podataka, mrežna rješenja, računalni stručnjaci koji za sve navedeno posjeduju ekspertna znanja i čije usluge sigurno nisu zanemariv financijski izdatak. Potrebna je edukacija službenika u javnim službama, „pametna rješenja“ vezana za digitalne potpise i certifikate, transparentna evaluacija znanja i vještina tih istih službenika.

Ali ako se takvim načinom organizacije postigne da jedan visoko profilirani, visoko stručni državni službenik obavlja višestruko više posla u istoj jedinici vremena, radi to brže i profesionalniji ili čak preuzima i odrađuje posao za koji su ranije trebala tri ili više službenika onda se može govoriti da je cilj postignut.

Smatramo da u ovom dijelu postoji jedan aspekt financiranja, koji Hrvatska nije dovoljno dobro iskoristila. Projekti Europske Unije PHARE i SIGMA (Support for Improvement in Governance and Management in Central and Eastern European Countries) čiji je glavni cilj bio pomoć institucijama RH da dosegnu europsku razinu su nedovoljno iskorišteni i tek je manji dio sredstava uspješno povučen iz njih.

VI. ZAKLJUČAK

U ovom radu nismo se dotakli još mnogih drugih projekata koji su u planu, u nekom od stupnjeva provođenja ili su već i zaživjeli; kao što je e-Zdravstvo, e-Uputnica, elektronske prijevozne karte u javnom prometu, uvođenje OIBa i mnogih drugih. Ali radi ograničenosti dostupnih podataka, kako na Internetu, tako i u radovima akademske zajednice, ali i u tromosti sustava državne uprave od kojih smo mnoge podatke tražili dopisima elektronske pošte, na koje nikad nismo primili odgovor, ostavili smo ih za razmatranje u nekom budućem radu.

U nadi da je proces informatizacije, iako se teškim početkom, ipak uzeo zamaha u Hrvatskoj i da ćemo uskoro dostići zemlje zapadne Europe ali i Skandinavije koje nam na ovom polju mogu biti istinski uzori, gledajmo u budućnost i nadajmo se najboljem.

Ipak, treba uzeti u obzir da se Hrvatska nalazi na jednom dugotrajnom putu i da nipošto sva dosadašnja nastojanja ne treba ni odbaciti ni omalovažavati, dapače treba ih studiozno proučavati kako bismo na temelju dosadašnjih iskustava, svojih i tuđih, mogli napraviti nužna poboljšanja i jednog dana ponosno reći da Hrvatska ima jedinstven informacijski sustav koji služi kao uzor drugim zemljama.

LITERATURA

1. Koprić, Ivan (1999) Struktura i komuniciranje u upravnim organizacijama. Zagreb: Pravni fakultet u Zagrebu
2. Kostakoplou, Dora: Floating Sovereignty: A Pathology or a Necessary Means od State Evolution?, Oxford Journal of Legal studies, 22, 2002.
3. Lauc Zvonimir: Lokalna samouprava u Republici Hrvatskoj.de lege ferenda, Zbornik Pravnog fakulteta u Zagrebu,
4. Lozina, Duško.; Klarić, Mirko.; Nova javna uprava; Sveučilište u Splitu, Pravni fakultet, Split 2003.
5. Lozina, Duško: Lokalna samouprava, poredbeno i hrvatsko iskustvo; Sveučilište u Splitu, Pravni fakultet, 2004.
6. Pusić, Eugen; Nauka o upravi, Školska knjiga, Zagreb, 2002.
7. Skrobotz Jan: E-Government, Chances, Checks and Concepts, Hrvatska javna uprava, br.1., 2007., slobodan prijevod
8. Šimović, Vladimir: Uvod u informacijske sustave, Golden marketing-Tehnička knjiga, Učiteljski fakultet Sveučilišta u Zagrebu, 2010.
9. Šimundić Slavko: Pravna informatika, Sveučilište u Splitu, Pravni fakultet, Split, 2007.
10. Šimundić Slavko: Upravna informatika, Sveučilište u Splitu, Pravni fakultet, Split, 2010.
11. Šimundić, Slavko; Barbarić, Danijel; Šimić, Josipa; modernizacija uprave s posebnim osvrtom na informatizaciju u kontekstu Europskih integracija; Zbornik radova- „Teritorijalni ustroj i regionalizacija Republike Hrvatske u kontekstu Europskih integracija“, Split,2010.
12. Žužul, Josip; Dragičević, Dražen: Informatika u upravi i pravosuđu, Zagreb, 1996.
13. Žužul, Josip; Teritorijalni ustroj i regionalizacija Republike hrvatske – informacijski pristup; Zbornik radova Pravnog fakulteta u Splitu, 2010.
14. Pavlić, Mile: Informacijski sustavi, Školska knjiga, Zagreb 2011.

Mogućnosti primjene računalstva u oblaku u tijelima lokalne uprave i samouprave

Željko Širanović, Bojan Nožica, Dunja Bjelobrk
Tehničko veleučilište Zagreb, Zagreb, Hrvatska
e-mail zsiranovic@tvz.hr, bnozica@tvz.hr, dbjelobrk@tvz.hr

Sažetak - Računalstvo u oblaku kao novi trend u informacijsko-komunikacijskim tehnologijama (ICT), nalazi široku primjenu od tzv. kućne pa do poslovne i javne primjene, kako u užoj tako i u široj zajednici. U radu se navodi nekoliko preporuka primjene tzv. oblak rješenja, s ciljem komunikacije tijela lokalne samouprave i njenih stanovnika, kao jednog novog demokratski neposrednijeg oblika komunikacije i suradnje. Primjenom specifičnih servisa koje nudi računalstvo u oblaku moguće je unaprijediti postojeće web stranice lokalne uprave i samouprave, čime se osim poboljšane funkcionalnosti mogu ostvariti i materijalne uštede. U radu se polazi od poslova koji su u djelokrugu lokalne uprave i samouprave, te mogućnosti primjene usluga iz računalnog oblaka, koje mogu pomoći građanima i drugim subjektima zajednice u boljem informiranju, te izravnijem i neposrednijem odlučivanju s ciljem ostvarivanja njihovih potreba.

I. UVOD

Internet je danas postao svakodnevno sredstvo komunikacije koje je značajno promijenilo način povezivanja i komunikacije ljudi. Osim toga korisnik koji je preko računala povezan sa Internetom ima mogućnost koristiti i druge oblike elektroničkih usluga i servisa, poznate pod nazivom računalstvo u oblaku. Značajan dio usluga i servisa iz računalstva u oblaku može se implementirati i primijeniti kroz postojeće internetske portale jedinica i upravnih tijela lokalne samouprave. Implementacija, primjena i korištenje određenih servisa računalstva u oblaku moglo bi poboljšati i unaprijediti postojeće internetske portale jedinica lokalne samouprave, a njihovim korisnicima omogućilo bi jedan novi neposredni demokratski oblik komunikacije i suradnje. Građani bi, kao korisnici u tim zajednicama, mogli postati aktivni sudionici zajednice, pri čemu bi koristi mogli imati i gospodarstveni subjekti.

Elektroničke usluge i servisi portala lokalne samouprave mogu na neki način biti pokazatelj koliko je ta zajednica informacijski razvijena. Kako su jedinice lokalne samouprave ujedno sastavni dio regije odnosno države, onda se kroz analizu implementiranih e-usluga i servisa može doći i do pokazatelja o informacijskoj razvijenosti neke regije i države. Ovi pokazatelji informacijske razvijenosti bitni su za donošenje i provedbu nacionalnih i regionalnih politika, vezanih za razvitak pojedinih javnih e-usluga i servisa potrebnih građanima i poslovnim subjektima. Razine elektroničkih usluga definiraju se kroz pet razina (skala od 1 do 5), a značenje im je [10] [11]:

- Informacija: Na Internetu je dostupna samo informacija.
- Jednosmjerna komunikacija: Dostupnost formulara u elektroničkom obliku za pohranjivanje na računalu. Prazne formulare moguće je i otisnuti na pisaču.
- Dvosmjerna komunikacija: Interaktivno ispunjavanje formulara i prijava uz autentifikaciju Ispunjavanjem formulara pokreće se pojedina usluga.
- Transakcija: Cijela je usluga dostupna na Internetu, pri čemu je omogućeno popunjavanje formulara, autentifikacija, plaćanje i isporuka potvrda, narudžbe ili drugi oblici potpune usluge putem Interneta.
- Targetizacija: Javna proaktivna ili automatizirana usluga bez potrebe da korisnici podnose zahtjev jer se potrebni podaci za realizaciju takve usluge prikupljaju po službenoj dužnosti.

Europska unija također promiče korištenje računalstva u oblaku od strane javnih tijela i tvrtki, vjerujući da se na taj način može potaknuti i rast BDP-a [1]. Kako RH treba 2013. postati dio Europske unije, onda su pokazatelji razvijenosti javnih elektroničkih usluga i servisa bitni radi donošenja strategija i politika daljnjeg razvitka i usklađivanja javnih elektroničkih servisa i usluga s onim iz EU. Politike razvitka usluga i servisa trebaju rezultirati standardizacijom Internetskih stranica tijela lokalne uprave i samouprave kako bi te usluge i servisi bili usmjereni i u potpunosti služili potrebama građana i poslovnih subjekata. U zakonskom vezano za pravo pristupa informacijama stoji da informacije koje posjeduju, raspolažu ili nadziru tijela javne vlasti, moraju biti dostupne fizičkim i pravnim osobama [14].

Polazeći s jedne strane od poslovnih funkcija koje su u djelokrugu zajednica lokalne uprave i samouprave, te s druge strane mogućnosti primjene usluga iz računalnog oblaka, u ovom radu se nastoje istaknuti neke od tih usluga koje bi se mogle implementirati kroz portale zajednica. Na taj način bi se moglo pomoći građanima i drugim subjektima zajednice u boljoj informiranosti, te izravnijem i neposrednijem odlučivanju u cilju ostvarivanja njihovih potreba.

II. DJELOKRUG RADA PODRUČNE I LOKANE UPRAVE I SAMOUPRAVE

Republika Hrvatska administrativno je podijeljena na 20 županija, 127 gradova[1] i 429 općina. Prema Ustavu RH [13] jedinice lokalne samouprave čine općine i gradovi. Županije predstavljaju više jedinice tzv. područne samouprave, ustrojene od općina i gradova [2][3]. Djelokrug rada općina, gradova i županija je reguliran njihovim statutima i Ustavom. Statutom kojeg donose općinska, gradska i županijska predstavnička tijela, podrobnije se uređuje samoupravni djelokrug, ustrojstvo, ovlasti i način rada i obavljanje poslova, kao i ostalih pitanja iz djelokruga dotične samouprave i pripadajućih upravnih službi koje su u njihovom sastavu [2]. Jedinice lokalne samouprave obavljaju poslove iz lokalnog djelokruga kojima se neposredno ostvaruju potrebe građana, a koji nisu Ustavom ili zakonom dodijeljeni državnim tijelima i to osobito poslovi koji se odnose na [2]:

- planove uređenje naselja i stanovanje
- prostorno i urbanističko planiranje u skladu sa županijskim planom
- komunalno gospodarstvo
- skrb o djeci
- socijalnu skrb
- primarnu zdravstvenu zaštitu
- dijelom na odgoj i osnovno obrazovanje
- kulturu, tjelesnu kulturu i šport
- zaštitu potrošača
- zaštitu i unapređenje prirodnog okoliša
- protupožarnu i civilnu zaštitu
- lokalni promet na svom području ne ubrajajući državne i županijske prometnice.

Veliki gradovi, kao i gradovi sjedišta županija u svom samoupravnom djelokrugu obavljaju poslove isto kao i općine, s tim da su u njihovu djelokrugu rada još i poslovi vezani na:

- izdavanje građevinskih i lokacijskih dozvola, drugih akata vezanih uz gradnju te provedbu dokumenata prostornog uređenja
- ostale poslove koji su im zakonom stavljeni u djelokrugu, s tim da mogu obavljati i neke poslove iz djelokruga županije.

Županije obavljaju poslove od regionalnog značaja koji nisu Ustavom i zakonima dodijeljene državnim tijelima, s tim da u njihovu djelokrugu rada spadaju poslovi [3]:

- gospodarskog razvoja
- planiranje i razvoj mreže obrazovnih, zdravstvenih, socijalnih i kulturnih ustanova
- izdavanje građevinskih i lokacijskih dozvola, drugih akata vezanih uz gradnju te provedbu dokumenata prostornog uređenja za područje županije izvan područja velikoga grada.

Veći gradovi, kao i gradovi sjedišta županija, mogu imati širi samoupravni djelokrug nego općine i gradovi. Uz to, veliki gradovi mogu obavljati i neke poslove iz djelokruga županije.

Specificiranje i prepoznavanje djelokruga poslova i nadležnosti lokalne uprave i samouprave bitno je radi analize, kreiranja strategije i politike razvitka, te funkcionalne standardizacije i implementacije potrebnih servisnih i aplikativnih oblak rješenja za web portale županija, općina i gradova. Primjer ne postojanja jedinstvene politike razvitka može se vidjeti već iz neujednačenosti naziva web sjedišta[2], a slično je i kod gradova i općina [4].

III. RAČUNALSTVO U OBLAKU

Računalstvo u oblaku (eng. Cloud Computing) je naziv za model i način sveobuhvatnog brzog i prilagodljivog pristupa konfiguriranim dijeljenim računalnim resursima (npr. mrežama, serverima, spremištima podataka, aplikacijama i servisima) koji se mogu brzo koristiti uz minimalno upravljanje i administriranja ili interakciju s davateljem usluge [5].

Model isporuka oblak usluga obzirom na programe, platformu i infrastrukturu, dijeli se na tri modela [6]:

- Model u kojem se za interakciju sa udaljenim servisima koriste određene instalirane desktop aplikacije ili web preglednici. Ovaj model je poznat pod nazivom „Softver kao servis“ (eng. Cloud Software as a Service - SaaS)
- Paket operacijskih i drugih softverskih servisa koji za korisnike, koji omogućuju razvoj i korištenje aplikacija smještenih u oblaku (aplikacija, baze podataka i sl.), model poznat pod nazivom „Platforma kao servis“ (eng. Cloud Platform as a Service - PaaS)
- Virtualne platforme odnosno virtualni serveri, skladišta podataka, model poznat pod nazivom „Infrastruktura kao servis“ (eng. Cloud Infrastructure as a Service - IaaS).

Aplikacije računalstva u oblaku mogu biti gotova programska rješenja opće namjene ili specijalizirane

[1] Grad Zagreb ima status kao i županija

[2] Neke web stranice županija u vršnoj domeni imaju .org ili . info (npr. www.edubrovnik.org, www.zupanija.info) umjesto .hr. Neke u poddomeni domeni imaju skraćenicu (nor. http://bbz.hr/), neke puni naziv (www.zadarska-zupanija.hr), a neke neodređeni naziv (npr. www.zupanija.info).

poslovne aplikacije poput praćenja odnosa s korisnicima. Programske platforme računalstva u oblaku omogućuju razvoj i smještaj korisničkih aplikacija u tzv. oblak okruženje [7:44-45]. Vlada RH promiče za javnu upravu informatička rješenja temeljena na otvorenom kodu, vjerujući da to može biti generator razvoja lokalne informatičke industrije [8][9][10].

Infrastruktura računalstva u oblaku zamjenjuje stvarnu korisničku fizičku infrastrukturu (poslužitelje, diskove, baze podataka, uređaje za osiguranje sigurnosti) [7]. S obzirom na to kako je razvijena infrastruktura, računarstvo u oblaku dijeli se na [5][6]:

- Privatni ili interni oblak (eng. private cloud) pripada isključivo jednoj organizaciji koja ima potpunu kontrolu nad podacima, sigurnosti i kvaliteti usluga. Resursi se ujedinjuju kroz cijelu organizaciju i obično je ovaj model prvi korak organizacije prema računarstvu u oblaku. Kada se organizacija želi „otvoriti prema vanjskom svijetu“ onda obično usluge oblaka implementira kroz virtualni privatni oblak (eng. virtual private cloud) kojem se pristupa putem sigurne veze (obično VPN) koje se mogu i naplaćivati: sam pristup može bit ograničen na fizičku lokaciju korisnika (npr. iza vatrozida).
- Javni oblak (eng. Public cloud) dobavljači usluge (eng. cloud provider) daju na raspolaganje nekom općenitom korisniku aplikacije, spremišta i ostalih resursa. Usluge se obično naplaćuju po tzv. potrošnji.
- Zajednički oblak (eng. Community cloud) partnerski dijeli nekoliko organizacija ili zajednica koje imaju određene zajedničke interese ili potrebe. Ovakvim oblacima mogu upravljati same organizacije ili netko drugi (npr. dobavljač usluga). Troškovi koji proizlaze iz ovakvog modela (npr. korištenje i održavanje), dijele se između organizacija ili zajednica. Zajednički oblak je oblik javnog oblaka koji je najčešće pod dobrim nadzorom.
- Hibridni oblak (eng. hybrid cloud) je sastavljen od dva ili više različitih privatnih, javnih ili zajedničkih oblaka. Kombinira javne i privatne modele oblaka i nudi prednosti različitih razvojnih modela. Hibridni oblak uvodi dodatnu složenost određivanja kako raspodijeliti aplikacije u javnom i privatnom oblaku.

Neovisno o modelima, za korištenje računalstva u oblaku važna je dostupnost servisa i usluga koje se na ovaj način isporučuju, kao i brzina pristupa Internetu. Isto tako treba voditi računa i o tome da neki udaljeni krajevi, npr. otočke općine, možda nemaju takav pristup ili koriste mobilni pristup Internetu. S druge strane, pouzdanost usluge ovisi o pouzdanosti davatelja oblak usluge, kao i o pouzdanosti svih internetskih veza između krajnjega korisnika i davatelja oblak usluge [7].

Prelaskom na računalstvo u oblaku moguće je osjetno uštedjeti na troškovima vezanim uz IT resurse, jer nema troškova kupovine hardvera, a troškovi softvera koji se najčešće plaćaju unaprijed puno su niži u odnosu na razvoj vlastitog rješenja ili kupnje sličnog. Uz to, troškovi održavanja ili nadogradnje su manji jer se ne mora plaćati prostor za smještaj poslužitelja niti zapošljavati osobe koje će se brinuti o njima.

Problem koji se javlja kod uvođenja računalstva u oblaku vezan je uz sigurnost i mogućnost gubitka kontrole jer su servisi i usluge smještene u javnom ili hibridnom oblaku. Odabir javnog ili hibridnog modela dovodi do toga da je veći dio aplikacija, mreže, računalnih sustava i podataka pod kontrolom trećih osoba davatelja usluga računalstva u oblaku.

IV. MOGUĆNOSTI PRIMJENE RAČUNALSTVA U OBLAKU U LOKALNOJ SAMOUPRAVI

Kada se govori o razini informatiziranosti jedinica lokalne i područne uprave i samouprave, ne postoji niti jedna jedinica koja nema vlastite web stranice. Općine, gradovi i županije, kao i njihovi pojedini upravni odjeli, na svojim web stranicama uglavnom imaju dostupne informacije o svojim uslugama (npr. razni pravilnici, obavijesti i sl.). Ovo se na skali mjerila praćenja informatiziranosti koju je utvrdila Europska komisija smatra prvom razinom informatiziranosti [10:15][11].

Sljedeća razina informatiziranosti predstavlja razinu jednosmjerne komunikacije [10:15][11]. Ova razina omogućuje građanima i privrednim subjektima preuzimanje dostupnih formulara u elektroničkom obliku, koji se mogu preuzeti, pohraniti lokalno na računalu korisnika, a kasnije ih otisnuti na pisaču. U prvom slučaju ne može se govoriti o nekom posebnom rješenju ili servisu računalstva u oblaku. Ovdje se praktično radi o informacijskim sadržajima koje se realiziraju klasičnim alatima za razvitak web sadržaja, pri čemu se određeni dijelovi informacija (npr. obrasci) mogu preuzeti i lokalno pohraniti na korisnikovo računalo.

Primjena aplikacija i servisa računalstva u oblaku prema skali informatiziranosti pripadali bi u višu razinu informatiziranosti [10:15][11]:

- jednosmjerna komunikacije
- dvosmjernu komunikaciju
- transakcijske usluge
- automatizirano proaktivna usluga u potpunosti prilagođena krajnjem korisniku.

Transakcijske i automatizirane usluge uglavnom se odnose na određene servise, obrasce i potvrde za koje se obično vežu autentifikacijski servisi računalstva u oblaku. Ovaj oblik informatizacije više je vezan uz realizaciju javnih servisa i usluga e-uprave, stoga nije u fokusu ovog rada.

Ovaj rad fokusiran je na usluge i servise računalstva u oblaku koji se zasnivaju na jednosmjernoj i dvosmjernoj komunikaciji koja bi se mogla relativno brzo i lako implementirati u postojeće portale lokalne i područne samouprave i njihovih upravnih tijela. S druge strane, očekuje se da bi efekt implementiranih usluga i servisa bio vrlo brzo vidljiv i prepoznat od strane zainteresiranih subjekata u zajednici. Građani bi imali koristi u smislu bolje informiranosti, izravnijem i neposrednijem odlučivanju i ostvarivanju potreba.

Uz računalstvo u oblaku moguća je osjetna ušteda troškova IT resursa, jer nema troškova kupovine hardvera, a troškovi softvera koji se najčešće plaćaju unaprijed puno su niži nego kada se razvija vlastito rješenje ili kupuje gotov proizvod. Pri tome su troškovi održavanja ili nadogradnje manji, a uz to nije potrebno dodatno plaćati prostor za smještaj poslužitelja, niti zapošljavati osobe koje će se brinuti o njima.

V. SERVISI JEDNOSMJERNE KOMUNIKACIJE

Jednosmjerna komunikacija najčešće obuhvaća online novosti i razne jednostavne online obrasce ili upitnike. Gotovo svaka jedinica lokalne samouprave posjeduje vlastite web stranice na kojima se mogu pronaći najnovije informacije vezane za rad tog lokalnog tijela. Uglavnom su to birokratsko-administrativne informacije koje se odnose na zaključke, odluke ili akte koje donose nadležna tijela lokalnih vijeća ili skupština.

Web stranice jedinica područne i lokalne samouprave trebale bi sadržavati informacije ili poveznice na druge privatne, javne i zajedničke oblake. Odnosno na one koji su od lokalnog značaja i u okviru samoupravnog djelokruga obavljaju ili neposredno ostvaruju potrebe građana i poslovnih subjekata. Ove vijesti obično se odnose na: uređenje naselja i prostorno-urbanističko planiranje, komunalno gospodarenje, skrb o djeci i socijalnu skrb, primarnu zdravstvenu zaštitu, obrazovanje, kulturu i sport, zaštitu potrošača, zaštitu i unapređenje prirodnog okoliša, civilnu i protupožarnu zaštitu, i sl. Web stranica lokalne samouprave trebala bi biti servis koji na pregledan i inteligentan način nudi poveznice prema drugim web stranicama, preko kojih se može doći do korisnih i potrebnih informacija npr. informacije o rasporedu rada liječnika primarne zdravstvene zaštite, komunalnog poduzeća, kulturnih i sportskih udruga i društava, koji djeluju u toj zajednici, upravnih i drugih tijela koja djeluju u zajednici.

S druge strane, online obrasci se mogu iskoristiti kao servisi preko kojih korisnici mogu slati određene informacije. Web obrasci se koriste gotovo na isti način kao i svaka druga web stranica. Ono što oni posjeduju su dodatni interaktivni elementi kao što su: polja za unos informacija, polja za izbor, akcijski gumbi i sl. Obrasce mogu koristiti građani i drugi zainteresirani subjekti kako bi izrazili svoja mišljenja, želja, sugestija, pohvale, prijedloge, kritike i sl. Obrasci se mogu koristiti i u slučaju prikupljanja mišljenja kroz anketiranje o nekoj određenoj temi. To mogu bit vrlo jednostavni obrasci koji odgovaraju na pitanja koja traže odgovor u obliku da/ne pa sve do složenijih pitanja. Uz svaku anketu moguće je postaviti mogućnost uvida u trenutno stanje glasovanja ili ankete.

Organizacija različitih događaja unutar zajednice, kao što su npr. kulturni ili športski događaji, mogu se jednostavno organizirati pomoću web baziranih kalendara. Najčešće se za ovakvo informiranje o događanjima unutar zajednice koriste plakati koji se postavljaju na oglasne ploče, nakon čega obično slijedi usmeno prenošenje informacija, a ponekad se koristi i elektronička pošta. Problem s ovakvim načinima informiranja je u tome što su plakati i obavijesti obično smještene samo na određenim mjestima[3] i na ovaj način ne mora svatko vidjeti te obavijesti ili eventualno dobiti informaciju od nekoga drugog. Problem kod usmenog prenošenja informacije je veliki entropijski učinak zbog kojeg informacija gubi vrijednost. Ovaj problem se jednostavno rješava korištenjem web baziranih kalendara. Ono što se može napraviti je kreiranje javnog korisničkog računa na nekom od servisa iz oblaka. Svi događaji koji se naprave na kalendaru mogu biti dostupni svim članovima zajednice. Nakon toga treba samo proširiti informaciju tako da svatko u zajednici zna za web bazirani kalendar (npr. Yahoo! Ili Google Calendar). Na taj način su sve informacije o bilo kojem relevantnom događaju u zajednici dostupne svima. U slučaju da dođe do bilo kakvih promjena, kao što je otkazivanje nekog događaja, pomak termina događaja i slično, te promjene se jednostavno rješavaju preko web baziranog kalendara. Bilo koja promjena se jednostavno zabilježi na kalendaru i svatko tko pogleda kalendar može ju odmah vidjeti. Prednosti ovakvih web baziranih kalendara su:

- mogućnost slanja alarma koji šalje obavijest pretplatnicima na e-poštu, mobilni uređaj ili messenger
- sinkronizacija kalendara (npr. Microsoft Outlook, CalDav i sl.), također i na raznim mobilnim uređajima koji imaju sinkronizacijske mogućnosti
- dijeljenje rasporeda između korisnika, uz mogućnost različitih načina i metoda dijeljenja kalendara (javno dijeljenje kalendara, dijeljenja kalendara s određenim grupama i sl.)
- mogućnost dodavanja više kalendara kojima može pristupati više korisnika
- kalendaru se mogu dodati različite razine pristupa, tako da im mogu pristupati samo određeni korisnici.

[3] U nekim slučajevima samo na oglasnoj ploči u općini

VI. SERVISI DVOSMJERNE KOMUNIKACIJE

Kako bi informacije brzo i jednostavno dolazile do građana u zajednici, bitna je adekvatna komunikacija. Svaki put kada bilo koja grupa ljudi u zajednici ima potrebu za međusobnom komunikacijom i suradnjom, web bazirane aplikacije mogu biti dobar način da se kroz komunikaciju aktiviraju građani i potakne njihovo kreativno razmišljanje. Slanjem i primanjem kreativnih informacija, lokalna zajednica kao cjelina može poboljšati reagiranja i dati nove poslovne prilike kao i načine rješavanja pojedinih lokalnih problema koji se mogu pojaviti u zajednici. Lokalne vlasti na taj način mogu postati dobri upravitelji koji brinu o interesima zajednice, a ne da budu statični birokratski gospodari koji pune proračun jedinice lokalne samouprave.

Jedno od tipičnih servisa računalstva iz oblaka koje se može svrstati u kategoriju dvosmjerne komunikacije je elektronička pošta. Ovim servisom znatno se može poboljšati komunikacija prema tijelima lokalne samouprave, lokalnim ispostavama državnih upravnih tijela, odnosno općenito u lokalnoj društvenoj zajednici. Adekvatan način primjene elektroničke pošte može dati velike prednosti za cijelu zajednicu, susjedstvo, sportske timove i edukacijske ustanove.

Korištenje tradicionalnih aplikacija elektroničke pošte (npr. Microsoft Outlook i Windows Mail i sl.), može biti problematično ako se ove aplikacije nalaze na fiksnom računalu u uredu ili u kući. Korisnik ne mora uvijek biti za svojim računalom. Bolje rješenje za komunikaciju putem elektroničke pošte u zajednici je korištenje web baziranih e-mail aplikacija (npr. Gmail, Microsoft Hotmail ili Yahoo Mail). Web e-mail baziranim aplikacijama, koje su u potpunosti riješene kao servis računalstva u oblaku, moguće je pristupati s bilo kojeg računala ili pametnog telefona, ukoliko postoji veza na Internet. Pomoću web preglednika moguće je provjeriti i slati e-mail poruke koje se spremaju unutar oblaka, jednako kao da se koriste tradicionalne desktop aplikacije za elektroničku poštu. Elektroničke poruke preko web baziranih e-mail aplikacija iz računalnog oblaka mogu se primiti i slati od kuće, posla ili bilo kojeg drugog mjesta. Neke zajednice čak naprave posebne web bazirane e-mail račune samo za komunikaciju sa zajednicom [12:62].

Jedan od korisnih servisa računalstva u oblaku koji može poslužiti za dvosmjernu komunikaciju u zajednici su internetski forumi. Sve poruke koje korisnik napiše i pošalje na forum vidljive su svim ostalim sudionicima foruma, pri čemu sve nalikuje na oglasnu ploču na kojoj sudionici ostavljaju svoje poruke. Zbog lakšeg snalaženja forum je obično podijeljen u nekoliko skupina prema aktualnim temama. Poruka poslana na jedan forum vidljiva je samo na tom forumu. Forum se uglavnom sastoji od više kategorija unutar kojih se nalaze teme koje otvaraju i započinju korisnici. Svaka tema ima svoju diskusiju u kojoj se sudjeluje pisanjem poruka, odnosno objavljivanjem postova. Svi razgovori i diskusije ostaju sačuvani na forumu dok ih moderator ne obriše. Također postoji mogućnost dodavanja prava korisnicima forumu, ograničavanje čitanja pojedinih ili svih diskusija i sl.

Pomoću programa za vođenje diskusije moguće je ostvariti bolju povezanost građana u lokalnoj zajednici čak do razine mjesnog odbora kao i olakšati rad unutar mjesnih odbora. Jednostavno i lagano se mogu dogovoriti termini skupština, međusobno dijeliti različiti podatci (različite vrste izvještaja, statistike itd.) i držati diskusije oko različitih problema unutar zajednice. U tim diskusijama mogu sudjelovati i građani. Ono o čemu treba voditi računa je da se na forumima, diskusijama i socijalnim grupama mogu susresti razni ljudi dobrih i loših namjera, zbog čega bi ovakav model komunikacije morao biti kontroliran.

Regionalna i lokalna samoupravna tijela generiraju mnogo izvješća i drugih dokumenata koja često moraju sastavljati i pisati u suradnji sa drugima. Dobro rješenje za suradnju na izvješću su servisi online kolaboracije. Umjesto slanja Microsoft Word dokumenta drugim suradnicima (npr. županijskim, gradskim ili općinskim vijećnicima) putem elektroničke pošte, mogu se koristiti web bazirani programi za obradu teksta iz računalnog oblaka. Većina tih aplikacija iz oblaka koje se koriste za obradu teksta dozvoljavaju ugradnju slika, tablica i grafova. Svaki takav servis omogućuje da svatko tko doprinosi izvješću, a ima za to ovlasti, može pristupiti oblaku, može nešto dopuniti ili promijeniti u dokumentu, pri čemu ostali sudionici mogu tu promjenu odmah vidjeti i eventualno komentirati. Rezultat je kvalitetan dokument koji odražava istinsku prirodu suradnje na projektu.

Jedinice lokalne i regionalne samouprave, odnosno njihova upravna tijela, sve više rade na raznim projektima. Većina tijela lokalne ili područne samouprave ima barem jedan veći projekt na kojem radi. Projekti ove vrste sastoje se od pojedinih manjih dijelova, od kojih svaki ovisi o završetku prethodnih zadataka. Praćenje svih individualnih zadataka, trošenje financijskih sredstava, tijek i dinamika projekta može biti vrlo važna informacija, ne samo za građane u toj zajednici, već i za privredne subjekte. Za vođenje i upravljanje projektima, kao i za vođenje i upravljanje projektnom dokumentacijom, mogu se koristiti za to pogodni alati iz računalnog oblaka. Svaki član projektnog tima može individualno raditi na svom djelu izvješća. Nakon što su svi individualni dijelovi izvješća gotovi, voditelj tima i ostali zainteresirani subjekti mogu pogledati cijelo izvješće ili pojedine dijelove. Često puta dinamika tih projekata može utjecati i na poslovne odluke privrednika u toj zajednici (npr. informacija o dovršenju određene komunalne infrastrukture). Praćenje projekta i dinamika izvještavanja zainteresiranih subjekata u zajednici, može se pojednostaviti uz korištenje web baziranih aplikacija za upravljanje projektima. S druge strane, članovi projekta mogu pristupiti aplikaciji s bilo kojeg mjesta i time dobili pristup određenim datotekama projekta. Ovlaštene

osobe potom mogu dodavati i ažurirati projektne zadatke, označiti zadatke kao završene, dodavati i objaviti informacije o naplati za pojedine zadatke i sl. Budući da je aplikacija za vođenje projekata u računalnom oblaku, svaki član tima vidi isti gantogram i isti popis zadataka. Podatci unutar datoteke odmah se osvježe čim neki član projektnog tima napravi promjenu [12:74]. Mnoge aplikacije za upravljanje projektima imaju dodatne funkcije kao što su web bazirano dijeljenje datoteka, oglasne ploče, praćenje vremena i troška projekta i sl., tako da uvid u projekt mogu imati legalno izabrani predstavnici građana, a po potrebi i građani.

VII. ZAKLJUČAK

Računalstvo u oblaku podrazumijeva niz internetskih ili intranetskih servisa, aplikacija i usluga koji mogu imati široku primjenu. Web stranice općina, gradova i županija trebale bi pružiti ili biti poveznica prema široj paleti elektroničkih usluga i servisa koji bi na jednom mjestu mogli zadovoljiti građane i druge zainteresirane korisnike. Vlasti lokalne uprave i samouprave mogu imati koristi od internetskih servisa računalstva u oblaku jer preko njih mogu još više poboljšati i proširiti informacijsku učinkovitosti i pružiti usluge svojim građanima.

Lokalnu zajednicu bi trebalo još više približiti građanima, ostvariti bolji kontakt sa zainteresiranim subjektima zajednice. To se može postići ako se prihvati internetski način komuniciranja i iskoriste mogućnosti koje nude suvremene informacijsko komunikacijske tehnologije. Ovakav način komunikacije može osnažiti suradnju i poslovanje lokalne zajednice.

Primjena aplikacija i servisa računalstva u oblaku u zajednici mogla bi omogućiti da što veći broj građana sudjeluje u različitim raspravama i na taj način podjeli svoje mišljene o stanju u zajednici i eventualno ponudi potencijalna rješenja.

Usluge i servisi računalstva u oblaku koji se zasnivaju na jednosmjernoj i dvosmjernoj komunikaciji, mogli bi se relativno lako i brzo implementirati u postojeće web portale lokalne i područne samouprave i njihovih upravnih tijela. Građani bi imali koristi u smislu bolje informiranosti, izravnijem i neposrednijem odlučivanju i ostvarivanju potreba. Implementirane usluge i servisi mogli bi biti prepoznati i prihvaćeni od strane građana i gospodarstvenika, te ostalih zainteresiranih subjekata u zajednici.

Građani lokalne zajednice mogli bi prihvatiti usluge i servise koje im nudi lokalna samouprava i uprava na svojim web stranicama i na taj način prepoznati onu zajednicu koja im nastoji pomoći. Prihvaćanjem i uvođenjem ovakvih novih modela komunikacije i suradnje u zajednici pruža se mogućnost okretanja građanima biračima, a ne birokraciji.

VIII. LITERATURA

[1] Brebrić, K., BUGONLINE, Europska unija potiče računalstvo u oblaku

http://www.bug.hr/vijesti/europska-unija-potice-racunalstvo-oblaku/118777.aspx (18.1.2013.)

[2] Wikipedija, Hrvatske općine i gradovi http://hr.wikipedia.org/wiki/Hrvatske_op%C4%87ine_i_gradovi (18.1.2013.)

[3] Wikipedija,Hrvatske županije

http://hr.wikipedia.org/wiki/Hrvatske_%C5%BEupanije (18.1.2013.)

[4] http://www.vlada.hr/hr/adresar_i_linkovi (21.1. 2013.)

[5] Mell, P., Grance, T., National Institute of Science and Technology, The NIST Definition of Cloud Computing

http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf. SP 800-145. (19.1.2013.)

[6] Wikipedia, Cloud computing

http://en.wikipedia.org/wiki/Cloud_computing. (15.1.2013.)

[7] Bartolinčić, N., Sigurnost u oblaku, Infotrend 183., 2011.

[8] Narodne novine, Strategija »Informacijska i komunikacijska tehnologija – Hrvatska u 21. stoljeću«. NN 109/2002.

[9]http://www.vlada.hr/hr/dodatno/o_portalu_vlade_rh/upoznajte_novi_portal_vlade_rh/otvoreni_kod_i_otvorenost (22.1.2013.)

[10.] Ministarsvo gospodarstva, Strategija razvoja elektroničke uprave u Republici Hrvatskoj za razdoblje od 2009. do 2012. godine

http://www.mingo.hr/default.aspx?ID=3330. (22.1.2013.)

[11] Vračić, T., Infotrend.hr – onLINE, Dostupnost elektroničkih javnih usluga na Internetu u Europskoj uniji u 2012. godini

http://www.infotrend.hr/files/pdf/casopis/2012/191/InfoTrend-191.pdf. (21.1.2013.)

[12] Miller, M., Cloud Computing Web-Based Applications That Change the Way You Work and Collaborate Online, Que Publishing, 2009.

[13] Narodne novine, Ustav Republike Hrvatske, NN 85/10

http://www.sabor.hr/Default.aspx?art=1891 (21.1.2013.)

[14] Narodne novine, Zakon o pravu na pristup informacijama, NN 172/03

Regulating IT projects and information systems audit in public and state administration of the Republic of Croatia

Biljana Cerin, dipl.ing.
doc. dr. sc. Goran Vojković

Ostendo Consulting Ltd.
London, UK
E-mail: biljana.cerin@ostendogroup.com

Centar za razvoj unutarnje plovidbe d.o.o. (CRUP d.o.o.)
Nova cesta 180, Zagreb, Republika Hrvatska
E-mail: vojkovic@crup.hr

Abstract – According to the State budget data, around a billion kunas (131.763.000 €) yearly is spent for IT projects in public and state administration in Croatia. Their success is of extreme importance for general development and fulfilment of state goals. Therefore, it is necessary to regulate the areas of preparation, monitoring and management of IT projects through legislation, and define the role of information systems audit in public and state administration, in order to ensure efficiency of investments in IT projects, avoid „double" or irrational procurements, supervise fulfilment of the procurement objectives, minimise dependence upon single IT solutions and services supplier, and avoid delivery of half-solutions which are later upgraded on the account of the customer. This paper will present the need to systematically regulate preparation, execution, management and supervision of IT projects in public and state administration, as well as the information systems audit, in order to ensure clear and transparent management of IT projects and enhance control over their realisation as well as over the quality of IT solutions implemented. Regulation of these questions would result in better use of resources intended for computerization of public and state administration, better support to the citizens and significant savings in areas of IT investments and maintenance of IT systems, as well as increased security of data and information systems themselves.

I. INTRODUCTION

General condition with IT projects in public administration in Republic of Croatia was clearly explained by Darko Parić, Assistant Minister of Public Administration for e-Hrvatska: „Condition of IT in the state is worse than we thought. There is no control or strategy and no one is in charge of a billion kunas that is allocated to all IT projects via state budget. Computerization projects are implemented throughout all ministries while it is not taken care of is there an overlap. Ministries behave as fortresses and complete the projects for themselves exclusively [1]."

Such enormous investments are not visible in practice – state and public administration of RoC has achieved something one would call initial computerization (using computers as sophisticated typing machine, using databases in electronic version instead o fin paper archives). However when electronic mail is mentioned, it has not become a form of official administrative communication, despite regulations of the General Administrative Procedure Act [2] and other regulations. One still communicates officially via classis, paper method with a signature and stamp by an authorized body.

Furthermore, interoperability for data is achieved in enormously small amount and in a few segments only (such as enrolments to schools and faculties). In practice, bases of similar data often overlap, even those with identical data should they be run by two or more various subjects. As stated by Darko Parić, Assistant Minister of Administration, "Every citizen is listed at least 16 times in various subjects of public administration with his basic data [3]."

Business practice of IT-oriented suppliers looked upon state and public administration as they were large source of money that orders a lot and asks very rarely in general. That was prior to the recession in 2009-2010. All negativities such as addiction to a particular supplier, non-standard solutions and purchase of non-tested programmes have become widely spread in the system of state and public administration.

Economic crisis as well as new understanding of state IT that is visible over the past year, opened two generally opposite demands:

a) It is necessary to ensure investment savings within state IT sector, due to general needs for rationalization.
b) It is necessary to increase efficiency of state IT sector, including investments, in order to have them as "engines" of stepping out of economical crisis.

These two conditions are opposite just on a first sight – because both regulation of prepare, realization, management and supervision of IT projects in state and public administration and revision of computerization systems, that are subject of this paper, may help in achieving both goals.

We would like to mention that Ministry of Public Administration of Republic of Croatia will be mentioned several times in this paper. The reason for such is that, according to present structure, the Ministry performs administrative and expert tasks that refer to development of computerization system of state administration and rationalization of IT resources in public administration bodies [4].

II. PROJECT PREPARATION

Most often, project preparation within system of state and public administration is not fruit of systematic consideration and particular methodologies, but is run due to the idea of "working things out" or "speed up on subject", however without honest intention for system reform and business administration processes in particular. Computerization is important, but itself cannot solve issues beyond its control such as usually unsettled business processes.

Another, even worse impulse for IT projects' start up, is "just to spend money" meaning not to "lose" money from budget. In such cases usually over dimensioned and useless IT solutions are being purchased.

All this goes into favour of importance of proper preparation of projects in state and public administration. How to make it?

Republic of Croatia does not have to develop some new methodology for preparation and leading of projects at all. All it has to do is to use an existing and good solution that European Union runs for its projects, which is Project Cycle Management (further: PCM) [5]. This methodology, that is actively used by European Commission for its projects and is generally used in EU-financed programmes, pays a great attention to project preparation phase and includes asset approve for project financing.

PCM includes Logical Framework Approach (LFA) that was originally developed in late 60's as assistance to the American Association for International Development to improve planning system and project evaluation. LFA was brought in due to the planning being too unclear and without clear defined goals that would be used for supervision and evaluation of efficiency (or non-efficiency) of a project.

LFA is a core tool used within Project Cycle Management.

• It is used during the identification stage of PCM to help analyse the existing situation, investigate the relevance of the proposed project and identify potential objectives and strategies;

• During the formulation stage, the LFA supports the preparation of an appropriate project plan with clear objectives, measurable results, a risk management strategy and defined levels of management responsibility [6].

It is exactly what we need in Croatian state and public administration – fully developed, checked and available methodology that may easily be implemented in practice.

When a concrete need and concrete project were identified for funding, via use of PCM and LFA; one essential area that is a part of project preparation phase – and that is bidding documentation preparation, in accordance with public procurement regulations.

Republic of Croatia has a modern and quality Public Procurement Act that is in force as of January 1, 2012 [7]. Current law on Public Procurement Act, more than previous regulations, provides public customer choice of a quality performer and generally clear explanation of quality of delivery of goods, works and services as conditions that are to be met by a bidder. According to Article 25, Para 1 of the Act, "public customer freely chooses between open and limited procedure of public procurement" (*open procedure* of public procurement is when any interested party may submit an offer while *limited* one is *procedure* where any interested party may ask for participation in procedure while only those recognized and approved by the customer may submit an offer).

Also, public customer may rank offers according to two criteria, naturally one must previously be determined for implementation: economically the most favourable offer and the lowest price. Economically the most favourable offer may include various criteria linked to the subject of procurement, such as quality, price, technical advantages, esthetical and functional characteristics, operative expenses, post-sale service and technical assistance, delivery date or deadline and realization deadline.

Existing methodology implied by European Union PCM (in national frames to be implemented with small changes, *cum grano salis*), with existing Public Procurement Act are a good frame for quality prepare of project, its financial approve and prepare of public procurement procedure for its realization.

Over the past several years Croatia is implementing projects in accordance with EU methodology (PHARRE, IPA...). Also, for system of public procurement provides certain education programme [8]. Taken into consideration, one would conclude that only a small effort from responsible bodies, firstly Ministry of Public Administration, in the field of adjustment and larger implementation of existing knowledge – in order to put state and public administration project on higher level. Minimum investments would pay off in the very first year.

III. PROJECT MANAGEMENT

A project can be considered as achievement of a specific objective, which involves a series of activities and tasks which consume resources. It has to be completed within a set specification, having definite start and end dates. In contrast, project management can be defined as the process of controlling the achievement of the project objectives. Utilising the existing organisational structures and resources, it seeks to manage the project by applying a collection of tools and techniques, without adversely disturbing the routine operation of the company [9].

Problem with managing projects may be represented by classic pyramidal administrative hierarchy where each person performs strictly defined job description. Such hierarchy is typical for state administration but also for the most of public administration.

Management of computerization project cannot successfully be performed in a classic administrative way, where a project manager, in the most part of state administration, would be "hidden" within IT department that is hierarchically under the secretarial cabinet. Internal

communication would be a classic one: Head of Department – Secretary to the Organization – Manager-in –charge of other hierarchy unit – person whose name is listed. It would be too slow and totally unadjusted to project management.

How to deal with it? There are two ways: one would be to define role of project manager in regulations that define work and internal management communication, while the other would be to give authorities to project manager to directly communicate via decision of Head of organization, Minister or Chief. The other way is an easy one, quick and does not require any particular adjustments. However, in order to run the project successfully, one must give authorities to project leader, because in classic administrative hierarchy, direct communication and horizontal leadership have never existed.

When executing EU funded or World Bank funded IT projects in public and state administration, project management organization is in most cases required to be established formally. Both the beneficiary and the supplier appoint project managers who collaborate closely during the entire project execution, perform regular reporting, establish clearly defined communication lines and escalation levels, document important decisions made on project meetings and continuously monitor project risks in order to timely implement risk management decisions. Independent third party supervision also plays an active role in the project, as well as independent third party quality assurance. This is the model that should be applied to management of IT projects in public and state administration regardless of the source of their funding, given the certain level of complexity of impact of the project that should require establishment of formal project management mechanisms. Implementing this approach to IT projects in public and state administration, although primarily established due to regular audits of the status of cost, schedule and achievement of project objectives by EU or World Bank bodies, would increase their efficiency.

However, in order to perform such activities in accurate manner, systematic education of project managers in public and state administration is required. It is not sufficient to simply appoint project managers and give them the responsibility for managing projects, without ensuring they have proper previous experience and level of education obtained. For example, in order to obtain a widely recognized professional certificate in project management, PMP (Project Management Professional), often sought for proposed key experts responsible for managing the project or the supplier's team leadership in the public tenders, the candidate has to satisfy the following educational background, project management experience and project management education requirements [10]:

- secondary degree, minimum five years unique non-overlapping professional project management experience during which at least 7,500 hours were spent leading and directing the project, 35 contact hours of formal education, or
- four-year degree, minimum three years unique non-overlapping professional project management experience during which at least 4,500 hours were spent leading and directing the project, and 35 contact hours of formal education.

Setting up the exact requirements for the project managers in public and state administration is a task that must be performed by responsible legal and regulatory bodies, taking into account experiences and trends from the professional project management community. Systematic approach to appointment of capable project managers would certainly demand additional financial and other resources to be reserved by the public and state administration bodies, but comparing the cost with the risk of failed and poorly implemented projects due to improper project management, the investment can be easily justified.

IV. PROJECT RISK MANAGEMENT

In order to ensure project management will be performed effectively, organisations need to have project management methodology defined, personnel pursuing project management responsibilities educated and experienced, and project management recognized as a necessary tool for obtaining the organizational goals. One of the most critical areas of project management is the management of project risks, which are an inherent part of every project. Only by timely recognizing, reporting, analysing, evaluating and making sound decisions on how to manage the project risks, the project objectives will be met. Risk conditions could include aspects of the project environment that may contribute to project risk such as poor project management practices, or dependency on external participants that cannot be controlled [11].

A necessary prerequisite for effective IT project risk management in public and state administration bodies is to develop appropriate risk awareness and risk management culture. This requirement for a proper risk management culture establishment is normally imposed by regulations in insurance and banking industry sectors, such as Basel II and upcoming Solvency II regulations, as a starting point for development of enterprise-wide risk management systems. Risk management culture in organization can be achieved by adopting formal policies, processes and procedures defining the framework for risk management system, organizational roles and responsibilities, supported by planned educational and awareness efforts, where all employees should recognize the role they have in identification, reporting, analysing and assessment of risks they may face in their field of work. The same approach is utilised in establishment of high risk awareness culture within the project team, with goal to enable each team member to identify and communicate the project risks he is aware of in executing the project tasks assigned to him.

Considering the formalised environment for projects in public and state administration bodies, best practice approach to establishing a high project risk management culture will be by clearly documenting the project risk management plan within the project management documentation and update it regularly with new risks reported and decisions made on their treatment in accordance with the organization's risk tolerance limits. The crucial point in this process is the reporting of risks, because project manager has to support the team members to state their concerns without fear of being dismissed as irrelevant, in conflict with some other initiatives or not taken seriously.

To overcome such problems, project managers together with the project team have to work on establishing effective communication mechanisms and ensure positive relations between the team members, external stakeholders and internal upper and lower hierarchical levels. Poor communication is the reason most IT projects fail, according to a Web poll released by the Computing Technology Industry Association (CompTIA). Nearly 28% of the more than 1,000 respondents to the survey said poor communication is the number one cause of project failure [12].

V. PROJECT MONITORING

When talking about supervision of IT projects, state and public administration know one model of supervision very little implemented in Croatian practice. However its range and consequences could be much larger and more efficient. We are talking about administrative supervision. In his classic paper "Administration Law" Borković says: "Among types of control done over administration, control within administration is pointed out (administration control). To organize such control means to regulate relationship between administration and authorities of higher and lower positioned bodies within it. The essence of such control is that it does not leave the frame of administration structure and is done in a way that one administrative body supervises the other one [13]."

Administration supervision is not based on idea of some internal administration police that shows up every once in a while should some difficulties in work occur [14]. Sometimes the term "supervision" is a cause for misunderstanding of administrative board as a model of control for "wrong procedure", even though the goal of it should be something different. More than 50 years ago Eugen Pusić defined it as follows: "Usually responsibility bears certain negative connotation of some gap made by the responsible party. On the contrary, control that follows continuous check up on successfulness of performance administrative programs, under changeable circumstances, does not have any implication of guilt, non-direct in any way from the side of those who were responsible for the implementation [15]."

The existing institute of administrative supervision is the way to control quality of implementation of IT projects, not in the negative meaning of "control" but in the way of help, assistance, consulting and addressing persons in charge of project implementation. Additionally, central bodies, precisely Ministry of Public Administration, may – should the shortage in personnel require so – engage external experts in accordance with their specialty. Those external experts would perform concrete supervisions for concrete, limited period of project and would not require increase of number of public servants. By such, along with existing positive regulations and resources, quality supervision over projects would be achieved.

VI. INFORMATION SYSTEMS AUDIT

Revision in state and public administration is regulated by the State Office for Revision Act [16]. According the Act, state income and expenses, financial reports and financial transactions of local self-governmental units, public sector units, legal persons financed by the budget, legal persons established by the Republic of Croatia or local (regional) self-government unit, units of local and regional self-government that have majority in stocks and shares and using assets of European Union and other international organizations or institutions for financing public needs – are all subjects to the state revision.

Revision is, according to the State Office for Revision Act, questioning documents, documentation, reports, systems of internal control and internal revision, accounting and financial procedures and other records that present financial reports and results of financial activities in accordance with accepted accounting regulations and standards. Revision is procedure of check up on financial transactions towards legal using of assets.

According to the strict regulation of the Act, revision covers estimating and marking on efficiency and economical activity performance and estimating efficiency of business goals or goals of individual financial transactions, programmes and projects [17].

So there is no legal obstacle for State Office for Revision to take over business of a revision of IT projects in state and public administration. An obstacle may be a factual one – which refers to skilled and trained state revisers and personnel team of the Office. State revisers should definitively be educated on performing such type of revision, because up-to-date revision was most usually account one and financial one. Such would assist in achieving revision of IT projects within existing revision system.

In the case of lack of skilled staff and extremely complex projects at the same time, State Office for Revision should contract external experts with proper international certificates.

One of the most widely accepted and recognized international professional certification for information systems audit is CISA, Certified Information Systems Auditor, administered by ISACA. ISACA is an independent, non-profit, global association, ISACA engages in the development, adoption and use of globally accepted, industry-leading knowledge and practices for information systems. In order to become a CISA, applicant has to fulfil certification requirements in terms of passing the CISA exam, adhering to the ISACA code of professional ethics and proving relevant professional experience in the following domains [18]:

- The Process of Auditing Information Systems,
- Governance and Management of IT,
- Information Systems Acquisition, Development and Implementation,
- Information Systems Operations, Maintenance and Support,
- Protection of Information Assets.

Incorporating requirement for information system audit function within organizations, who perform regular, risk based audits of information systems, is also a regulatory requirement already established for financial institutions. This function may be outsourced to external third parties who obtain approval by the relevant supervisory body for performing such service, or educated personnel can be employed within the organization for this function. The

information system auditor should be independent of the area under the review and base the findings on objective evidence. Having information system audits systematically undertaken for public and state administration information systems would enable timely recognition, reporting and management of weaknesses in information system controls' design and implementation. The situation in public and state administration is that existing internal auditors in most cases do not have specific IT audit knowledge, but, internal auditors can help IT departments address on-going business changes and growth by providing recommendations that enhance key areas of the IT project management life cycle [19].

Enhancing public and state bodies' internal auditors with specific IT audit knowledge, appointing dedicated IT auditors or establishing this function within organization, mandating it with relevant laws and regulations, will minimise the risk of over sizing or under sizing information system controls (and other resources) procured together with the new information systems.

VII. INFORMATION SYSTEMS GOVERNANCE

Successful preparation, execution, management and monitoring of IT projects in public and state administration require well established information system governance principles. Effective information system governance enables an organization to maintain high quality information to support decision making process, generate value from the IT investments, achieve operational excellence through efficient application of information technology, maintain IT related risk at an acceptable level and optimise the cost of information systems. It also enables the organization to comply with relevant laws, regulations, contractual agreements and policies. A widely accepted information systems governance framework in organizations worldwide is COBIT framework. COBIT has evolved from an audit framework in 1996 to a governance and management of enterprise IT framework in 2012.

COBIT 5 defines a set of enablers to support the implementation of a comprehensive governance and management system for enterprise IT [20]. Enablers are broadly defined as anything that can help to achieve the objectives of the enterprise. The COBIT 5 framework defines seven categories of enablers:

1. Principles, policies and frameworks,
2. Processes,
3. Organisational structures,
4. Culture, ethics and behaviour of individuals and of the enterprise,
5. Information,
6. Services, infrastructure and applications,
7. People, skills and competencies.

The COBIT 5 process reference model subdivides the governance and management processes of enterprise IT into two main areas of activity - governance and management. Each domain contains a number of processes. Although most of the processes require ‘planning’, ‘implementation’, ‘execution’ and ‘monitoring’ activities within the process, they are placed in domains in line with what is generally the most relevant area of activity regarding IT at the enterprise level [21]:

- To evaluate, direct and monitor,
- To align, plan and organize,
- To build, acquire and implement,
- To deliver, service and support,
- To monitor, evaluate and assess.

COBIT 5 also introduces the goals-cascade principle, which allows the definition of priorities for implementation, improvement and assurance of governance of enterprise IT based on strategic objectives of the enterprise and the related risk. This overarching framework covers in essence all the activities necessary for preparation, execution, management and monitoring of IT projects in public and state administration. Having such a powerful tool at hand, IT managers in public and state administration bodies can utilize COBIT 5 framework to ensure successful and cost-effective delivery and implementation of IT projects in alignment with the organizational goals and public expectations.

The additional benefit is found in COBIT Process Assessment Model which serves as a basis for assessing an organization's IT governance and management processes against COBIT 5. The model provides an enterprise-level view of IT process capability, an end-to-end business view of IT’s ability to create value, and helps IT managers gain upper management members buy-in for information system change and improvement initiatives.

VIII. CONCLUSION

Quality of preparation, execution, management and monitoring of IT projects in public and state administration of the Republic of Croatia and information systems audit is often under the requirements and standards for large and important projects and resources invested in this sector. None of the segments analysed in this paper is adequately covered in today’s Croatian public and state administration, which is not only methodically and from the project management aspect unacceptable, but also presents irrational use of limited government resources.

It is necessary to emphasize again that the amount discussed is around a billion kunas yearly – only 10% savings in these projects means a yearly amount large enough for building a couple of schools (bigger school building with practicums and sport gym costs around 34 million kunas) [22]. Additionally, poorly managed projects do not achieve result for which they have been initiated, which is primarily facilitating and accelerating administrative procedures.

Use from the current situation, where the system for management and monitoring of projects and IT audit is almost non-existing, and interoperability is not introduced adequately, have only large hardware and software vendors – procurement of hardware and software in such a system are much larger than necessary, and the lack of interoperability requires for each segment to develop its own software.

In this work we tried to prove that such a situation can be significantly improved with reasonable investments in implementation of standards, codes of professional practice

and education. We as well dare to say that visible positive developments would emerge very quickly, within one to two years, and their effects would be many times greater than the required investments.

Obstacle to improving the current situation can be just the lack of will, not even the lack of experts or the lack of methodology.

REFERENCES

[1] Business.hr, „Država rasipa milijardu kuna za IT bez nadzora i strategije," http://www.business.hr/ekonomija/drzava-rasipa-milijardu-kuna-za-it-bez-nadzora-i-strategije-108863 (visited: 20.1.2013)

[2] G. Vojković, "Novi Zakon o općem upravnom postupku i elektronička isprava," Pravo i porezi, vol. XVIII 11, pp. 38-47, 2009.

[3] D. Parić, „IT u državnoj upravi", PPT-presentation, http://www.vlada.hr.

[4] Zakon o ustrojstvu i djelokrugu ministarstava i drugih središnjih tijela državne uprave, Narodne novine*, No. 150/11, 22/12.

[5] "Project Cycle Management Guidelines", vol. 1, March 2004.

[6] Ibid., pp. 53-54.

[7] Zakon o javnoj nabavi, Narodne novine, No. 90/11

[8] Pravilnik o izobrazbi u području javne nabave, Narodne novine, No. 06/12

[9] A. K. Munns, B.F. Bjeirmi, "The role of project management in achieving project success," International Journal of Project Management, vol. 14, No. 2, pp. 81-87, 1996.

[10] PMP Handbook, https://www.pmi.org/Certification/~/media/PDF/Certifications/pdc_pmphandbook.ashx (visited 20.1.2013.)

[11] Project Management Institute, "A guide to Project Management Body of Knowledge (PMBOK Guide)," 2000

[12] "Survey: Poor communication causes most IT project failures", http://www.computerworld.com/s/article/9012758/Survey_Poor_communication_causes_most_IT_project_failures (visited 21.1.2013.)

[13] I. Borković, "Upravno pravo," Narodne novine, Zagreb, pp. 123, 2002.

[14] G. Vojković, "Administrative supervision of work of port authorities in Republic of Croatia," Promet - Traffic & Transportation vol. 18, 2, pp. 73-76, 2006.

[15] E. Pusić, "Uprava," Knjiga I., Školska knjiga, Zagreb, pp. 318, 1961.

[16] Zakon o državnom uredu za reviziju, Narodne novine, No. 80/11.

[17] Read more: G. Vojković; B. Cerin, "Upravljanje rizicima IT-projekata u državnoj i javnoj upravi," MIPRO 2012, 35. International Convention / Biljanović, Petar (ur.), Rijeka, pp. 1899-1903, 2012.

[18] "How to Become CISA certified", http://www.isaca.org/Certification/CISA-Certified-Information-Systems-Auditor/How-to-Become-Certified/Pages/default.aspx (visited 21.1.2013.)

[19] Auditing IT project management, http://www.theiia.org/intAuditor/itaudit/archives/2008/may/auditing-it-project-management/ (visited 21.1.2013.)

[20] ISACA, "COBIT 5 - A Business Framework for the Governance and Management of Enterprise IT," 2012.

[21] ISACA, "COBIT 5 – Enabling processes," 2012.

[22] http://www.slobodnadalmacija.hr/Hrvatska/tabid/66/articleType/ArticleView/articleId/158803/Default.aspx (visited 3.2.2013.)

* "Narodne novine" is "Official Gazette" od the Republic of Croatia.

SP

International Conference on

MIPRO Junior - STUDENT PAPERS

Steering Committee

Chairs: Ines Bombek, KSET – University of Zagreb, Croatia
Goran Cetušić, KSET – University of Zagreb, Croatia

PAPERS

Upravljanje vozilom na električni pogon

Denis Damjanović [1], Vedran Kirinčić [2], Srđan Skok [3]

Zavod za elektroenergetiku, Tehnički fakultet, Sveučilište u Rijeci
Vukovarska 58, 51000 Rijeka. Hrvatska

[1] ddamjan@riteh.hr
[2] vedran.kirincic@riteh.hr
[3] srdjan.skok@riteh.hr

Sažetak - U radu je opisana regulacija električnog motora u sprezi s baterijom, koja je provedena za potrebe studentskog projekta razvoja vozila na električni pogon pri Tehničkom fakultetu Sveučilišta u Rijeci. Za kontrolu momenta sinkronog motora SPM132 - 01 s permanentnim magnetima upotrijebljena je metoda vektorske regulacije statorskim tokom. Provedena je simulacija rada motora reguliranog korištenjem navedene metode. Korištenjem programskog paketa Matlab Simulink modelirani su upravljački krugovi za regulaciju brzine vrtnje i krugovi regulacije preko metode slabljenja polja te je prikazana ovisnost brzine vrtnje o poprečnoj komponenti statorske struje. Promatrane su prijelazne pojave te je objašnjena njihova dinamika. Izvršen je odabir metode koja rezultira boljim karakteristikama te je objašnjena njena učinkovitost i veza s odabranim mikrokontrolerom, ugrađenim u vozilo na električni pogon u cilju praćenja parametara i lakšeg otkrivanja kvarova.

Ključne riječi – električna vozila, regulacija, vektorska regulacija statorskim tokom, dinamički odziv sustava.

I. UVOD

Kako bi se ublažile negativne posljedice klimatskih promjena i učinci globalnog zatopljenja, postavljeni su ciljevi smanjenja emisija stakleničkih plinova, integracije obnovljivih izvora energije (OIE) i povećanja ukupne energetske učinkovitosti [1], [2]. Elektroprivrede diljem svijeta stoga razvijaju tehnologije, koncepte i rješenja koja će omogućiti postupan prijelaz današnjeg elektroenergetskog sustava (EES) u viziju poznatu kao napredna elektroenergetska mreža (eng. Smart Grid). Dio tog rješenja i važno područje interesa predstavlja automobilski sektor, kao značajan čimbenik u ukupnoj potrošnji i zagađenju [3].

Povećana upotreba električnih vozila (eng. Electric Vehicle – EV) ima pozitivan utjecaj na okoliš zbog smanjenja emisije CO_2. Pored toga električni uređaji korišteni u EV bi trebali imati visoku razinu učinkovitosti kako bi se gubici i potrošnja minimizirali. Također, EV mogu koristiti mogućnost reverzibilnog kočenja te štede energiju u slučaju zastoja u prometnim gužvama. Samim time se najveća ušteda energije, u odnosu na vozila pogonjena motorom s unutrašnjim izgaranjem (eng. Internal Combustion Engine Vehicle - ICEV), ostvaruje u uvjetima gradske vožnje.

Razmatrajući navedene čimbenike vezane za automobilski sektor, studentski projekt pri Tehničkom fakultetu Sveučilišta u Rijeci kao krajnji cilj ima razvoj EV preradom postojećeg ICEV. Kao primjer rješenja je ponuđeno malo električno vozilo marke Smart Fortwo s dovoljno autonomije za gradske potrebe. Ovaj se rad fokusira na odabir metode regulacije korištenog sinkronog motora s permanentnim magnetima.

II. VEKTORSKA REGULACIJA STATORSKIM TOKOM

Kod sinkronog motora s permanentnim magnetima porastom opterećenja rotor počinje usporavati te senzor pozicije rotora preko upravljačke logike daje nalog za sporije prebacivanje vođenja struje s jednog namota na drugi. Senzor položaja rotora preko upravljačke logike direktno utječe na frekvenciju izmjenjivača, odnosno u konkretnom slučaju se frekvencija pretvarača smanjuje. Dinamički moment, koji nastaje rasterećenjem motora, vraća pogon u radnu točku neopterećenog stroja, a to znači da impulsi od senzora prema upravljačkoj logici pristižu brže pa se i frekvencija rada izmjenjivača povećava.

Suvremeni motori koji se koriste u električnom pogonu napajaju se iz energetskih pretvarača. Za upravljanje pretvaračima koriste se sofisticirane tehnike koje u najvećoj mjeri ovise o podacima prikupljenim senzorima. Najkorištenije metode vektorske regulacije su regulacija statorskim tokom (eng. Field Oriented Control - FOC) i direktna regulacija momenta (eng. Direct Torque Control - DTC) koja koristi modulacijsku tehniku s prostornim vektorom (eng. Space Vector Modulation - SVM) [4], [5]. Primjenom DTC metode moguće je matematičkim transformacijama izraziti moment u ovisnosti o iznosima statorskog i rotorskog toka te njihovom međusobnom kutu. Budući je ovisnost toka statora o naponu statora izravna, upravljanjem vektorom napona moguće je regulirati moment motora. Glavni nedostatak je promjenjiva frekvencija sklapanja [6]. Izvedbom DTC-SVM metode omogućen je rad i pri konstantnoj frekvenciji sklapanja. Time ova metoda postaje sve važnija za rad sinkronog motora s permanentnim magnetima (naročito kod njegovog zaleta).

U programskom paketu Matlab Simulink izrađena su dva upravljačka modela, jedan s regulatorom struje i drugi koji je proširen s nadređenim regulatorom brzine. Modeli su korišteni kako bi se ispitala FOC metoda za upravljanje sinkronih motora s permanentnim magnetima. Dinamički

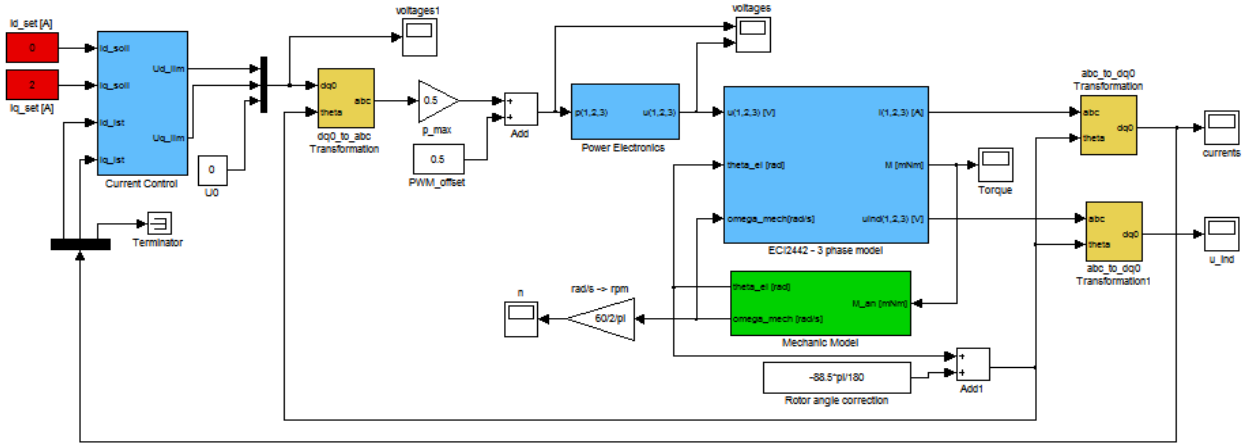
model sinkronog motora s permanentnim magnetima je razmatran uz sljedeće pretpostavke:

- motor je električki i geometrijski simetričan u sve tri faze,
- parametri motora su koncentrirani,
- raspodjela magnetskog polja u zračnom rasporu je sinusna (zanemareni viši harmonici),
- zasićenje i gubici u željezu se zanemaruju,
- na rotoru nema prigušnog kaveza,
- magnetski tok rotora, otpori i induktiviteti imaju konstantnu vrijednost,
- efektivna duljina zračnog raspora je konstantna po cijelom obodu stroja,
- priključak motora na trofazni izvor napona mreže je bez nul-vodiča.

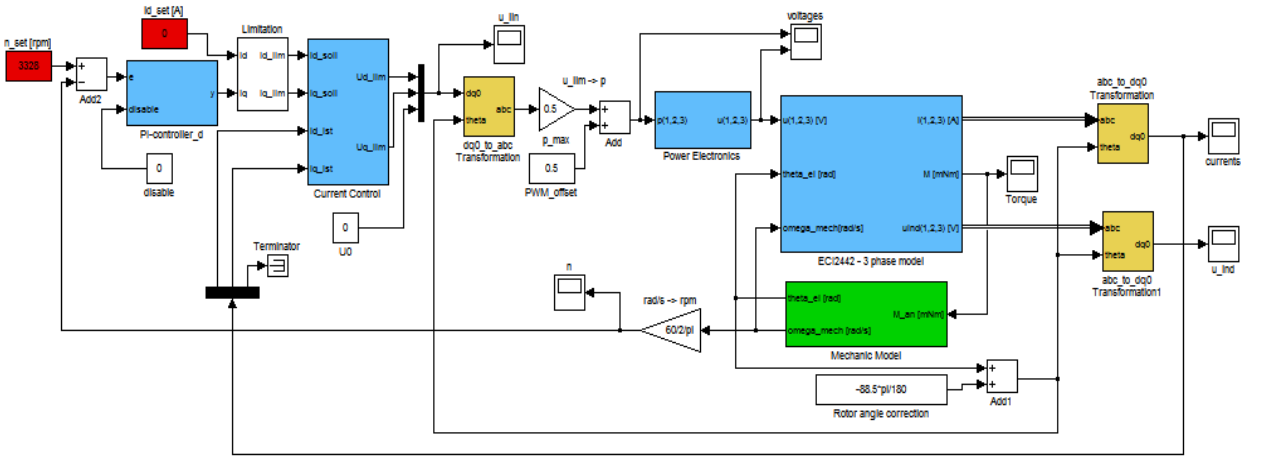
III. SIMULACIJSKI MODELI

Na Slici 1. prikazan je upravljački krug s jednostavnim proporcionalno-integracijskim (PI) regulatorom struje [7]. Kako bi se postigla regulacija povratnom vezom, osmatrane su dvije fazne struje statora motora (i_1 i i_2). Budući se u procesu zaleta motor pokreće od stanja mirovanja do nazivne brzine vrtnje, za regulaciju je nužan frekventni pretvarač. Upravljanje pretvaračem je ostvareno modulacijom širine impulsa (eng. Pulse Width Modulation - PWM). Pritom je nužno provesti transformaciju iz statorskog u rotorski koordinatni sustav, obzirom da poznavanje kuta rotora omogućuje realizaciju vektorske orijentacije Regulacija momenta je ostvarena pomoću poprečne komponente statorske struje i_{sq}.

Slika 2. prikazuje upravljački krug koji je proširen dodatnim vanjskim PI regulatorom brzine okretaja [7], čiji je izlaz ograničen kako bi se spriječilo nekontrolirano upravljanje.



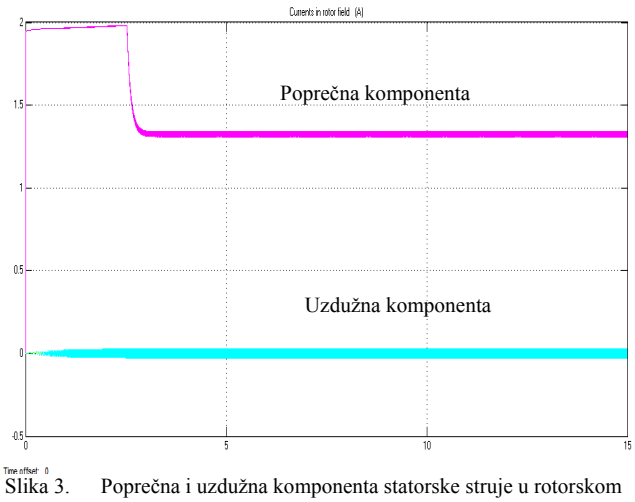
Slika 1. Simulacijski model upravljačkog kruga s regulatorom struje [7]



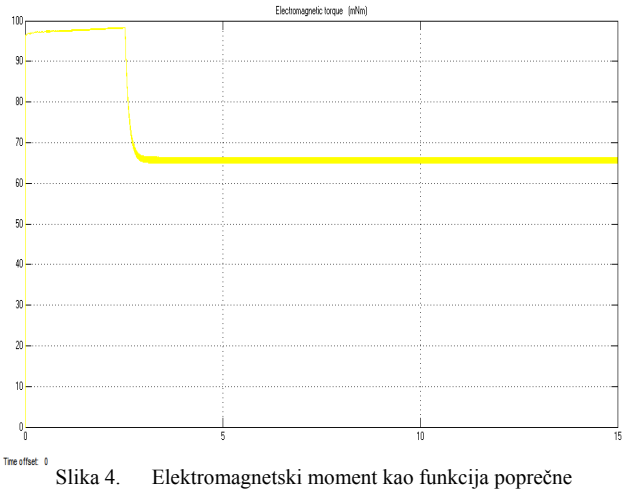
Slika 2. Simulacijski model upravljačkog kruga s dodatnim regulatorom brzine vrtnje [7]

Za upravljački krug prikazan Slikom 1. provedene su simulacije za vrijednosti poprečne komponente statorske struje. Simulacija je provedena za vrijednost struje $i_{sqs} = 2$ Ampera. Veličine koje su promatrane i koje su nužne za regulacijski krug su moment i brzina rotora te struje i naponi statora izraženi u rotorskom koordinatnom sustavu. Slika 3. - Slika 6. prikazuju rezultate dobivene simulacijom modela koji koristi samo regulator struje.

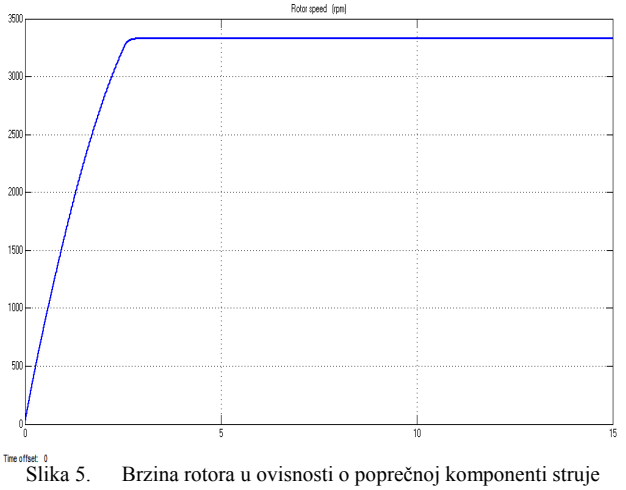
Slika 3. prikazuje uzdužnu i poprečnu komponentu struje statora u rotorskom koordinatnom polju. Uzdužna komponenta struje je približno jednaka nuli i ne sudjeluje u stvaranju magnetskog polja na rotoru zbog toga što je sastavljen od permanentnih magneta. Poprečna komponenta struje je u početku simulacije konstantnog iznosa i sudjeluje u stvaranju zaletnog momenta, a kad motor postigne nazivnu vrijednost brzine, ona opada na nižu vrijednost manju od 1,5 Ampera. Slika 4. prikazuje vanjsku karakteristiku elektromagnetskog momenta, pri čemu se mogu razlikovati područja konstantnog momenta i konstantne snage. U početku motor ima konstantni zaletni moment, dok nakon postizanja nazivne brzine moment opada na nižu vrijednost. Iznad nazivne brzine vrtnje motora, odnosno u području upravljanja konstantne snage, napon napajanja motora je konstantnog nazivnog iznosa. Brzinom vrtnje motora, koja je prikazana Slikom 5., može se upravljati smanjenjem ili slabljenjem magnetskog toka u zračnom rasporu. Slika 6. prikazuje napone regulacije u rotorskom polju. Uočava se linearan porast poprečne komponente induciranog napona do konstantnog iznosa, dok je uzdužna komponenta jednaka nuli.



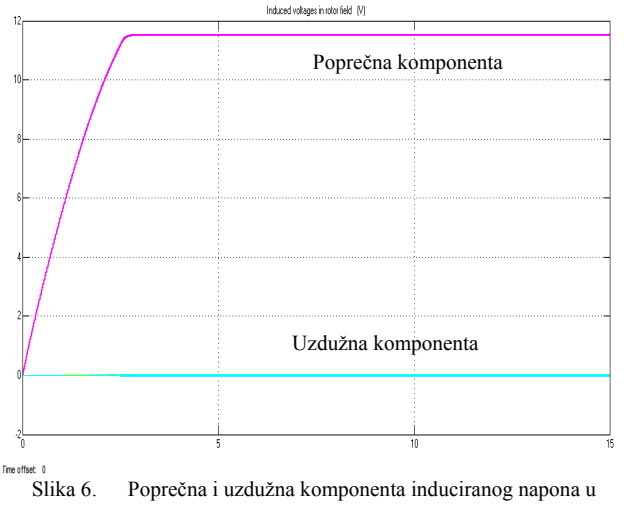
Slika 3. Poprečna i uzdužna komponenta statorske struje u rotorskom polju



Slika 4. Elektromagnetski moment kao funkcija poprečne komponente statorske struje



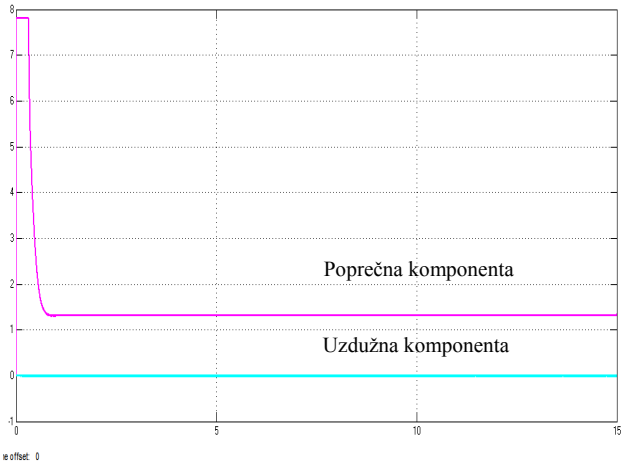
Slika 5. Brzina rotora u ovisnosti o poprečnoj komponenti struje



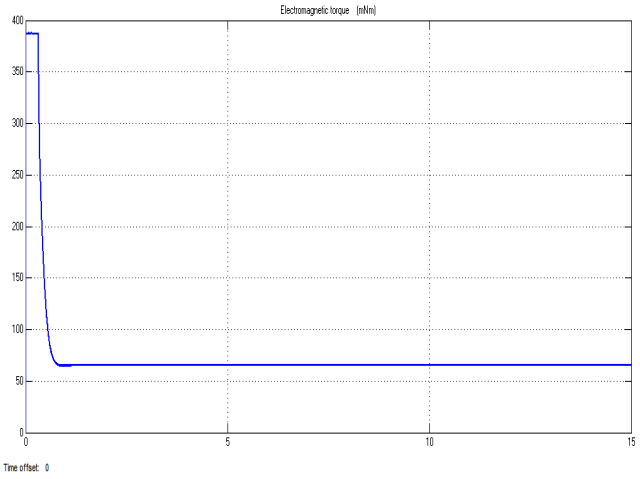
Slika 6. Poprečna i uzdužna komponenta induciranog napona u rotorskom polju

U nastavku su dani rezultati dobiveni simulacijom modela, prikazanog Slikom 2, u kojem je uz regulator struje dodan vanjski regulator brzine vrtnje s ograničenjem izlaza regulatora brzine. Korištenjem navedenog modela je ostvareno vektorsko upravljanje s

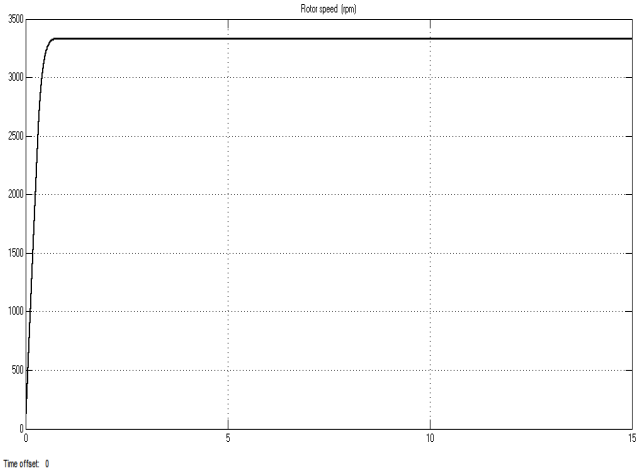
većom učinkovitošću. U stvaranju momenta sudjeluje poprečna komponenta statorske struje stoga ona ima isti vremenski odziv kao i moment, što prikazuje Slika 7. Slika 8. prikazuje momentnu karakteristiku na kojoj se vidi isto tako brzi odziv sustava i veća vrijednost početnog momenta što ukazuje na pokretanje motora s maksimalnom snagom pri zadanom opterećenju. Na Slici 9. je vidljivo da brzina vrtnje linearno raste te motor brzo postiže zalet do nazivne vrijednosti. Slika 10. prikazuje brzi aperiodski odziv inducirane poprečne komponente napona u rotorskom polju s jako malim nadvišenjem što ukazuje na brzi odziv sustava i stabilnu regulaciju. Uzdužna komponenta napona je jednaka nuli te se ne generira unutar magnetskog polja.



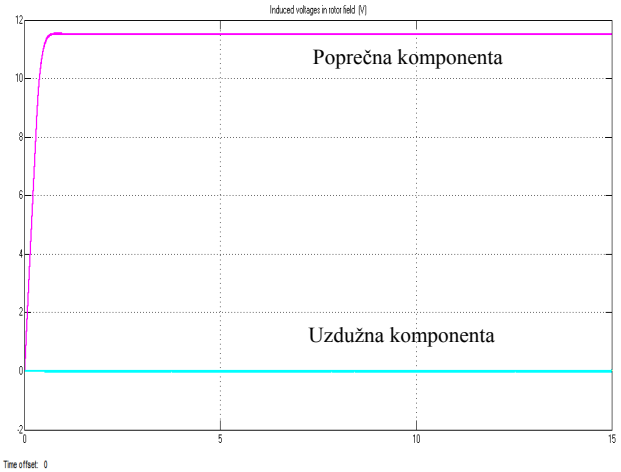
Slika 7. Poprečna i uzdužna komponenta statorske struje u rotorskom polju



Slika 8. Elektromagnetski moment kao funkcija poprečne komponente statorske struje



Slika 9. Brzina rotora u ovisnosti o poprečnoj komponenti struje



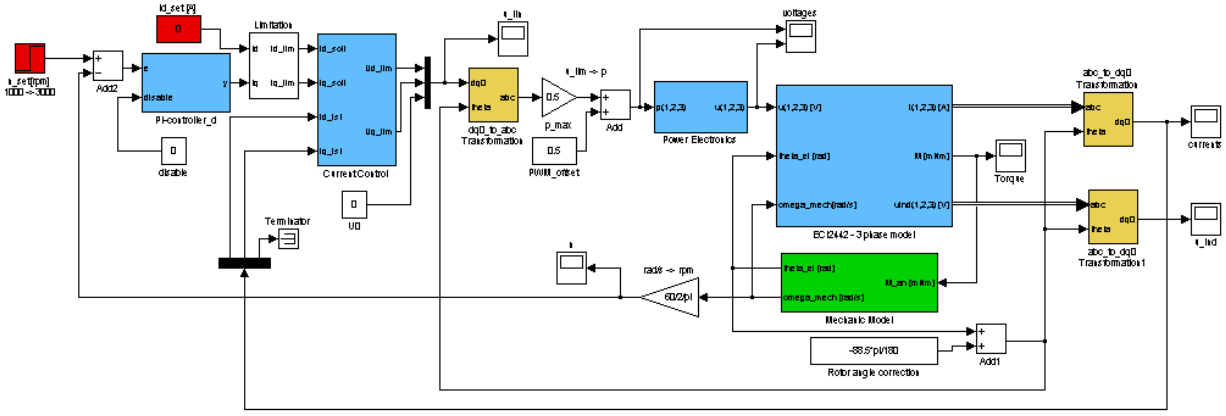
Slika 10. Poprečna i uzdužna komponenta induciranog napona u rotorskom polju

Kako efekt slabljenja polja u području konstantne snage kod sinkronih motora s vanjskim permanentnim magnetima ovisi o sinkronoj reaktanciji stroja (koja ovisi o zračnom rasporu i iznosi do 1 mm) odnosno reakciji armature, upravljanje je ograničeno i ne može se primijeniti za široki raspon brzine. Opisanim modelom simulirano je upravljanje metodom slabljenja polja pri brzini od 3000 o/min, pri kojoj je moment bio konstantan, a motor još nije dosegnuo maksimalnu snagu. Rezultati simulacije ukazuju da je nemoguće ubrzati motor upravo zbog malog zračnog raspora i djelovanja reaktancije stroja.

IV. SIMULACIJA MODELA S PROMJENOM BRZINE VRTNJE

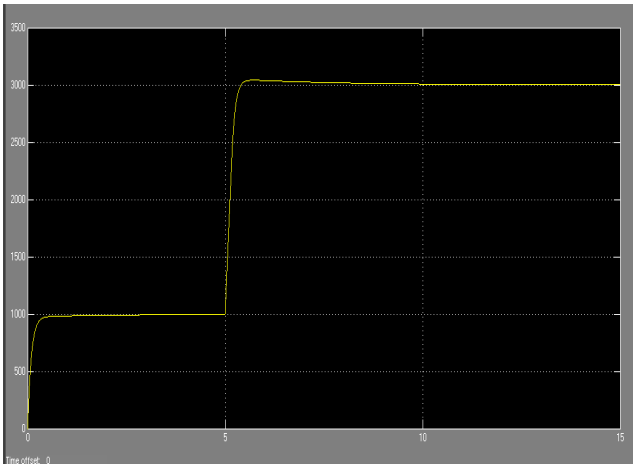
Model prikazan na Slici 11. je dodatno korišten je i za simulaciju skokovite promjene brzine vrtnje s 1000 o/min na 3000 o/min, kako bi se nakon ugađanja parametara regulatora dobilo brzo vrijeme postavljanja brzine vrtnje. Pri upravljanju brzinom vrtnje na izlaz je stavljen limiter. PI regulator brzine ima već u sebi ugrađen anti-windup model koji sprječava efekt zaleta upravljačke veličine te

ne kvari vladanje zatvorenog kruga upravljanja, obzirom da ne dozvoljava da izlaz iz regulatora izađe iz ograničenja.

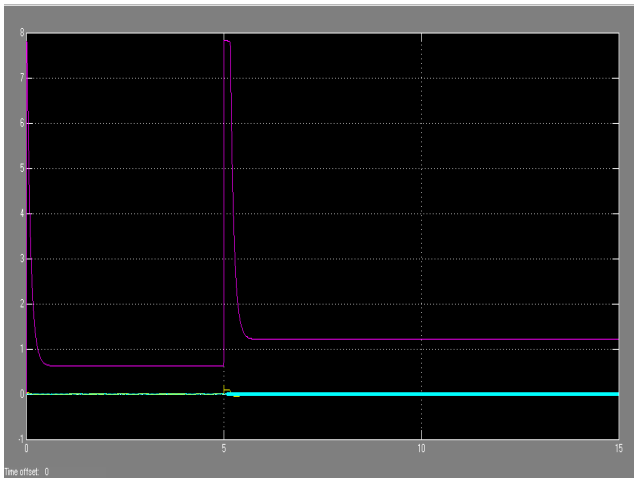


Slika 11. Simulacija modela s promjenom brzine okretaja

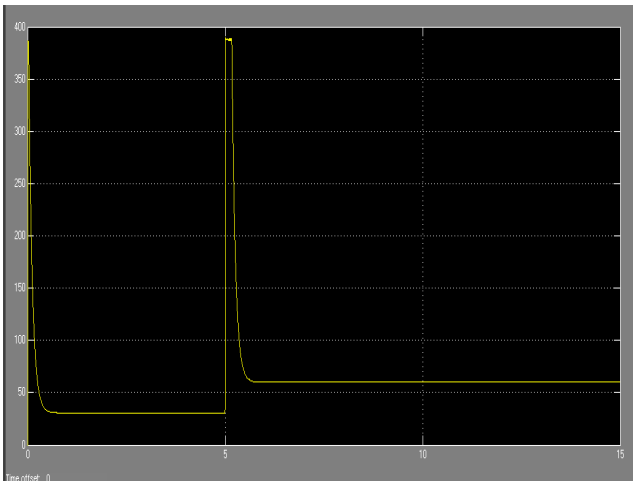
Na Slici 12. prikazane su komponente statorske struje koje sudjeluju u regulacijskom krugu pri promjeni brzine vrtnje s 1000 na 3000 okretaja u minuti. Treća komponenta označena žutom bojom na Slici 12. nastaje zbog nagle i skokovite promjene brzine koja se kasnije eliminira u krugu. Na Slici 13. prikazan je skokovit prijelaz brzine koja je nakon prijelaznog vremena regulirana na konstantnu vrijednost. U trenutku skokovitog prijelaza brzine vrtnje elektromagnetski moment „skače“ na nazivnu vrijednost od 390 Nm kao i na početku startanja motora Utjecaj njegovog ponašanja je vidljiv na Slici 14. Poprečna komponenta induciranog napona na Slici 15. mijenja oblik pri promjeni brzine i na taj se način može vidjeti da i ona sudjeluje pri stvaranju okretnog magnetskog polja u rotoru motora.



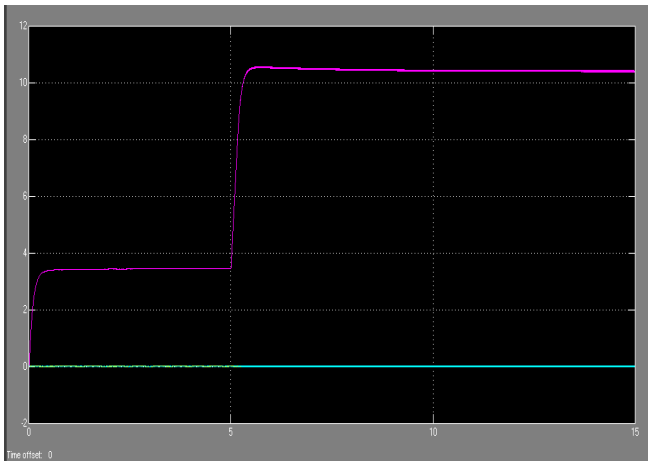
Slika 13. Brzina rotora u ovisnosti o poprečnoj komponenti struje



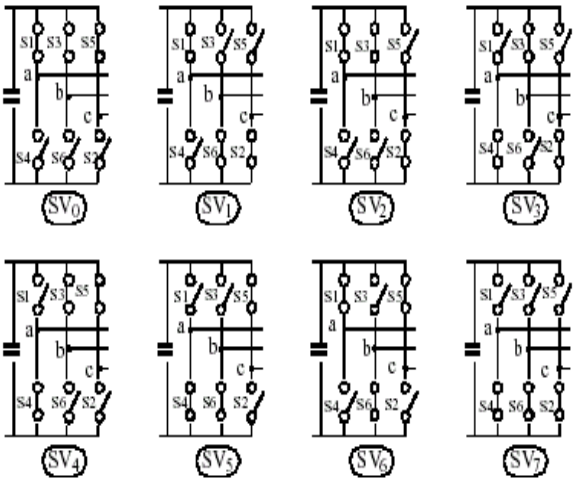
Slika 12. Poprečna i uzdužna komponenta statorske struje u rotorskom polju



Slika 14. Elektromagnetski moment kao funkcija poprečne komponente statorske struje



Slika 15. Poprečna i uzdužna komponenta induciranog napona u rotorskom polju



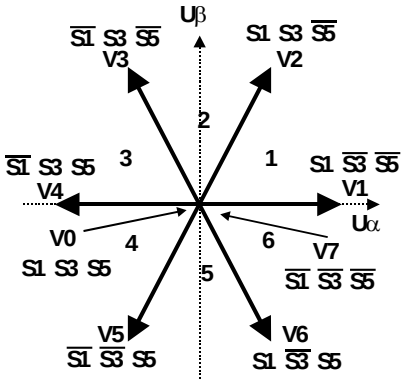
Slika 16. Moguće kombinacije faznih napona trofaznog invertera

V. ODABIR VRSTE REGULATORA

Napretkom tehnologije modulacijska tehnika s prostornim vektorom je u suštini digitalna modulacijska tehnika pa je pojavom digitalnih procesora signala postala jedna od najčešće korištenih PWM tehnika za upravljanje trofaznim izmjeničnim (AC) sinkronim motorima s permanentnim magnetima.

Za studentski projekt izrade EV odabran je regulator koji koristi navedenu metodu s mikroprocesorom koji ima integriranu brzu ´look up´ tablicu za preračunavanje i provodi fazno upravljanje sekvencijom napona (eng. Sequential Phase Control - SPC).

Trofazni naponski inverter, prikazan Slikom 16., ima osam različitih stanja koja su definirana stanjima prekidača invertera, a svako stanje odgovara jednom skupu vrijednosti faznih napona. To praktično znači da inverter može generirati osam stacionarnih vektora SV0 do SV7, gdje broj u indeksu označava odgovarajuće stanje invertera. Od ovih osam vektora dva su nulta SV0 i SV7, a ostali čine pravilni šesterokut kao što je prikazano na Slici 17 [8].



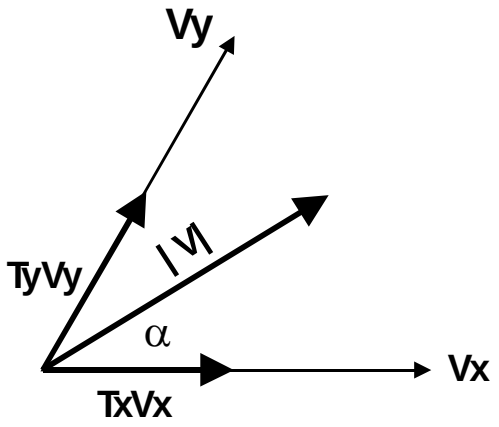
Slika 17. Stacionarni vektori 3Φ invertera

Amplituda V_m stacionarnog ne nultog vektora je dana s:

$$V_m = \frac{2}{3} V_{DC} \tag{1}$$

gdje je V_{DC} vrijednost ulaznog DC napona invertera. Problem upravljanja frekvencijom i amplitudom napona na izlazu invertera svodi se na generiranje prostornog vektora v koji se okreće željenom kutnom brzinom ω i ima amplitudu $|V|$, koja odgovara zadanom stupnju modulacije.

Dva susjedna stacionarna vektora čine jedan segment, koji u vremenskoj domeni gledajući odgovara 1/6 periode izlaznog napona invertera. Princip rada SVPWM tehnike (eng. Space Vector Pulse Width Modulation - SVPWM) prikazan je na Slici 18.



Slika 18. Princip generiranja prostornog vektora

Uzmimo proizvoljni segment fazne ravnine koji je omeđen dvama susjednim prostornim vektorima **Vx** i **Vy**. Promotrimo jedan prekidački ciklus i pretpostavimo da je referentni prostorni vektor **V** konstantan u toku prekidačkog ciklusa. Položaj prostornog vektora **V** u segmentu određen je komponentama Tx**Vx**, Ty**Vy** i Tz**Vz** (**Vz** jedan od dva nulta vektora $\mathbf{V}_0$ ili $\mathbf{V}_7$). Vrijednosti Tx, Ty i Tz su vremena trajanja stacionarnih stanja invertera **Vx**,**Vy** i **Vz** od kojih se generira prostorni vektor i koje treba odrediti. Stacionarne vektore **Vx**, **Vy** i **Vz** možemo predstaviti kao:

$$\begin{aligned} V_x &= V_m e^{j\theta} \\ V_y &= V_m e^{j\left(\theta+\frac{\pi}{3}\right)} \\ V_z &= 0 \end{aligned} \tag{2}$$

Referentni prostorni vektor možemo izraziti kao:

$$\mathbf{V} = MV_m e^{j(\theta+\alpha)} \tag{3}$$

gdje je M stupanj modulacije trajanja pojedinih stanja. Tx, Ty i Tz se mogu odrediti razdvajanjem realnog i imaginarnog dijela referentnog vektora te su dani izrazima:

$$\begin{aligned} T_x &= MV_m \sin(\frac{\pi}{3}-\alpha) \\ T_y &= MV_m \sin(\alpha) \\ T_z &= 1 - T_x - T_y \end{aligned} \tag{4}$$

Jedna od osnovnih prednosti SVPWM tehnika nad PWM tehnikama s modulacijskom funkcijom je manji broj promjene stanja prekidača invertera po broju prekidačkih intervala, što je direktna posljedica fleksibilnosti izbora stanja invertera.

SPC koristi direktno integriranu tehnologiju upravljanja te mjeri struju u svakoj fazi u svakom PWM ciklusu osiguravajući potpuno uravnoteženu glatku i učinkovitu kontrolu nad vozilom. SPC pruža poboljšanu performansu nad standardnom kontrolom proklizavanja, posebno u dinamičnim opterećenjima. Načelna prednost takve tehnologije je u učinkovitom napajanju MOSFET prekidača koji se mogu montirati direktno na PCB pločicu (eng. Printed Circuit Bord) čime se ostvaruje brzo i efikasno hlađenje. Prema tome, pouzdanost i učinkovitost su značajno poboljšane hlađenjem snažnih sklopnih uređaja. Ovaj pristup također vodi ka znatnim poboljšanjima kontinuiranog prijenosa snage, omjera vršne snage, s kontinuiranim radom upravljača kao jednom od najvažnijih aspekata u određivanju performansi vozila.

Periferijski uređaj ADMC201 koji je ugrađen u regulator za EV posjeduje 12-bitni A/D konverter sa sinkroniziranim procesom odmjeravanja (100 ns), jedinicu za regularnu (uniformnu) i SVPWM modulaciju, kao i modul za prihvaćanje i obradu signala sa inkrementalnog enkodera. Pored spomenutih periferijskih modula, ADMC201 uključuje i sklopovski implementiranu Clarkovu i Parkovu transformaciju potrebnu za SVPWM algoritme, čime se rasterećuje DSP (eng. Digital Signal Processor) i znatno uvećava brzina rada. Korištenjem frekvencije sustava od 10 MHz, sklopna frekvencija PWM modulacije može se podesiti u rasponu od 2,5 kHz do 20 kHz [9].

Odabrani regulator (mikrokontroler), koji služi regulaciji motora u vozilu na električni pogon ima sučelje povezano s kalibratorom. Ručni kalibrator ima programski alat koji se koristi za konfiguriranje i praćenje svih parametara regulatora, kao i otkrivanje kvarova 0.

VI. ZAKLJUČAK

Korištenjem prikazanih modela regulacije sinkronog motora s permanentnim magnetima, koji se koristi u EV razvijanom u sklopu studentskog projekta pri Tehničkom faktuletu Sveučilišta u Rijeci, provedene su simulacije dinamičkih pojava u motorskom načinu rada. Prikazani su odzivi praćenih veličina prilikom zaleta motora i njegovog daljnjeg rada. Metode s dodanim regulatorom brzine vrtnje ukazuju na bolje rezultate te otvaraju mogućnost daljnje primjene u radu. Problem povišenja simuliranih regulacijskih veličina iznad nazivnih nije značajan i ne predstavlja ograničenje u primjeni ove metode kod sinkronih motora. Problem nastaje jedino kod proširenja opsega brzine vrtnje motora u području

konstantne snage stoga je moguće nastaviti istraživanja na način da se u dinamiku uključi i regulacija d-komponente struje statora (koja je do sad zadavana kao konstantna vrijednost) i tako za još jedan stupanj poveća dimenzija regulacije.

Odabrani regultator omogućuje jednostavno i lako usklađivanje motora s karakteristikama regulatora te optimiranje željenih karakteristika vozila.

LITERATURA

[1] D. Radulovic; S. Skok; V. Kirincic; "Energy efficiency public lighting management in the cities," Energy, Elsevier, vol. 36, no. 4, pp. 1908-1915, 2011.

[2] D. Radulovic; S. Skok; V. Kirincic; "Cogeneration – investment dilemma," Energy, Elsevier, accepted for publication, July 2012.

[3] P. Sasa, S. Skok, V. Kirincic, „Urban electric vehicle development as part of Smart Grid“, International Conference on Innovative Technologies (IN-TECH) Rijeka, 26 – 28 September 2012, Rijeka, Croatia.

[4] M. Zelechowski, “Space Vector Modulated – Direct Torque Controlled (DTC-SVM)” Inverter- Fed Induction Motor Drive, PhD Thesis, Warsaw University of Technology, 2005

[5] C. Mi, “Field-oriented Control of Induction Motor Drives with Direct Rotor Current Estimation for Applications in Electric and Hybrid Vechiles”, Journal of Asian Electric Vechiles, Volume 5, Number 2, December 2007.

[6] Z. Maljković, M. Šundrica, “Numerička simulacija vektorske regulacije sinkronog stroja”, HRO CIG*RE*, Cavtat, studeni 2009

[7] W. Gruber: "Synchronous Machines", Materials for laboratory exercise, University of Rijeka, Faculty of Engineering

[8] S. Lubura, Mile K. Stojčev, Pregled PWM tehnika i načina njihove digitalne implementacije“, INFOTEH-JAHORINA, Vol. 4, Ref. E-III-14, p. 379-386, March 2005.

[9] ANALOG DEVICES Datasheet Motion Coprocessor ADMC201 web: www.datasheetcatalog.com

DMC GmbH, „PC Interface manual – Revision 1.2“ 01/2011, web: www.dmcde.de

The development of charging stations for electric vehicles: a solution or a problem?

Ivana Grdic [1], Vedran Kirincic [2], Srdjan Skok [3]

Department of Power Systems, Faculty of engineering, University of Rijeka
Vukovarska 58, 51000 Rijeka. Croatia
[1] igrdic@riteh.hr
[2] vedran.kirincic@riteh.hr
[3] srdjan.skok@riteh.hr

Abstract - Trends in automotive industry support global goals of reducing greenhouse gases and increasing overall energy efficiency. Increased use of electric vehicles (EVs) has to be supported with adequate charging infrastructure. This paper presents a plan of technical solutions and design approach of charging stations, with some particular solutions of existing public charging stations, as well as the description of the situation in Croatia, compared with other countries. A special attention is paid to the development of a charging system as a distributed source of electrical energy. Simultaneous recharging of the large number of EVs will have a significant impact on the power system operation, such as generation capacity, transformer loading levels, and load profiles. There is a growing need for widely distributed publicly accessible charging stations, some of which support faster charging at higher voltages and currents. The optimization of the time and the place of charging are also investigated. Using a combination of standardized protocols and wide-area communication technologies, all the charging points can be connected in an intelligent network that provides valuable real-time information about the charging station, which results in higher reliability and stability of the power system.

Keywords: charging station, electric vehicle, optimized charging, power system, Smart Grid

I. INTRODUCTION

Wider offer of charging services for electric vehicles (EVs) encourages people to drive the EVs and vice versa. The EVs could play an important role in the government's overall energy policy because of their potential to reduce our dependence on fossil fuels and reduce greenhouse gases emissions, along with other initiatives that include integration of renewable energy sources (RES) and increase of overall energy efficiency [1]-[2]. The transport sector, one of the biggest single consumers of final energy, is recognized as the fastest growing sector in the last decade.

The EV topic raises many dilemmas. Standpoints vary from those who consider the EVs and the RES as a realistic solution for the environmental problems to those who claim that the technology is still immature. Disposal of batteries or solar panels, which affects the environmental state, is yet to be discussed. Furthermore, introduction of a large number of charging points will significantly change the power system operation and probably influence its stability and reliability. On the other hand, there are ideas of returning the energy from the EVs to the grid in load peak periods.

The EVs require some basic changes in driving habits. Current obstacles of the EV technology are a limited driving range on electric power, long-lasting recharging and a relatively high starting investment in comparison with conventional vehicles of similar performances, which use an internal combustion engine (internal combustion engine vehicle - ICEV). The life expectancy and replacement cost of the battery packs are still an important drawback, and in most regions the ability to recharge in public areas is limited. As life habits of the majority of urban population assume local shorter distances, the need for charging mid-trip is reduced. Nevertheless, longer trips between cities and adjacent urban areas require a network of public charging stations or another method to extend the range of EVs beyond the normal daily commute.

The charging stations can leverage the existing ubiquitous electrical grid and home recharging is a realistic option. Anxiety regarding available driving range and finding the charging station can be a major concern for potential EV drivers. This can be supported with online directories such as EV-Networks or some charging station providers, which publish availability of their charging locations online in real time.

As the EV technology is an interesting topic, students at the Faculty of Engineering, University of Rijeka started the EV project. Intention of this paper is to focus on the charging infrastructure, as an important part of the EV solution.

II. ENVIRONMENTAL SITUATION

Modern public transportation is one of the most important parts in a city plan and the mean for millions of citizen for their daily commute. In the recent years there have been a lot of discussions concerning the future of personal ICEVs. Main issues were global warming, pollution and the dependency upon the non-renewable oil. The environmental advantages of the EVs are beneficial for both people and nature.

Considering the CO_2 emissions, the transport sector contributes to the overall emissions with about 17 %. In comparison with the ICEVs, the emissions from the EVs are still considerably lower. Table I shows the potential of the EVs that is charged from the RES to reduce the CO_2 emissions drastically [3]. For example if we analyze the ICEV of 1500 kg and which consumption is 8 l / 100 km, its emission of the CO_2 equals to 217 g / km. On the other hand, for the EV with the same weight the emission of CO_2 is lower, especially if the electric energy is produced from a wind power plant. In this case the CO_2 emission is 0.9 g / km that represents reduction of 99.60 %. The average value of the CO_2 emission from an ICEV is 203 g / km. Although it is supposed to be reduced to 175 g / km by 2014 and further to 147 g / km by 2020, the values are still relatively high. The industry estimates that 10 % of an ICEV's lifetime CO_2 emissions are from the production phase, 85 % from the in-use phase and 5 % from the end-of-life phase. On the other hand, the CO_2 emissions from an EV reach the average value of 105 g/km.

TABLE I. CO_2 EMISSIONS [3]

CONVENTIONAL VEHICLE						
CHARACTERISTICS	WEIGHT	GASOLINE	EMISSION CO2			
	kg	L/100 km	g/km			
	1600	9	244			
	1500	8	217			
	1400	7	190			
	1300	6	167			
	1200	5	136			
ELECTRIC VEHICLE						
POWER PLANT	COAL	NATURAL GAS boiler	NATURAL GAS comb. cycles	HYDRO ELECTRIC	NUCLEAR	WIND TURBINES
EMISSION CO2 g/kWh	1075	653	461	18	15	5.5
EMISSION CO2 g/km	172	104	74	2.9	2.4	0.9
SAVING CO2 g/km	45	113	143	214.1	214.6	216.1
EFFICIENCY	20.7%	52.1%	65.9%	98.7%	98.9%	99.6%
(EV: 1500 kg, consuming 16 kWh/100 km from the power plant to the wheels)						

III. INVESTMENT DILEMMA

When considering the investment for buying an EV, there are some factors that should be taken into account.

How far do you drive? If you will be using the vehicle only for running local errands or you have a relatively short commute, for instance 50 km a day, an EV would probably fit your needs and provide enough buffer to allow you some side trips.

Will an EV save you money? Depending on your electricity rates, driving an EV can be less expensive than filling up at the gas station. An important factor are much lower maintenance costs in comparison with an ICEV, but the life expectancy and the price of battery packs should be carefully planned.

How will you charge your vehicle? You can charge an EV at any household outlet, but the type of circuit dramatically affects the charging time. A standard 220 V outlet could charge an EV overnight.

How much will you pay for charging? Many utilities offer a wide variety of rate plans, so it is important to find out which is the optimal for you. Off-peak rates are usually lower because you are taking advantage of the excess capacity that power plants have at night.

IV. CHARGING PLACE AND TIME

Just as an ICEV has a fuel tank, fuel and an ICE engine so does an EV have batteries, electricity and an electric motor. Instead of refueling the fuel tank at the gas station, the EV is recharged at a special station or simply at home.

Charging stations can be found and will be needed where there is an on-street parking, at taxi stands, in parking lots (at places of employment, hotels, airports, shopping centers, convenience shops, fast food restaurants, coffeehouses, etc.), as well as in driveways and garages at home. Existing filling stations may also become or may incorporate charging. The charging stations can be private or public.

Table II describes the most important characteristics of the charging stations designed and operated until now.

TABLE II. CHARGING STATION CHARACTERISTICS [4]

TYPE	**SLOW CHARGE**	**NORMAL CHARGE**	**FAST CHARGE**	**BATTERY SWAPPING**
POWER	<3.7 kW	3.7-44 kW	>44 kW	
SOURCE	AC 1F; 16 A; 220 VAC	AC 1F; 32 A; 220 VAC	DC; 100-125 A; <500 V	
		AC 3F; 32 A; 380 VAC		
TIME (battery capacity 16 kWh)	6-8 h	4 h-30 min	30-3 min	<3 min
AUTONOMY/h	<20 km	20-110 km	>110 km	
PRICE €	3000	4000-10000	20000-50000	
MODE	**1**	**2**	**3**	**4**
	slow charging from a regular electrical socket (1F/3F)	slow charging from a regular socket equipped with some EV specific protection	slow or fast charging using a specific EV multi-pin socket with control and protection	fast charging using some special charger technology such as CHAdeMO

In comparison with the charging station, a battery switch station allows the driver to swap a discharged battery or battery pack for a fully charged one, saving the time of waiting for the battery to be charged. The driver has to check-in by an identification card, which will send information to the main centre about the vehicle model and the battery type. If the station disposes with the required battery, the change will start. At the battery switch station, the driver does not need to get out of the vehicle while the battery is being swapped. The swapping process last less than a few minutes, which is faster than refueling an ICEV. The idea is that the drivers do not own the battery; they just pay for the energy they use to change the battery pack.

V. COMMUNICATION

Every charging station should be station network-enabled and embedded with an on-board computer, a display, a standards-based RFID (Radio-Frequency identification) reader and a utility-grade meter that provides precise, bi-directional energy measurement. The communication can be realized through the GSM service

between the vehicle and the charging station, between the charging station and the customer, and between the charging station and the energy supplier.

Owing to the communication links between all the parts of the charging station and between the vehicle owners, it is known exactly the state and status of EVs, with real-time charging location information, EV charging summaries and usage report. Fig. 1 gives the steps during the charging of an EV [4]-[6].

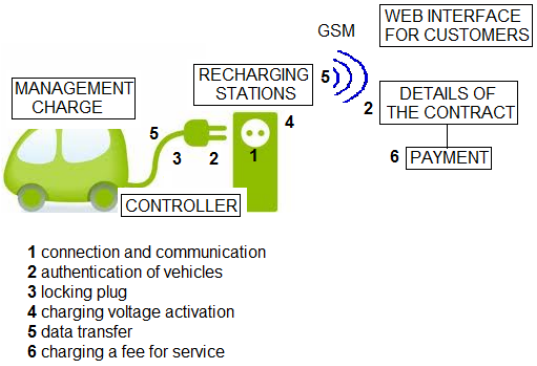


Figure 1. Steps during the charging of an EV

Using a combination of standards-based protocols and wide-area communication technologies, all the charging points are connected over an intelligent network that provides valuable, real-time information about the charging stations. The EV drivers can find a publicly available charging station by searching for geographic location, address, connector type and more, using the appropriate website, smart phone applications, and navigation systems. The applications determine the station availability in real time, so drivers do not waste time driving to the stations they cannot use. They also can make a reservation to avoid going to an unavailable station. Specialized software packages allow the control and flexibility to optimize the performance of the EV's charging operations, to monitor and track the charging station usage and costs.

Charging points cloud provides charging statistics, which could be useful for fleet managers to inform them on which vehicle is ready to roll, and which one needs recharging. By creating a charging account, SMS and e-mail alerts for a vehicle fully charged, plug-out detection, ground fault detection and over-current detection could be set to alarm the user. All usage history for the public charging including start and end times, energy used (kWh), greenhouse gases saved, occupied time and location could also be presented to the user.

The charging stations have multiple payment options which depend on the owner. The charging stations allow flexible billing, by setting charging fees by time, session, or kWh, or any combination, and setting fees for different drivers or groups of drivers, or different times of the day. The EV drivers can also receive e-mails or messages that announce special offers to drivers such as discounts or coupons drive traffic to the stations [5].

VI. CHARGING STRATEGIES

Most of the excess electricity needed to recharge the EVs is nowadays being produced in coal plants, creating even more pollution than the average ICEV. Recharging a large battery pack presents a high load on the power system, but this can be scheduled for periods of reduced load or reduced electricity costs. In order to schedule the recharging, either the charging station or the vehicle should be able to communicate with the grid. A smart grid is an electrical grid that uses information and communication technologies (ICT) to gather and act on information, such as information about the behavior of suppliers and consumers, in an automated fashion to improve the efficiency, reliability, economics, and sustainability of the production and distribution of electricity. Furthermore, in a Vehicle-to-grid (V2G) scenario the vehicle battery can supply energy to the grid at periods of peak demand. This requires additional communication between the grid, charging station, and vehicle electronics. Society of Automotive Engineers (SAE) is an international association that is developing a range of standards for energy transfer to and from the grid [6].

In terms of the issues that still need to be solved; the main topics are availability of the infrastructure, battery technology and costs as well as education of customers. However, with the increasing popularity of the EVs worldwide, the charging system must also protect the grid and minimize its operating expenses. An efficient EV charging infrastructure must scale rapidly to accommodate the growing demand, integrating the EV charging into the smart grid operations, while monitoring and reporting on energy use, and actively manage energy load (demand/response), all in a single platform. A possible charging strategy, which is given in Fig. 2 [6]-[7], supposes charging of the vehicle during the night when the power system load is at its minimum so the EV has the lowest impact on the distribution network. The upper diagram represents the not-optimized charging during a week, where the charging period coincides with the peak demand periods. The lower diagram represents the optimized charging, where the charging process is optimized during the low energy demand periods. The advantages of the proposed concept are better control of the network, reduction of investments into the distribution network and optimal dispatching with minimization of charging time.

The smart grid technologies promise a major overhaul of the aging energy infrastructure. The changes will be realized with the addition of alternate, sustainable energy sources to smart meters in homes and businesses. However, an EV's requirements for energy could increase a household's electricity consumption up to 50 % or even more.

Standards represent a critical first step toward interoperability and seamless integration of the EVs into the grid. The EV owners can pay less for their charging if they follow the strategy of charging the vehicles in the periods of the reduced demand, at night when the costs of energy are less. Another example of the coordination

beetwen the EVs and the grid is that the energy used for overnight charging is during the day returned to the grid, to help the energy load peaks. This is possible if the owners connect their vehicles to the charging station during their working hours, so the energy from the battery could be taken. The batteries in this case represent a form of energy storage. This option is compatible only for those owners that can drive from their job to the home with an almost empty battery. In this case the EVs and the charging stations could significantly contribute to the modern grid operation [7]-[9].

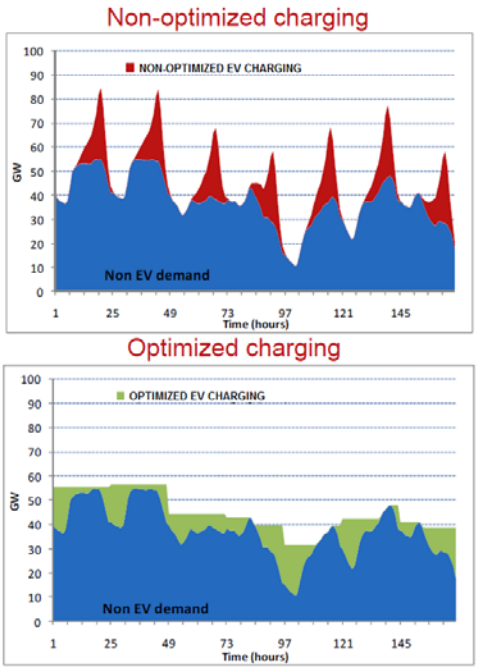


Figure 2. Charging optimization [7]

VII. Global Solutions

The most attractive solutions for the charging stations are those that use the energy from the RES, which significantly reduces the CO_2 emissions, as shown in Table I. This scenario assumes a large number of the RES integrated into the power system, which are intermittent by their nature and therefore introduce uncertainties in the power system operation. In order to maintain the integrity of the power system, the Wide Area Monitoring, Protection and Control (WAMPAC) systems are developed [10]-[11].

Regarding examples of the charging stations, a small wind powered charging station was built in Spain [12]. A solar charging station, built in front of one University in the USA, can plug twenty vehicles at the same time [13]. Estonia has the biggest charging station network in Europe [14], with more than 200 DC and 500 AC charging stations. Ireland promotes a healthy lifestyle by building green hotels and charging points for the EVs.

VIII. Croatian Reality

The idea of supporting electro-mobility as a new concept of urban mobility in Croatia emerged in 2010. The national integrated E-mobility Initiative was launched by the Energy Institute “Hrvoje Pozar” (EIHP) and first Croatian charging station for EVs was build in 2011 in Croatian capital Zagreb [15]. The charging station built in 2011 in Zadar was the first solar charging station in Croatia [16]. The Croatian power utility built the charging station in Labin [17], which is in touristic region Istria. Fig. 3 gives locations of the charging stations in Croatia. These charging points are not sufficient, because they only symbolically link all the parts of the country and the EVs with such a driving range are still not commercially available.

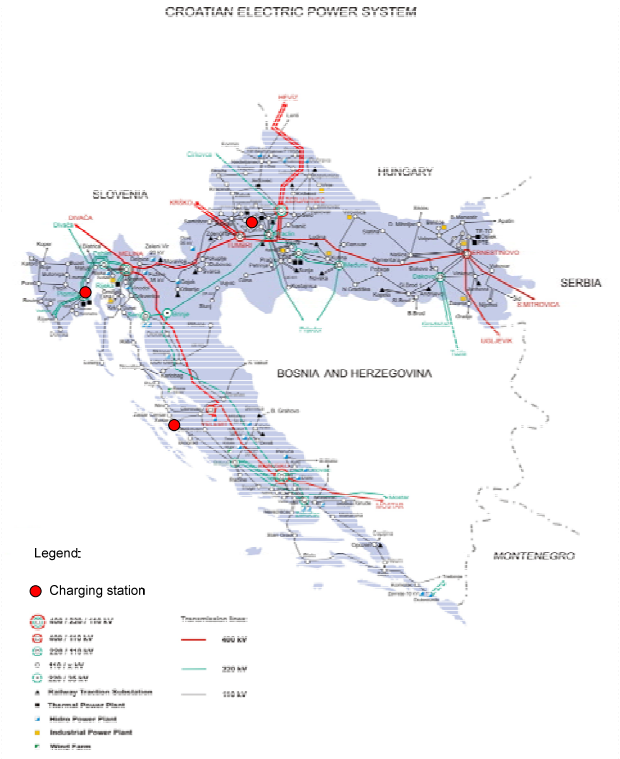


Figure 3. Locations of the charging stations in Croatia

IX. Student Electric Vehicle at the University of Rijeka

The students of the Faculty of Engineering, University of Rijeka in Croatia are working on the project of transforming the ICEV into the EV. Important part of the project is to design and build a charging station next to the Faculty building. Fig. 4 presents the Smart ForTwo model in the working phase [18]-[19].

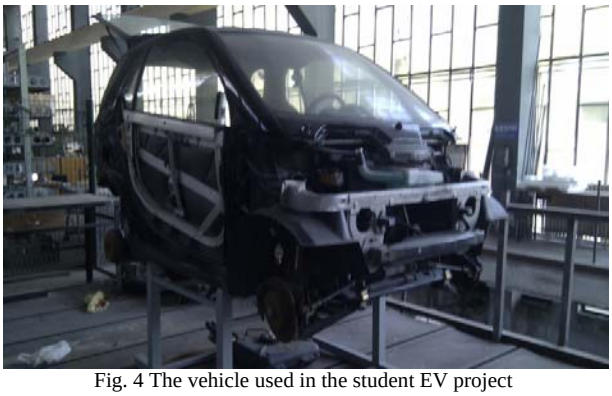


Fig. 4 The vehicle used in the student EV project

Regarding the charging of the EVs, there are two basic concepts. The first approach, which is chosen for the student project, is to build an infrastructure for charging/recharging the EV, while the second one supposes stations for the battery replacement, which is not so popular due to large investment costs.

The charging station that is planned to be built is the fast charging station of type 3, as given in Table II. This type of the charging station guarantees DC charging at higher power, which speeds up the recharging to last up to half an hour. Fig. 5 gives a schematic overview of the planned charging station.

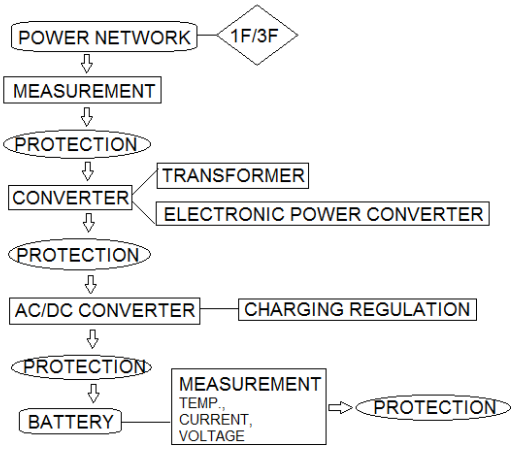


Figure 5. Charging station scheme

X. CONCLUSION

The automotive industry significantly contributes to the overall energy consumption. Regarding the rising trend of fossil fuel prices and the activities to reduce greenhouse gases emissions there is an open challenge in development of the EVs. The educational goal of the student EV project at the University of Rijeka, Faculty of engineering, is to support students in innovative thinking and application of modern technologies. The final goal of the project is to transform the ICEV to the EV and to build the charging station. The paper investigates the actual situation regarding the EVs and the charging stations. It can be a starting point for students from other universities to get involved in similar projects and design their EVs with the charging stations.

REFERENCES

[1] D. Radulovic; S. Skok; V. Kirincic; "Energy efficiency public lighting management in the cities," Energy, Elsevier, vol. 36, no. 4, pp.1908-1915, 2011.

[2] D. Radulovic; S. Skok; V. Kirincic; "Cogeneration – investment dilemma," Energy, Elsevier, vol. 48, no. 1, pp. 177-187, 2012.

[3] Green transport and energy. Available online: http://greentransportandenergy.blogspot.com/2009_05_01_archive.html [December 2012]

[4] T. Jakasa, S. Sucic: "Electric vehicles charging infrastructure," 10th Symposium HRO CIGRE, Cavtat, November 2011.

[5] Network of electric vehicle (EV) charging stations. Available online: http://www.chargepoint.com [14 th July 2012]

[6] Scholer, R., Maitra, A., Ornelas, E., Bourton, M. et al., "Communication between Plug-in Vehicles and the Utility Grid," SAE Technical Paper 2010-01-0837, 2010, doi:10.4271/2010-01-0837.

[7] http://bib.irb.hr/datoteka/532327.DOIE2011_Skrlec.pdf [January 2013].

[8] MERGE – Mobile Energy Resources in Grids of Electricity, FP7 project. Available online: http://www.ev-merge.eu/ [January 2013].

[9] G4V – Grid for Vehicles, FP7 project. Available online: http://www.g4v.eu/ [January 2013].

[10] P. Sasa, S. Skok, V. Kirincic, „ Urban electric vehicle development as part of Smart Grid", International Conference on Innovative Technologies (IN-TECH) Rijeka, 26 – 28 September 2012, Rijeka, Croatia.

[11] T. Sencic, V. Kirincic, K. Lenic, „Electric vehicle modelling for real application",The international Congress Energy and the Environment 2012, 17-18 October 2012, Rijeka, Croatia.

[12] UGE installs its first wind-powered EV charging station. Available online: http://www.gizmag.com/sanya-skypump-wind-powered-ev-charging-station/23738/ [16 th August 2012].

[13] Solar electric vehicle charging station. Available online: http://facilities.uiowa.edu/uem/renewable-energy/solar-energy.html [June 2012] .

[14] Electric Mobility Programme for Estonia. Available online: http://www.mkm.ee/electric-mobility-programme-for-estonia/ [January 2013].

[15] Available online : http://www.automotoportal.hr/2011/07/07/u-sklopu-e-mobilnosti-ispred-eihp-postavljena-prva-punionica-elektro-automobila-u-hrvatskoj/ [7 th July 2011].

[16] Available online: http://www.monitor.hr/clanci/u-zadru-otvorena-prva-solarna-punionica-elektricnih-automobila/160896/ [June 2012].

[17] Available online: http://www.istarski.hr/node/5686 [15 th November 2012].

[18] S. Skok, D. Brnobic, and V. Kirincic, "Croatian Academic Research Wide Area Monitoring System - CARWAMS", The International Journal on Communications Antenna and Propagation - IRECAP, vol. 1, no. 4, pp. 72-78, Aug. 2011.

[19] S. Skok, V. Kirincic, K. Frlan, "Dynamic Analysis of Wind Farm Operation Integrated in Power System Based on Synchronized Measurements," Engineering review, 30 (2010), no. 1, pp. 73-83, 2010.

Monitoring of low voltage facilities using synchronized phasor measurements

Darjan Bulic [1], Ranko Jancic [2], Vedran Kirincic [3], Srdjan Skok [4]
Department of Power Systems, Faculty of Engineering, University of Rijeka
Vukovarska 58, 51000 Rijeka. Croatia
[1] darjan.bulic@riteh.hr
[2] ranko.jancic@riteh.hr
[3] vedran.kirincic@riteh.hr
[4] srdjan.skok@riteh.hr

Abstract - Modern power systems are facing increase of overall energy consumption and aging of the main infrastructure. On the other hand, integration of intermittent renewable energy sources and development of the energy market are introducing uncertainties and strains to the power system operation, which is closer to stability limits. As a consequence, large power system blackouts occur. Synchronized phasor measurements have already been recognized as one of key technologies for developing Smart Grid applications. In recent years power utilities all over the world focused their attention to deployment of Phasor Measurement Units (PMUs) and development of Wide Area Monitoring, Protection and Control (WAMPAC) systems, which operate in real-time. This paper describes the student project of establishing a laboratory model of the Wide Area Monitoring System (WAMS) at the Faculty of Engineering, University of Rijeka. The developed model will be used for testing of various applications based on the synchronized phasor measurements, developed as a part of student projects and final thesis. Lessons learned from the low voltage laboratory model will be used to simulate larger scale power systems, which will be used for educational purpose of future generations of students.

Keywords: power system, laboratory model, Phasor Measurement Unit (PMU), synchronized phasor measurements, Wide Area Monitoring System (WAMS).

I. INTRODUCTION

Power systems are facing increase of consumer needs and overall energy consumption. Main infrastructure is aging and needs to be replaced with modern solutions. Integration of intermittent renewable energy sources (RES) introduces uncertainties into the power system operation, therefore development of modern power system monitor and control solutions is needed [1]. A Smart Transmission Grid is an advanced power system, which gathers technical concepts that enable active infrastructure capable of withstanding modern day problems and challenges in power systems.

Synchronized phasor measurements are recognized as one of fundamental technologies for building the Smart Grid [2]-[6]. Key feature of Wide Area Monitoring Systems (WAMS) is their ability to provide a complete overview of the power system in real time (refresh constant is about 200 ms), with an ability to provide common time stamp for each measurement, which offers a great advantage over commonly used SCADA (Supervisory Control And Data Acquisition) systems [7].

The idea of building a laboratory model of the WAMS at the Faculty of Engineering, University of Rijeka was conceived after the completion of the project named: „Wide Area Monitoring, Protection and Control of Power Systems in Deregulated and Liberalized Energy Market“ in 2009, and was included as an integral part of the project: „Intelligent Systems in Transmission Power Grid“, which began in 2010. The plan was to build a model of the power system which would then be monitored using synchronized phasor measurements. Project was co-founded by Croatian transmission system operator (TSO) and Croatian science foundation. This paper describes all the basic skills, materials and work required to build a laboratory model of the WAMS.

II. WAMS IN CROATIA

Implementation of the WAMS in the Croatian TSO began in 2003. Development of the WAMS started as a pilot project of monitoring the line between two 400 kV nodes Zerjavinec and Tumbri. Both nodes were selected due to their importance to the Croatian transmission power system and for their connection to power systems in neighboring countries. A Phasor Data Concentrator (PDC) was installed in the National Dispatching Center in Zagreb. In addition to standard PDC functions, two advanced applications were installed: Line thermal monitoring and Voltage stability monitoring [8].

First version of the WAMS installed in Croatia was biased towards research and was used to acquire basic knowledge about real time monitoring and control of power systems. Second stage of the WAMS implementation introduced three new Phasor Measurement Units (PMUs) installed at the remaining 400 kV nodes. With every 400 kV node being monitored, full observability was achieved for the 400 kV voltage level, which completed the second stage of the WAMS implementation in Croatia in 2007. Over the last few years a number of PMUs in Croatia was increased to 14 units installed at all the 400 kV voltage nodes and a part

of 220 kV voltage nodes [9]. The collected synchronized phasor measurements in Croatian transmission system are currently being used only for monitoring purposes, although automatic power system regulation is being developed for future use.

Work on the idea of the laboratory model of the WAMS started after positive experience and knowledge acquired during the project "Croatian Academic Research Wide Area Monitoring System – CARWAMS" [10]. The CARWAMS is intended to support small R&D teams in development of applications based on the synchronized measurement technology. The originally developed portable PMUs called WAMSTER were implemented in laboratories of 4 Croatian universities located in different transmission areas. Display of real time measurements is available online and is being used to analyze and compare the measurements with those from the WAMS installed in the National Dispatching center. Main benefit of this system is deployment simplicity and communication infrastructure independence, as the measured data is sent by using a General Packet Radio Service (GPRS) protocol.

III. Overview of Laboratory WAMS Model

Initial setup of the model included two PMUs connected to the PDC, which will also be installed in the laboratory. Following donations from various sponsors made realization of the WAMS laboratory possible:

- Schweitzer Engineering Laboratories, Inc:
  - Three PMUS SEL 421 (devices also function as distance relays)
  - SEL – Synchrowave Software package
- Croatian TSO:
  - Four PMU-s Arbiter 1133A
  - GPS antennas with accessories
- Croatian TSO – Transmission area Rijeka:
  - Rest of the materials and supplies required for the laboratory

Since the number of the PMUs was too large for laboratory purposes, two PMUs SEL 421 and two PMUs Arbiter 1133A were inspected for their complete functionality, and after some minor repairs the devices were installed in the Croatian transmission system. The laboratory WAMS model architecture includes.

- PMUs
- PDC which will collect data from all PMUs installed in faculty building
- Switchboards used to connect PMUs to the faculty network

Following chapters give description of all the elements.

A. *Phasor Measurement Units*

The PMUs are devices that measure electric waves in the electric grid. Measurements are done in real time, and have a common time source which allows for synchronized measurements made in multiple remote points in the power network. All the PMUs have to be connected to the Global Positioning System (GPS) that supplies the common time source. Measurement output consist of voltage and current amplitudes and angles in phasor form – representation of a sinusoidal function with amplitude (A), frequency (ω) and phase angle (Θ). As shown in Fig. 1 [11], inputs to PMU are analog signals which are converted into digital form. Some proposed functions of the PMU include monitoring, alarming and control operations [12].

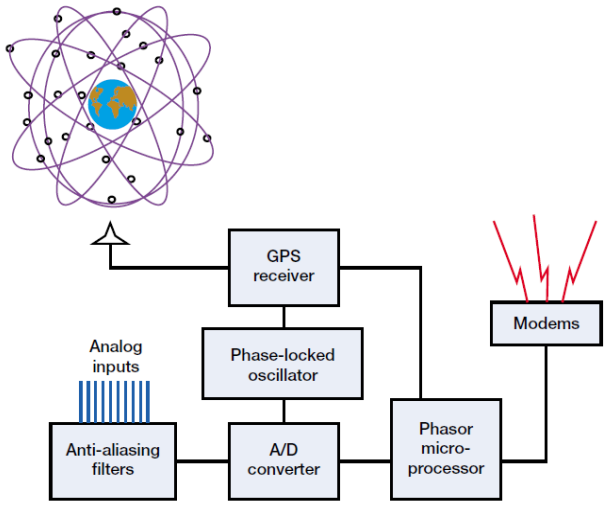


Figure 1. Block diagram representation of PMU [11]

B. *Phasor data concentrator*

A PDC collects and processes all the measured data received from the PMU. With use of appropriate algorithms and applications a snapshot of the power grid state is obtained in real time [13]. The PDC has additional monitoring and display functions which in our case will be used to display data on several computers installed in the laboratory.

C. *Switchboard*

The function of the switchboard is to transform current and voltage measurements into appropriate levels which will not damage the devices. It is important to note that in low voltage environment some PMU models can gather data directly from the voltage buses without signal transformation on the switchboard, and checkup of device inputs is recommended before installation of current and voltage transformers. The switchboard consists of three current transformers (one transformer per each phase) and a three phase voltage transformer. Secondary values of the current and voltage transformers are In=5 A and Un=63 V respectively, which is well below factory recommended PMU limits. The transformer outputs are connected to the current and voltage inputs of the PMUs. The configuration with three current and one voltage transformer offers full three phase current and voltage observability, although other options with fewer current transformers are also possible.

IV. IMPLEMENTATION OF LABORATORY WAMS MODEL

This chapter describes the actual laboratory implementation at the Faculty of Engineering Rijeka. First step is to acquire all the parts and items needed for proper functioning of the power system model. List of basic parts includes:

- One or more PMUs,
- GPS antennas equal to the number of PMUs,
- Voltage and current transformers,
- Circuit breakers,
- One computer which will be used as the PDC (computer must be equipped with an Ethernet card),
- Additional computers (for data displaying purposes - optional),
- Mounting plate or electrical cabinet,
- Mounting rack for the PMUs (optional),

The highest investment in this list are the PMUs, with price several times higher than the rest of the parts combined, so tracking of various sponsors is recommended when equipping a laboratory for educational purposes. After acquiring all the necessary parts, next step is to choose the laboratory location, with several factors which should be considered:

- Location in the power network which will offer the best operability of the monitored power system.
- Availability of the three phase connection - connecting to electrical cabinet is recommended, although there are some PMUs that can be connected directly to the power socket, e.g. portable PMU solutions such as WAMSTER.
- Access to the roof of the building or windows for the GPS connection.

After choosing the most suitable available location for the laboratory, the next phase is the PMUs installation and connection to the power network. In the WAMS laboratory at the Faculty of Engineering, University of Rijeka, the following 4 PMUs are used: 2 Arbiter 1133A units, 1 SEL 421 unit and 1 WAMSTER unit. All the PMUs were installed in the laboratory, except for one Arbiter 1133A unit which will be installed on the main power input node near the faculty entrance. The Arbiter 1133A requires the current and voltage transformers to transform input signals to level that will not damage the device. The SEL 421 has dual functions: as a PMU and as a distance relay and is capable of withstanding higher voltages and currents. The WAMSTER as a most versatile device from the group can be connected directly to the power socket and requires no additional work on connection. The Arbiter 1133A was connected to the switchboard that we built according to the description in the previous chapter. The switchboard was installed on the wall near the installing point of the PMUs to minimize length of the cables. The current and voltage transformer configuration used on the switchboard is shown in Fig. 2. It is important to note that the transformers are connected to the high voltage, and extreme caution should be used when connecting the switchboard to the power system.

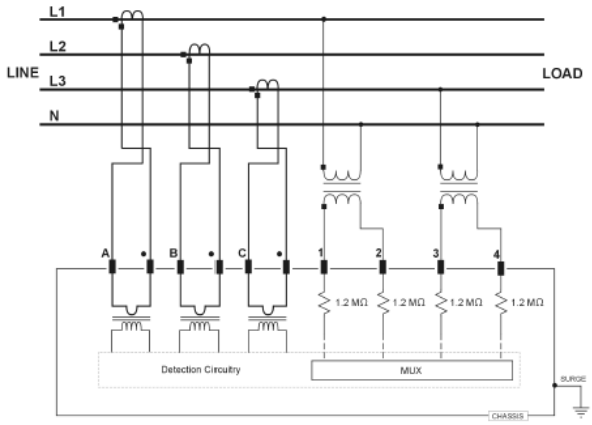


Figure 2. Transformer configuration on the switchboard

The PMUs will not function properly without the GPS connection, meaning that we cannot setup the laboratory in a room that has no access to the roof. Alternatively, it is possible to use windows but prior testing for GPS satellites visibility is strongly recommended. For complete coverage, the antenna needs to have a clear view of the sky from 15 degrees above the horizon to directly overhead for all points of the compass. Although minimal installations where the antenna is mounted in a less favorable location may work, however reception may be somewhat limited during certain hours of the day. It is possible to simply use a longer coaxial cable to connect the antennas mounted on the roof, but longer cables bring further expenses and greater possibility for cable malfunction (in our case coaxial cables have proven to be somewhat fragile). Since our laboratory room does not have direct access to the roof, the GPS antennas were mounted on the faculty roof as shown in Fig. 3 and then connected to the PMUs through the window.

Figure 3. Antenna placement

After the completion of the initial phase it was clear that the SEL 421 and the Arbiter 1133A units represent backbone of the laboratory WAMS model, which is given in Fig. 4. Being a portable PMU solution, we mostly used the WAMSTER unit as a supplementary PMU for redundancy check of other three PMU and various testing situations on different locations in the faculty power system. As the SEL 421 also has distance protection functions it offers full protection and control for two breakers. Implementation of this function is being developed for future use and could be used in conjunction with relay inputs on the Arbiter 1133A to achieve full observability and control over two breakers.

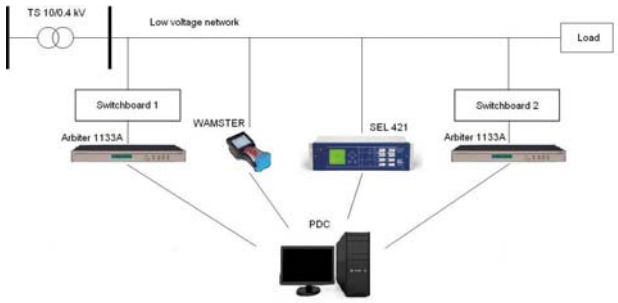


Figure 4. WAMS model used in the laboratory

All the PMU units are connected to the PDC (Fig. 5) by Ethernet connection using LAN protocol, except the WAMSTER unit which uses GPRS as primary connection when using real time measurements and flash memory for backup and later analysis. Since each PMU comes with its own distributed software one of the main problems when setting up the laboratory model is to display all the measurements on a single platform. Several computers connected to the main PDC using LAN are equipped with appropriate software and are capable of displaying measurements received from the PMUs. The developed laboratory WAMS model will be used in laboratory exercises at the Faculty of Engineering, University of Rijeka, and will give students opportunity to familiarize themselves with one of the leading technologies and concepts used in modern power systems.

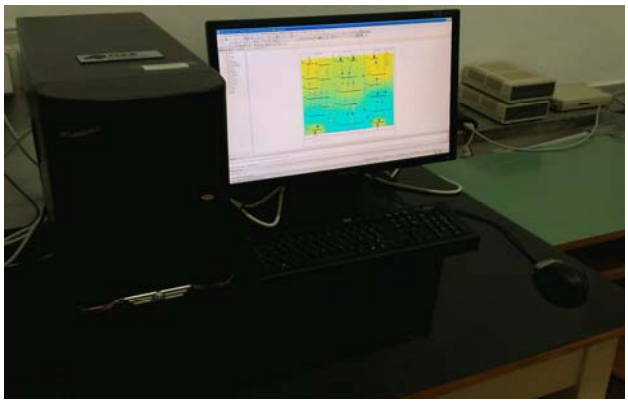


Figure 5. PDC installed in the laboratory

V. CONCLUSION

The paper describes development of the WAMS laboratory model at the Faculty of Engineering, University of Rijeka. The synchronized phasor measurements are collected with several PMUs, which offers opportunity for students to acquire working experience with devices made by some of the leading manufacturers in the world. Using the PMU measurements as inputs for future applications, the students will have ability to observe and interact with real power system. As it is expected that similar laboratory models will be developed at the universities all over the world, a detailed description given in the paper presents a thorough insight and guidance into the process of building the laboratory WAMS model.

REFERENCES

[1] S. Skok, V. Kirincic, K. Frlan, "Dynamic Analysis of Wind Farm Operation Integrated in Power System Based on Synchronized Measurements," Engineering review, vol. 30, no. 1, pp. 73-83, 2010.

[2] A. G. Phadke, J. S. Thorp, "Synchronized phasor measurements and their applications", Springer, 2008.

[3] A. G. Phadke, "Synchronized phasor measurements - a historical overview," Transmission and Distribution Conference and Exhibition 2002: Asia Pacific. IEEE/PES, pp. 476- 479, 6-10 October 2002.

[4] A. G. Phadke, J. S. Thorp, "History and Applications of Phasor Measurements", Proceedingss of the IEEE Power Systems Conference and Exposition, pp. 331 – 335, 29 Oct. – 1 Nov. 2006.

[5] IEEE Standard for Synchrophasors for Power Systems, IEEE Standard C37.118-2005 (Revision of IEEE Standard 1344-1995), 2006.

[6] V. Terzija, G. Valverde, D. Cai, P. Regulski, V. Madani, J. Fitch, S. Skok, M. Begovic, A. G. Phadke, "Wide Area Monitoring, Protection and Control of Future Electric Power Networks", Proceedings of IEEE, vol. 99, no. 1, pp. 80-93, 2011.

[7] V. Kirincic, S. Skok, I. Pavic, „Power system state estimation based on PMU measurements vs SCADA measurements," International Review on Modelling and Simulations - IREMOS, accepted for publication , ISSN 1974-9821, 2012.

[8] S. Skok, R. Matica, I. Sturlic, „Wide Area Monitoring Enhanced Architecture Opened for Multiple Phasor Data Concentrators", European Transactions on Electrical Power, Volume 21, Issue 4, pages 1531–1540, May 2011

[9] V. Kirincic, „Synchronized phasor measurements as a function of advanced power networks guidance", PhD qualifying exam, Faculty of Electrical Engineering and Computing, University of Zagreb 2011.

[10] S. Skok, D. Brnobic, and V. Kirincic, "Croatian Academic Research Wide Area Monitoring System - CARWAMS", The International Journal on Communications Antenna and Propagation - IRECAP, vol. 1, no. 4, pp. 72-78, Aug. 2011.

[11] R. F. Nuqui, „State Estimation and Voltage Security Monitoring Using Synchronized Phasor Measurements," PhD dissertation, Virginia Polytechnic Institute and State University, 2001.

[12] Skok, S.; Ivankovic, I.; Cerina, Z.: "Applications Based on PMU Technology for Improved Power System Utilization," Power Engineering Society General Meeting, 2007. IEEE, 2007

[13] J. Thorp, A. Abur, M. Begovic, J. Giri, R. Avila-Rosales, „Gaining a Wider Perspective," Power and Energy Magazine, IEEE, vol. 6, no. 5, pp. 43-51, 2008.

Studentski projekt razvoja vozila na električni pogon kao dijela elektroenergetskog sustava budućnosti

Predrag Sasa [1], Vedran Kirincic [2], Srdjan Skok [3]
Zavod za elektroenergetiku, Tehnički fakultet, Sveučilište u Rijeci
Vukovarska 58, 51000 Rijeka. Hrvatska
[1] predrag.sasa@riteh.hr
[2] vedran.kirincic@riteh.hr
[3] srdjan.skok@riteh.hr

Sažetak - U radu je opisan studentski projekt koji obuhvaća izradu vozila na električni pogon, konverzijom postojećeg vozila pogonjenog motorom s unutrašnjim sagorijevanjem. Primjena obnovljivih izvora energije (OIE), u konkretnom slučaju fotonaponskih ćelija na krovu vozila, omogućuje dopunu punjenju baterija. U projektu sudjeluju studenti završnih godina elektrotehničkog usmjerenja, koji na taj način imaju priliku primijeniti znanja stečena dosadašnjom fakultetskom izobrazbom. Obzirom na blisku suradnju studenata i nastavnog osoblja na Tehničkom fakultetu Sveučilišta u Rijeci, stečena se znanja i iskustva mogu učinkovito prenijeti budućim generacijama kroz nastavni proces. Time će se motivirati mlađe generacije studenata za nastavak rada na tematici električnih vozila, čime će se stečeno znanje akumulirati. Bavljenje problematikom električnih vozila omogućuje studentima uključenim u projekt rad na području koje je aktualno na svjetskoj razini, obzirom da se sve više poznatih proizvođača odlučuje na slična rješenja. S druge strane, otvara se prostor za razvoj rješenja punionice takvih vozila, odnosno distribuiranih izvora električne energije, kao dio naprednih elektroenergetskih mreža budućnosti.

Ključne riječi: električna vozila, napredne elektroenergetske mreže, obnovljivi izvori energije.

I. UVOD

Pored značajnije izgradnje i integracije obnovljivih izvora energije (OIE) u elektroenergetski sustav (EES), zatim proizvodnje energije iz visokoučinkovitih izvora poput kogenercijskih postrojenja te općenito provođenja mjera energetske učinkovitosti [1]-[2], bitnu ulogu u izgradnji naprednih elektroenergetskih mreža budućnosti (eng. Smart Grid) ima tehnologija vozila na električni pogon (eng. Electric Vehicle – EV) [3]. Nakon niza desetljeća dominacije vozila koja koriste motore s unutrašnjim sagorijevanjem (eng. Internal Combustion Engine Vehicle - ICEV) dolazi do promjene u načinu razmišljanja proizvođača i potrošača [4]. Naime, uslijed rasta cijena fosilnih goriva, klimatskih promjena i širenja svijesti o važnosti zaštite okoliša i održivom razvoju, autoindustrija današnjice postupno kreće ka tehnologiji EV [5], odnosno popularizaciji EV i brige za okoliš kao stila života, što je vidljivo kroz organizaciju natjecanja s jakom medijskom vidljivošću [6]. S vremenom su do izražaja došle tri osnovne izvedbe:

1) Hybrid Electric Vehicle (HEV) – vozilo koristi motor s unutrašnjim sagorijevanjem u kombinaciji s električnim motorom, nema mogućnost priključka na EES već se sva energija dobiva iz fosilnih goriva,

2) Plug-in Hybrid Electric Vehicle (PHEV) – vozilo je moguće priključiti na EES te se energija može pohraniti u baterije uz korištenje energije fosilnih goriva,

3) Battery Electric Vehicle (BEV) - energija se preuzima iz EES-a i pohranjuje u baterije.

Uvođenje tehnologije EV, odnosno većeg broja vozila i pripadnih punionica, ima niz prednosti u usporedbi s primjenom ICEV:

- smanjenje ovisnosti o fosilnim gorivima,
- smanjenje emisije staklenički plinova i lokalnog onečišćenja, naročito ukoliko je dio energije za EV proizveden korištenjem OIE,
- smanjenje buke u urbanim sredinama,
- relativno niske investicije u infrastrukturu za punjenje, obzirom na široku rasprostranjenost elektroenergetske mreže,
- niži troškovi održavanja zbog jednostavnije konstrukcije.

S druge strane, obzirom da se radi o još relativno novoj i komercijalno nedovoljno zastupljenoj tehnologiji, postoji više faktora koje je potrebno uzeti u obzir:

- nedostatak adekvatne infrastrukture za punjenje vozila,
- tehnološka ograničenost baterijskog sustava – ograničenje dometa vozila, vrijeme punjenja, masa i životni vijek baterija,
- relativno visoka cijena,
- stav naftne industrije prema tehnologiji EV.

U ovom je radu opisan studentski projekt na Tehničkom fakultetu Sveučilišta u Rijeci, koji okuplja studente elektrotehničkog usmjerenja. Cilj projekta je

izrada vozila na električni pogon i pripadne punionice, konverzijom postojećeg ICEV u EV. Projekt ima edukacijsku komponentu, obzirom da potiče razvoj projektne i poduzetničke kulture u studentskoj populaciji. Pregled projekta prikazan u radu predstavlja motivaciju za studente na drugim fakultetima prilikom promišljanja o rješenjima za izradu vlastitog EV.

II. PROJEKT IZRADE ELEKTRIČNOG VOZILA

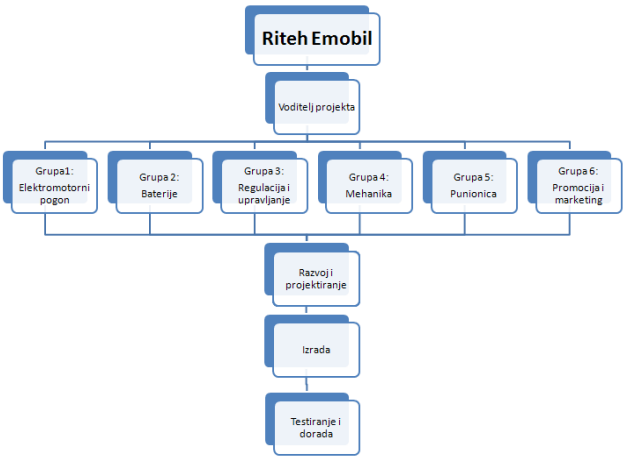
Osnovna ideja vodilja projekta je aktivni rad studenata na realnom problemu izrade vozila na električni pogon s pripadnom punionicom. U projektu sudjeluju studenti završnih godina, koji na taj način imaju priliku primijeniti znanja stečena dosadašnjom fakultetskom izobrazbom. Problemi na koje studenti nailaze tijekom različitih faza provedbe projekta omogućuje im stjecanje novih znanja i vještina te ih motivira za timski, ali i samostalan rad. Naglasak je pritom na tehničkih znanjima te inovativnosti i kreativnosti u pronalaženju rješenja. Svatko od sudionika na ima priliku pronaći područje rada koje mu najviše odgovara, razvijajući time vještine poput organizacije rada, vođenja većeg ili manjeg tima osoba, komunikacijske i prezentacijske vještine prilikom promocije projekta i slično. Bitan dio rada na projektu predstavlja i prikupljanje financijske podrške te prijava na natječaje za sufinanciranje, odnosno izrada cjelokupne popratne dokumentacije. Studenti radom na projektu dobivaju novu dimenziju obrazovanja, kakvu nemaju priliku steći u postojećem sustavu.

Projekt vozila na električni pogon i pripadne punionice ima posebnu važnost, jer prvi puta na Tehničkom fakultetu okuplja veći broj studenata elektrotehničkog usmjerenja. Obzirom na prirodu problema, također postoji mogućnost uključivanja ostalih studenata Sveučilišta u Rijeci, zainteresiranih za različite aspekte vozila na električni pogon.

Obzirom na blisku suradnju studenata i nastavnog osoblja na Tehničkom fakultetu, stečena se znanja i iskustva mogu učinkovito prenijeti budućim generacijama kroz nastavni proces. Time će se motivirati mlađe generacije studenata za nastavak rada na projektu, čime se stečeno znanje akumulirati. Treba istaknuti kvalitetniju edukaciju, stjecanja novih znanja i vještina kao i mogućnost kreativnog izražaja pojedinog studenta u području njegovog interesa. Zadaci na projektu su strukturirani, što dovodi do stjecanja organizacijskih sposobnosti kao i sudjelovanje u timskom radu.

Primarni korisnici projekta bit će studenti Tehničkog fakulteta Sveučilišta u Rijeci. Osim studenata direktno uključenih u projekt izrade električnog automobila, vozilom će se moći služiti i ostali djelatnici fakulteta. Obzirom da će se vozilo koristiti u edukacijske svrhe, buduće će generacije moći raditi na novim varijantama, promišljajući inovativna rješenja već pri prvim danima dolaska na fakultet. Odlaskom na studentska natjecanja i izložbe studenti će uspostaviti međunarodne kontakte te steći uvid u ostvarenja ostalih timova. Vozilo će biti u funkciji promocije Tehničkog fakulteta i Sveučilišta u Rijeci te će se koristiti kao službeno vozilo Tehničkog fakulteta, prikladno za transport i prezentacije. Sekundarni korisnici ovog projekta će biti sponzori, tvrtke i pojedinci koji su pripomogli razvoju projekta. Pod tercijarnim korisnicima podrazumijeva se šira javnosti jer će se projektom izrade električnog automobila upoznati s mogućnostima razvoja izrade vozila na električni pogon, kao što će se istovremeno podići svijest građana o predmetnoj tematici.

Slikom 1. prikazana je organizacijska struktura projekta, gdje su vidljive cjeline prema kojima su određeni timovi studenata te njihova međusobna povezanost [7]:



Slika 1. Organizacijska struktura studentskog projekta

Realizacija projekta i njegov dovršetak iziskuje određena novčana sredstva. U samom početku projekta pristupilo se izradi financijskog plana i troškovnika. Troškovnikom su obuhvaćeni troškovi provedbe aktivnosti, troškovi opreme, troškovi komunikacije kao i putni troškovi. Nakon izrade troškovnika pristupilo se prikupljanju ponuda te prikupljanju novčanih sredstava putem natječaja i donacija. Projekt je u skladu s trendovima smanjenja potrošnje primarnih izvora energije, povećanja udjela OIE, povećanja energetske učinkovitosti i smanjenja emisije stakleničkih plinova. EU je donijela direktivu o promociji korištenja OIE 2009/28 [8], koja razrađuje ciljeve „20-20-20 do 2020.“, kojom se želi povećati udio OIE na 20 %, smanjiti ukupna potrošnja za 20 % upotrebom mjera energetske učinkovitosti i smanjiti emisije CO_2 za 20 %. Stoga je tematika projekta pogodna za korištenje pristupnih europskih fondova, čime bi se ostvarili preduvjeti za nastavak rada budućih generacija studenata na projektu.

Obzirom na opseg aktivnosti i broj sudionika, određeno je više faza projekta kako bi se pojednostavnilo upravljanje i nadzor nad provedbom. U nastavku je dan popis glavnih faza projekta [7]:

1. Definiranje polaznih točaka, određivanje problematike pojedinih cjelina projekta, prikupljanje literature i informacija o sličnim projektima, stupanje u kontakt s osobama koje su radile na sličnim projektima.
2. Projektiranje, dokumentiranje i iznalaženje tehničkih rješenja te izrada projektne dokumentacije.
3. Nabavka vozila, odnosno odabir najprikladnijeg vozila za konverziju, skupljanje ponuda i realizacija kupovine.

4. Mehanički radovi na vozilu i priprema – izvođenje radova na vozilu u radionici. Faza prvenstveno obuhvaća demontažu nepotrebnih sustava i dijelova kao i zamjenu dotrajalih i neispravnih dijelova vozila koji ostaju u funkciji.
5. Nabavka potrebnih dijelova i komponenata – prikupljanje ponuda i realizacija kupnje.
6. Slaganje vozila u cjelinu – ugradnja novih sustava pogona, baterijskog sustava i pripadajućih komponenti.
7. Izrada punionice – izvođenje radova u radionici.
8. Testiranje vozila – testiranje i kontrola komponenata i sustava u cjelini kada je vozilo u potpunosti završeno.
9. Ispitivanje, homologacija i tehnički pregled vozila – prikupljanje potrebne dokumentacije, obrazaca, proučavanje zakonske regulative te sukladno tome provedba aktivnosti potrebnih za uključivanje vozila u promet.

III. OPIS VOZILA

Odabir vozila za konverziju na električni pogon jedna je od prvih točaka razrade i realizacije projekta. Postoji nekoliko kriterija na koje pritom treba obratiti pažnju. Kako se projekt temelji na vozilu koje će zadovoljiti gradske potrebe korisnika, potrebno je uzeti u obzir vanjske dimenzije i masu istog. Upravo je masa vozila jedan od glavnih kriterija koji određuje snagu električnog motora, kapacitet i masu baterija, vrstu upravljanja te samim time performanse i radijus kretanja vozila. Nadalje, potrebno je obratiti pažnju na raspoloživi prostor za smještaj propulzijskog električnog sustava kao i uklanjanje nepotrebnih dijelova i komponenata postojećeg pogonskog motora s unutrašnjim sagorijevanjem. Potrošače električne energije koji ostaju na vozilu potrebno je dodatno prilagoditi i učiniti ekonomičnijim kako bi se smanjila cjelokupna potrošnja energije. Vozilo odabrano za projekt konverzije je model Smart ForTwo serije 451, čiji su originalni podaci, kao i očekivane karakteristike nakon konverzija prikazani Tablicom I.

TABLICA I. TEHNIČKI PODACI VOZILA

Podatak	**Prije konverzije**	**Poslije konverzije**
Vrsta	osobno vozilo	
Marka	SMART	
Model	ForTwo	
Masa praznog vozila [kg]	730	~ 730
Maksimalna brzina [km/h]	135	~ 80
Vrsta motora	Otto motor 4 takta	Sinkroni električni motor s permanentnim magnetima
Snaga motora [kW]	37	19
Brzina vrtnje motora [min^{-1}]	5250	3600
Radni obujam motora [cm^2]	698	-
Potrošnja	6 l/100 km	~ 15 kWh/100 km

Korištenjem podataka o potrošnji iz Tablice I te uz pretpostavku cijene goriva 10,34 kn/l [9], odnosno električne energije 0,56 kn/kWh [10], moguće je odrediti prosječnu cijenu vožnje prije i nakon konverzije, kako navodi Tablica II.

TABLICA II. PROSJEČNA CIJENA VOŽNJE

Vozilo	Potrošnja	Cijena energenta	Cijena vožnje
ICEV	6 l / 100 km	10,34 kn/l	62,04 kn / 100 km
EV	15 kWh / 100 km	0,56 kn/kWh	8,40 kn / 100 km

Iako prilikom proračuna nisu uzeti u obzir gubici prilikom konverzije energije, kao i postupno smanjenje kapaciteta baterije, jasno je da EV značajno pridonose smanjenju potrošnje primarnih izvora energije te da postoji veliki potencijal za rast njihovog udjela u ukupnom broju vozila na tržištu [11]. Navedenom u prilog ide i smanjenje emisija CO_2 [12]. Razmatrajući potencijalne uštede prilikom usporedbe EV i ICEV, treba istaknuti da je u krajnjoj cijeni goriva za ICEV u Republici Hrvatskoj sadržano niz naknada i trošarina [13], dok je cijena električne energije za EV zasad bez navedenih dodataka. Stoga u budućim detaljnijim analizama i usporedbama navedeno treba uzeti u obzir.

Prilikom izrade vozila na električni pogon postoji više varijanti pogonskog sustava uključujući vrstu električnog motora, vrstu baterija, način upravljanja sustavom (mogućnosti regulatora) te vrstu i način punjenja baterija. Financijske mogućnosti uvelike pridonose odluci o izboru pogonskog sustava kao i njegovoj namjeni.

Nakon pripremnih radnji uklanjanja motora s unutrašnjim sagorijevanjem te pripadajućih uređaja i opreme (rezervoar goriva, ispušni sustav, rashladni sustav) iz vozila, potrebno je naći optimalna rješenja za smještaj električnog motora i svih pripadnih komponenata.

Posebnu pažnju treba obratiti na izvedbu spoja pogonskog vratila električnog motora s mjenjačkom kutijom koja u ovom slučaju ostaje na vozilu radi jednostavnosti izvedbe. Projektom je potrebno obuhvatiti kompletno ožičenje sustava, pravilan smještaj regulatora obzirom na potrebno hlađenje te sigurnosne i tehničke propise.

Obzirom na cijenu i važnost u radu vozila, baterijski sustav predstavlja jednu od bitnijih cjelina [14]. Kako bi se zadovoljile potrebe energije dostatne za pokretanje i dovoljan radijus kretanja vozila na kratkim i srednjim relacijama, za projekt su odabrane LiFePo4 baterije, koje će biti ugrađene u podnicu vozila, Slika 2. U usporedbi s ostalim tehnologijama, baterije na bazi litija imaju visok elektrokemijski potencijal, što omogućuje visoku specifičnu snagu baterije.

Slika 2. Model vozila s prikazom smještaja baterija

Primjena OIE u studentskom projektu izrade električnog vozila je zamišljena kroz smještaj fotonaponske folije na krovu vozila, Slika 3. Obzirom da tako zamišljeno rješenje nije dostatno za punjenje baterija korištenih za pogon vozila, energija dobivena na ovaj način bi služila kao dopuna sekundarnog sustava napajanja odnosno baterijske grupe koja je namijenjena kao izvor električne energije za manje kritične potrošače, poput radija, sustava navigacije i slično. Prednost fotonaponskih folija je njihova fleksibilnost, koja omogućuje prilagodbu zakrivljenim površinama kao što je krov vozila.

Slika 3. Model vozila s prikazom smještaja fotonaponske folije

IV. IZRAČUN SNAGE MOTORA

Prilikom odabira električnog motora potrebno je provesti niz proračuna, kako bi se uzele u obzir karakteristike vozila odabranog za konverziju, kao i željene performanse nakon prerade [15]-[18].

Koristeći drugi Newtonov zakon možemo pisati relaciju za primjer EV:

$$F_i = C_i \cdot m \cdot a \tag{1}$$

gdje je:

C_i – faktor dodatne inercije rotacijskih dijelova (za male automobile $C_i = 1{,}06$),

m – masa vozila,

a – ubrzanje, odnosi se na ubrzanje do 100 km/h (u ovom slučaju je odabrano $a = 2{,}77$ m/s^2),

Uvrštavanjem navedenih vrijednosti i mase vozila, dobivamo:

$$F_i = 1{,}06 \cdot 730 \cdot 2{,}77 = 2143{,}4 \text{ N} \tag{2}$$

Kreće li se vozilo uzbrdo potrebno je dodati silu:

$$F_h = F_g \cdot \sin\alpha \tag{3}$$

gdje je:

F_g – težina vozila,

α – kut kosine kojim se vozilo kreće,

Uvrštavanjem vrijednosti (uz pretpostavku $\alpha = 10°$) dobivamo:

$$F_h = 730 \cdot 9{,}81 \cdot \sin 10° = 1243{,}54 \text{ N} \tag{4}$$

Potrebno je u obzir uzeti i aerodinamičku silu trenja:

$$F_d = \frac{C_d \cdot A \cdot v^2 \cdot \rho}{2} \tag{5}$$

gdje je:

C_d – koeficijent trenja odabranog vozila,

A – efektivna površina prednjeg dijela vozila [m^2],

v – željena maksimalna brzina vozila [m/s],

ρ – gustoća zraka.

Za odabrano vozilo može se pisati:

$$F_d = \frac{0{,}38 \cdot 2{,}265 \cdot 22{,}22^2 \cdot 1{,}225}{2} = 260{,}28 \text{ N} \tag{6}$$

Doprinos sile vjetra na vozilo je dan izrazom:

$$F_w = F_d \cdot C_w \tag{7}$$

gdje je:

C_w – relativni koeficijent doprinosa vjetra (za prosječne vrijednosti brzine vjetra vrijedi $C_w = 1,4$).

Tada se može pisati:

$$F_w = 260,28 \cdot 1,4 = 364,39 \text{ N} \tag{8}$$

Zatim je potrebno izračunati silu trenja kotrljanja:

$$F_r = F_g \cdot C_r = 730 \cdot 9,81 \cdot 0,015 \cdot = 107,42 \text{ N} \tag{9}$$

Za željene podatke proračun se može nastaviti na temelju dobivenih sila, te je ukupna sila:

$$F = \sum_{j=1}^{n} F_j = 4119,03 \text{ N} \tag{10}$$

Proračun potrebne snage motora se provodi uz pretpostavku konstantne brzine i gibanje po horizontalnoj podlozi pa se sile F_i i F_h u tom slučaju ne uzimaju u obzir što daje ukupnu silu:

$$F = \sum_{j=1}^{n} F_j = 732,09 \text{ N} \tag{11}$$

Uzimajući u obzir radijus kotača odabranog vozila (r = 0,286 m) moguće je izračunati moment na kotačima:

$$M_{kot} = F \cdot r = 209,37 \text{ Nm} \tag{12}$$

Poznavajući moment može se napisati relacija za snagu motora:

$$P = \frac{M_{kot} \cdot n \cdot 2 \cdot \pi \cdot \eta}{60} \tag{13}$$

Uz odabir brzine vrtnje kotača n = 740 min^{-1} i korisnost η = 90 % dobivamo:

$$P = \frac{209,37 \cdot 740 \cdot 2 \cdot \pi \cdot 0,9}{60} = 14594,76 \text{ W} \tag{14}$$

Za odabrano vozilo smo dobili minimalnu potrebnu snagu motora od 14,6 kW kako bismo ostvarili željene karakteristike [19].

V. ZAKLJUČAK

Tematika vozila na električni pogon dobiva sve veći značaj obzirom na niz čimbenika, poput povećanja energetskih potreba i rasta cijene fosilnih goriva, kao i težnje ka smanjenju emisija stakleničkih plinova te promišljanja o održivom razvoju. Opravdanost izrade vozila na električni pogon u sklopu studentskog projekta potvrđuje niz modela hibridnih ili u potpunosti električnih vozila poznatih svjetskih proizvođača.

Očekivani rezultat projekta je izrada tehnički ispravnog i registriranog vozila s ugrađenim električnim pogonom. Konverzija vozila koje koristi motor s unutrašnjim sagorijevanjem u vozilo s električnim pogonom će se ostvariti ugradnjom sinkronog motora s permanentnim magnetima, u sprezi s postojećim automatskim mjenjačem te baterijskim sustavom. U planu je postavljanje fotonaponske folije na krov vozila, kako bi se osiguralo napajanje manje zahtjevnih sustava i time produžila autonomija malog gradskog vozila

Projekt ima važnu edukativnu komponentu, obzirom da studentima omogućuje rad na realnom problemu te omogućava stjecanje niza znanja i vještina. Pored promocije tehnologije unutar studentske populacije, cilj projekta je razvoj svijesti o brizi za okoliš te održivom razvoju šireg društva. Opisani projekt predstavlja pilot-projekt te omogućuje povezivanje studenata s različitih fakulteta, zainteresiranih za istraživanje niza aspekata izrade EV te njihovog utjecaja na razvoj društva općenito.

LITERATURA

[1] D. Radulovic, S. Skok, V. Kirincic, "Energy efficiency public lighting management in the cities," Energy, Elsevier, Vol. 36, No. 4, pp. 1908-1915, 2011.

[2] D. Radulovic, S. Skok, V. Kirincic, "Cogeneration – investment dilemma," Energy, Elsevier, Vol. 48, No. 1; pp. 177-187, 2012.

[3] T. Jakasa, S. Sucic: "Electric vehicles charging infrastructure," 10th Symposium HRO CIGRE, Cavtat, November 2011.

[4] „Who Killed the Electric Car?," dokumentarni film, SAD, 2006. Dostupno online: http://www.whokilledtheelectriccar.com/ [siječanj 2013.].

[5] J. Larminie, J. Lowry. : „Electric Vehicle Technology Explained," John Wiley & Sons Ltd., West Sussex, England, 2003.

[6] Future Car Challenge, The Royal Automobile Club, Velika Britanija. Dostupno online: http://www.futurecarchallenge.com [siječanj 2013.].

[7] P. Sasa, S. Skok, V. Kirincic: „Urban electric vehicle development as part of Smart Grid", International Conference on Innovative Technologies (IN-TECH) Rijeka, 26 – 28 September 2012, Rijeka, Croatia.

[8] DIRECTIVE 2009/28/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 23 April 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC. Dostupno online: http://eur-lex.europa.eu [siječanj 2013.].

[9] Cjenik goriva, Ina d.d. Dostupno online: http://www.ina.hr [veljača 2013.].

[10] Tarifni modeli, HEP-Operator distribucijskog sustava d.o.o. Dostupno online: www.hep.hr [veljača 2013.].

[11] T. Sencic: „Električna vozila - san ili realnost," predavanje. Tehnički fakultet, Sveučilište u Rijeci, Rijeka, 2012.

[12] Green transport and energy. Dostupno online: http://greentransportandenergy.blogspot.com/2009_05_01_archive.html [prosinac 2012.].

[13] Pravilnik o utvrdivanju najviših maloprodajnih cijena naftnih derivata, Ministarstvo gospodarstva Republike Hrvatske, prosinac 2012. Dostupno online: http://narodne-novine.nn.hr/clanci/sluzbeni/2012_12_145_3102.html [ožujak 2013.].

[14] S. Dhameja: „Electric Vehicle Battery Systems," Butterworth–Heinemann, USA, 2002.

[15] P. D. Blair: „Modeling energy and power requirements of electric vehicles,“ Energy Conversion, Volume 18, Issue 3, 1978, Pages 127-134

[16] J. R. Bumby, P. H. Clarke, I. Forster: „Computer modelling of the automotive energy requirements for internal combustion engine and battery electric-powered vehicles, IEE Proceedings Science, Measurement and Technlogy,“ Vol. 132, No 5, 1985, Pages 265-279

[17] S. Campanari, G. Manzolini, F. G. de la Iglesia,: „Energy analysis of electric vehicles using batteries or fuel cells through well-to-wheel driving cycle simulations,“ Journal of Power Sources 186, 2009, Pages 464–477

[18] T. Sencic, V. Kirincic, K. Lenic, „Electric vehicle modelling for real application,“ The international Congress Energy and the Environment 2012, 17-18 October 2012, Rijeka, Croatia.

[19] C. Mi, A. Masrur: „Hybrid electric vehicles, principles and applications with practical perspective,“ Wiley, 2011.

The educational model of the system integrity protection scheme

Ingrid Sterpin [1], Vedran Kirincic [2], Srdjan Skok [3]

Department of Power Systems, Faculty of engineering, University of Rijeka
Vukovarska 58, 51000 Rijeka. Croatia
[1] ingrid.sterpin@gmail.hr
[2] vedran.kirincic@riteh.hr
[3] srdjan.skok@riteh.hr

Abstract - Lifestyle of modern society results in rising need for electric energy. On the other hand, reduction of pollution and increase of energy efficiency leads to integration of renewable energy sources. Dynamic and unpredictable loads make the power system operation a challenging task, which results in lower security margins and large power system blackouts. Therefore, power utilities are collaborating with experts from industry and researchers from academia on development of technical solutions and concepts that should support secure and economical power system operation. Regarding the power system protection, instead of focusing on only the faulted element, the power system is considered as a whole, which brings us to variety of System Integrity Protection Schemes (SIPS). The paper investigates application of the synchronized measurement technology for development of the SIPS. Simulation of the test power system is performed in the Siemens PSS/E. Among further improvement of the proposed solution and development of new schemes, the model has educational value, as it will be used as part of laboratory exercises at The University of Rijeka, Faculty of Engineering, in order to educate future generations of engineers about emerging technologies.

Keywords: Power system, System Integrity Protection Scheme (SIPS), Synchronized Phasor measurements, Laboratory model

I. INTRODUCTION

Due to modern trends, which include integration of renewable energy sources (RES) and liberalization of the energy market, the power system operates with lower security margins that can result in large power system blackouts. To prevent cascading outages, energy utilities all over the world are developing System Integrity Protection Schemes (SIPS), also known as System Protection Schemes (SPS) or Remedial Action Schemes (RAS) [1]-[3]. While conventional protection systems focus on isolating the faulted element, the SIPS takes adequate actions to maintain the power system integrity. The SIPS can be based on synchronized measurement technology, which uses GPS signal to synchronize the sampling processes of geographically dislocated phasors or voltage and current. The collected synchrophasors can be used in various applications for Wide Area Monitoring, Protection and Control (WAMPAC) systems.

II. SYNCHRONIZED MEASUREMENT TECHNOLOGY

Although the fundamentals of the synchronized measurement technology are set several decades ago, wider application of the synchrophasors started in the last decade. In comparison with the conventional Supervisory Control And Data Acquisition (SCADA) system, the Wide Area Monitoring System (WAMS) offers several advantages, as presented in Table I [4].

TABLE I. WAMS AND SCADA CHARACTERISTICS [4]

Characteristic	**WAMS**	**SCADA**
Synchronization accuracy	1 μs	1 s or more
Data sending period	20 ms	1 s - 1 min
Data archiving period	100 ms	1 s - 1 min

The WAMS architecture consists of:

- Phasor Measurement Units (PMUs),
- Phasor Data Concentrator (PDC),
- Telecommunication infrastructure.

The PMUs are installed in power facilities in order to collect the synchrophasors of voltages and currents. To ensure data availability, the adequate level of redundancy is needed. The block diagram of the PMU is given in Fig. 1 [5].

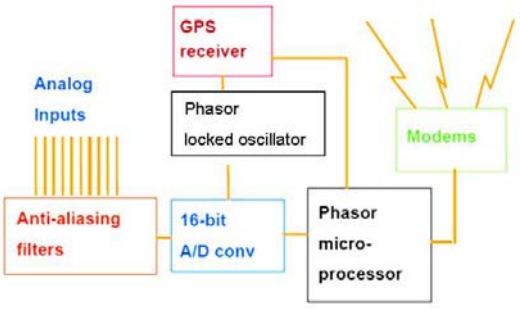


Figure 1. Block diagram of the PMU [5]

The synchrophasors measured by the PMUs are collected in the PDC, which consists of software and hardware. The software is installed in the equipment in power system management centers. The hardware consists of two types of server-client configuration:

- server-client,
- main server-sub servers-client.

The first configuration collects data on one level and it is recommended for systems with a smaller number of PMUs. The second configuration collects data on two or more levels and a larger number of PMUs, therefore a Super Phasor Data Concentrator (SPDC) could be used.

The communication infrastructure for the WAMS uses one of the following options:

- Telephone lines,
- Optical links,
- Satellite connections,
- Power Line Communication – PLC,
- Microwave connections.

Advantages of the telephone lines are lower costs and simplicity. However, the use of optical links is rapidly increasing because of their superior characteristics, such as faster data transmission and lower interference.

III. Synchronized Measurement Technology in Croatia

The first pair of PMUs was accommodated in the Croatian power system in 2003 to monitor the 400 kV power line Tumbri-Zerjavinec. The ABB PSGuard software package was based on point to point communication architecture. The used modem had a 19200 bit/s speed with less than 100 ms of data delay. The detailed overview of the first phase of the Croatian WAMS is given in literature [1]-[3]. The second phase followed in 2007 with installation of additional PMUs in other substations. At the moment (January 2013), there are 14 PMUs deployed in the Croatian transmission power system, with several portable PMUs that collect low-voltage measurements [7]. Fig. 2 gives geographical locations of the PMUs, which observe the complete 400 kV and the part of the 220 kV power system [2]. Table II gives a list of the substations where the PMUs are deployed. Currently used applications include:

- Data archiving,
- Monitoring of voltage angles,
- Monitoring of voltage stability,
- Monitoring of thermal line ratings,
- Monitoring of power oscillations.

Figure 2. Overview of the PMUs in the Croatian transmission power system [2]

TABLE II. LOCATIONS OF THE PMUS IN THE CROATIAN TRANSMISSION POWER SYSTEM

Substation	Monitored branch toward
Melina	Tumbri
	Velebit
Ernestinovo	Zerjavinec
Zerjavinec	Ernestinovo
	Tumbri
Tumbri	Zerjavinec
	Melina
Velebit	Melina
	Konjsko
Konjsko	Velebit
	Brinje
Brinje	Mraclin
	Konjsko
Mraclin	Brinje

IV. Educational Model of the System Integrity Protection Scheme

When a fault occurs in the power system, the power-system protection isolates the faulted element which may lead to the overload of another element in the grid. On the other hand, the SIPS should detect the overload and apply some of the corrective actions, such as load shedding, redirecting power flows, changing generation levels, etc [8]-[9].

Educational model of the SIPS is tested using the test power system modeled in program package Siemens PSS/E [10]. The test power system that is given in Fig. 3 is divided into three subsystems, namely: “STUDY”, “EAST” and “WEST”. With an assumption that the

synchrophasors are collected by the PMUs deployed in the system, they could be used as inputs for the SIPS.

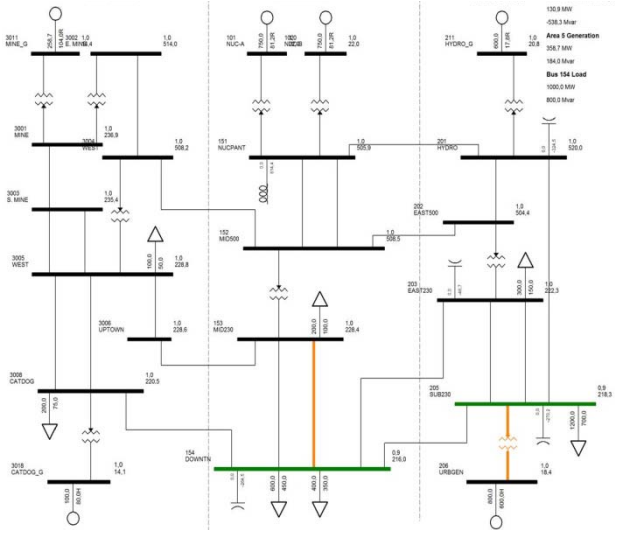


Figure 3. The test power system

Fig. 4 presents results of the power flow analysis, with a graphical representation of the lines' loading levels.

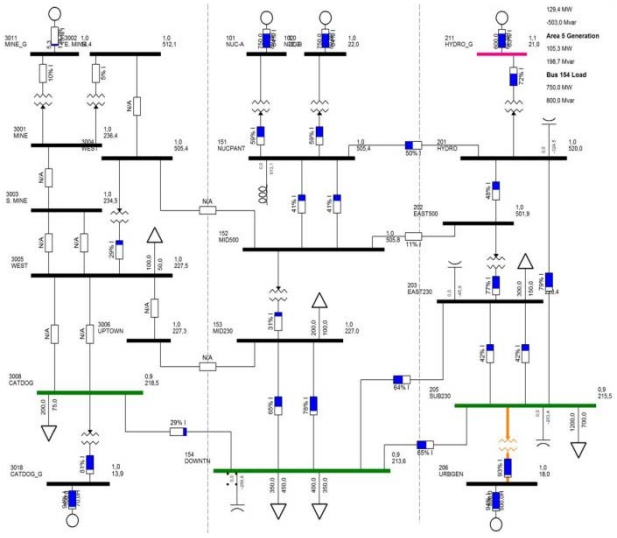


Figure 4. Power flow analysis

The contingency analysis was performed for each element in all the three subsystems. However, the results will be given for one case in each subsystem.

A. Subsystem "STUDY"

Fig. 5 presents the results obtained by simulating the outage of one of two parallel branches connecting the buses #153 and #154. The outage caused overload of the remaining branch between the buses #153 and #154, and high loading levels of the branches #154 - #203, #154 - #205, #201 -#205, #202 - #203 and #205 - #206.

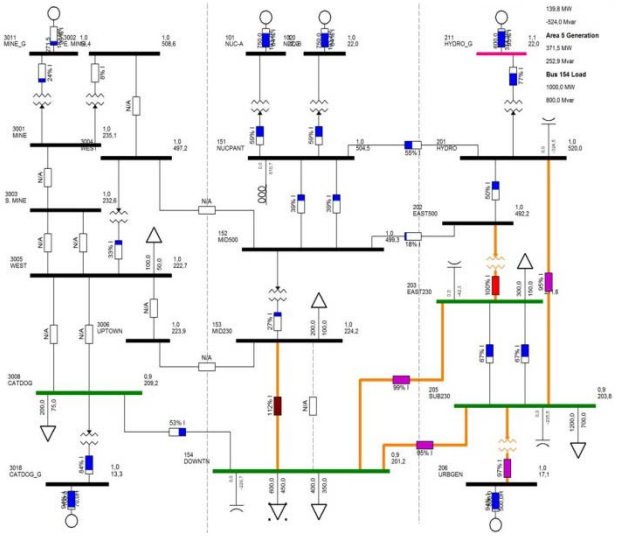


Figure 5. Subsystem "STUDY" – outage and overload

After performing the contingency analysis, the SIPS could be deployed to mitigate the overloads in the power system. Fig. 6 gives the flow diagram of the SIPS for the described contingency.

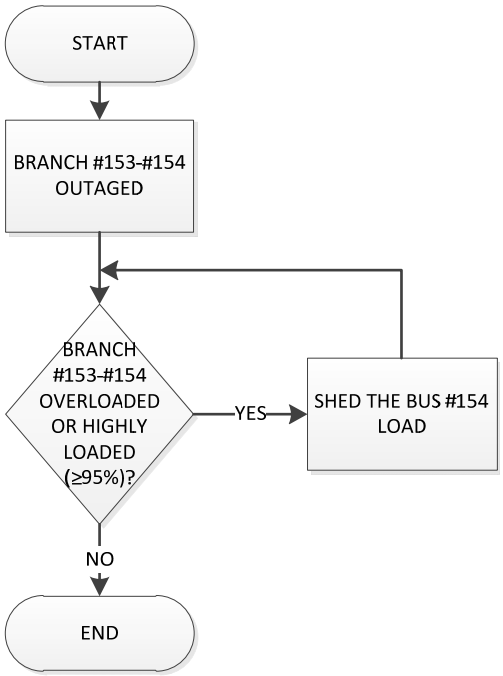


Figure 6. Flow diagram of the SIPS

Fig. 7 presents the power flows in the system after the described SIPS is applied. After shedding the load connected to bus #154 for 250 MW, the remaining branch #153 - #154 is loaded below its nominal loading level.

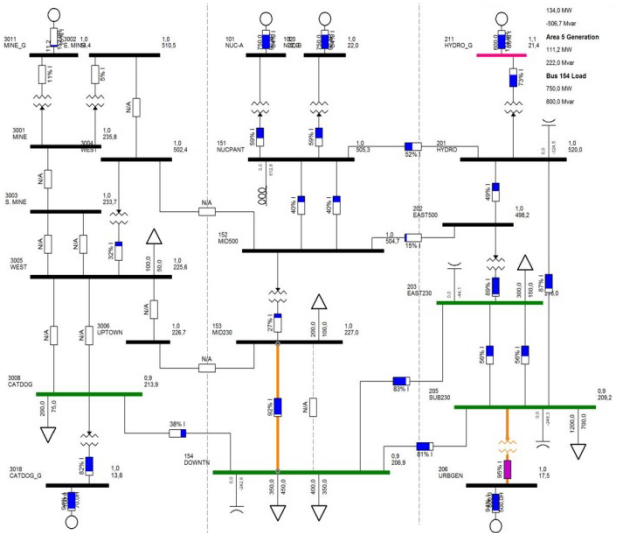


Figure 7. Subsystem "STUDY" –SIPS applied

B. Subsystem "EAST"

Fig. 8 gives the results after simulating the outage of the branch connecting the buses #203 and #205. This outage caused high loading level of one of the parallel branches between the buses #153 and #154.

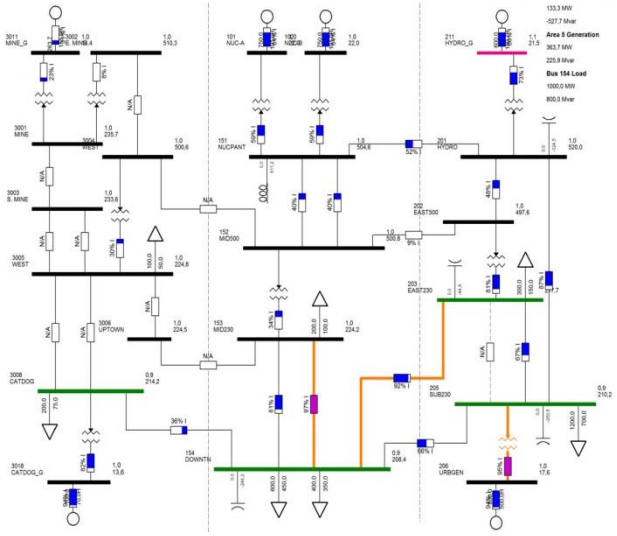


Figure 8. Subsystem "EAST" – outage and overload

Fig. 9 gives the flow diagram of the SIPS for the described contingency.

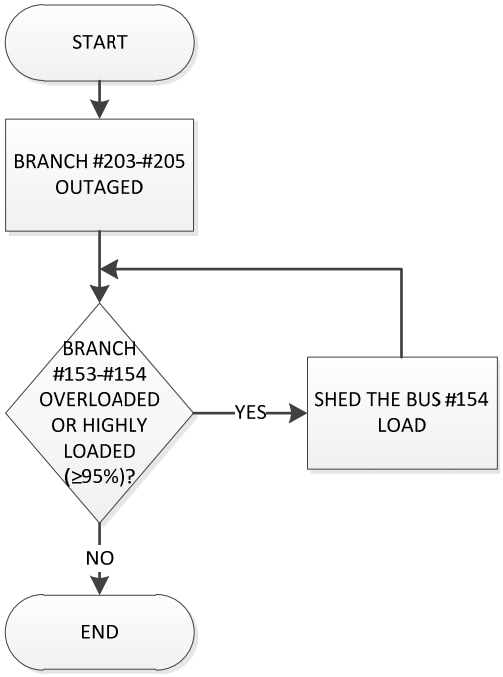


Figure 9. Flow diagram of the SIP

Fig. 10 presents the power flows in the system after the described SIPS is applied.

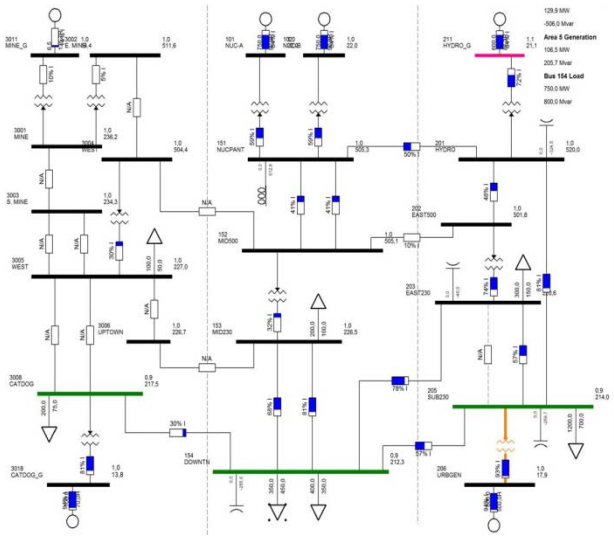


Figure 10. Subsystem "EAST" – SIPS applied

C. Subsystem "WEST"

Fig. 11 represents the outage of the branch between the buses #3005 and #3008. As in the previously simulated cases, one of the parallel branches between the buses #153 and #154 is highly loaded.

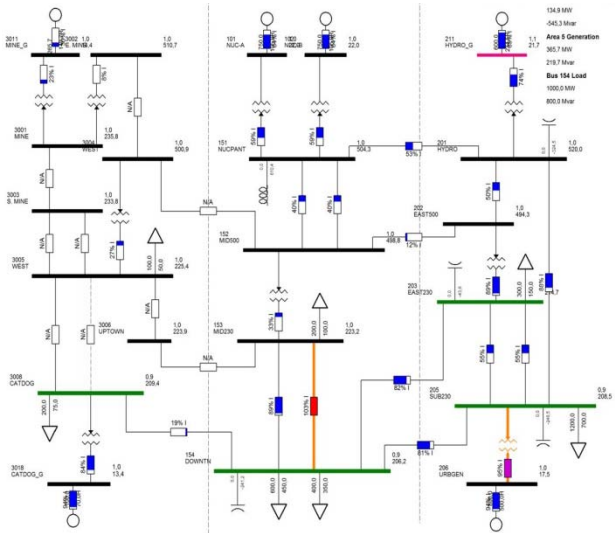


Figure 11. Subsystem "WEST" – outage and overload

Fig. 12 gives the flow diagram of the SIPS for the described contingency.

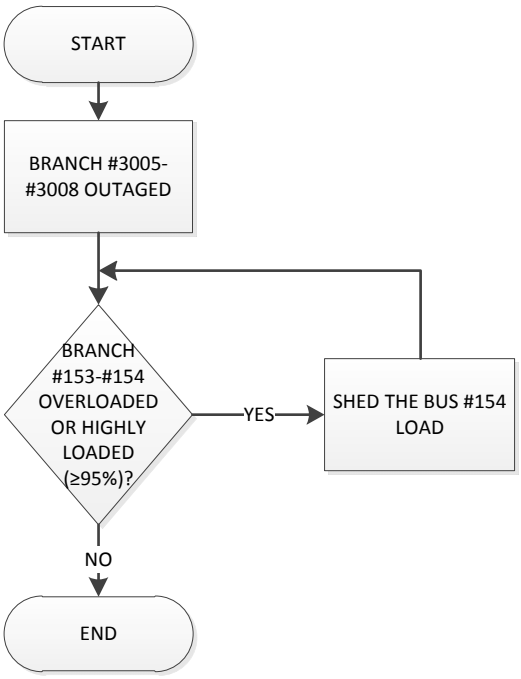


Figure 12.Flow diagram of the SIPS

Fig. 13 presents the power flows in the system after the described SIPS is applied.

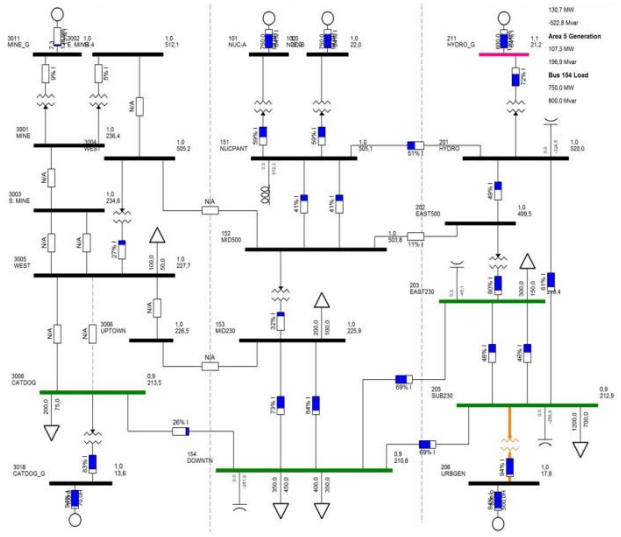


Figure 13. Subsystem "WEST" – SIPS applied

D. Overall flow diagram of the SIPS

After performing the contingency analysis for different cases, it can be concluded that several single outages in the test power system cause overload or high loading of one of two parallel branches between the buses #153 and #154. A remedial action for these scenarios is to shed the load connected to the bus #154. Therefore, the flow diagram for the simulated contingency is given in Fig 14.

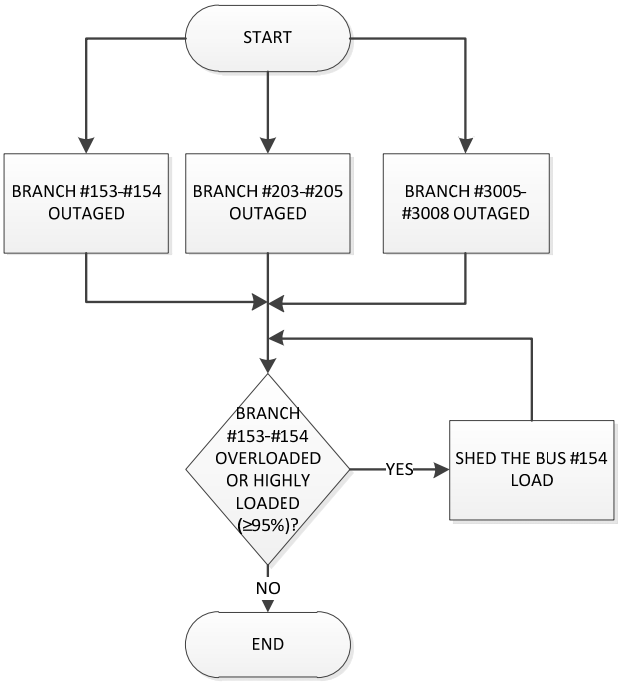


Figure 14. Flow diagram of the SIPS for simulated contingency

The presented results and the flow diagram is an example of the procedure that should be performed in

order to develop the SIPS for other possible contingencies in the power system.

V. CONCLUSION

Dynamic and unpredictable loads make the power system control a challenging task. Considering the large amount and the slow refreshing rate of the data collected by the SCADA system, the operators are unable to get a full insight into the dynamic situation in the grid.

The synchronized measurement technology provides more accurate information about the power system state in real time. Availability of the current and voltage synchrophasors is a driving factor for development of WAMPAC applications. While conventional protection systems focus on the faulted element, the SIPS maintains the power system integrity.

The educational model of the SIPS, based on the synchrophasors, is presented. The results indicate that predefined actions can be effective in mitigating overloads of the power system elements and eventually prevent the cascading power system blackouts. The model will be used for educational purposes, as part of the laboratory exercises at the University of Rijeka, Faculty of Engineering. Also, further development of the applications based on the synchronized measurement technology is planned.

REFERENCES

[1] R. Matica, „Wide Area Monitoring, Protection and Control for enhancement of the power system operation “, qualifying doctoral exam, Faculty of computing and engineering, University of Zagreb, Croatia.

[2] S. Skok, V. Kirincic, K. Frlan, “Dynamic Analysis of Wind Farm Operation Integrated in Power System Based on Synchronized Measurements,” Engineering review, vol. 30, no. 1, pp. 73-83, 2010.

[3] S. Skok, R. Matica, I. Sturlic, "Enhanced open architecture of phasor data concentrator," European Transactions on Electrical Power, 2010.

[4] V. Kirincic, S. Skok, I. Pavic, „Power system state estimation based on PMU measurements vs SCADA measurements,” International Review on Modelling and Simulations - IREMOS, accepted for publication, ISSN 1974-9821, 2012.

[5] R. F. Nuqui, „State Estimation and Voltage Security Monitoring Using Synchronized Phasor Measurements,“ PhD dissertation, Virginia Polytechnic Institute and State University, 2001.

[6] S. Skok, V. Kirincic, I. Ivankovic, "Automated control for power transmission system in urban area," IEEE GCC Conference and Exhibition, pp.649-652, UAE, 2011.

[7] S. Skok, D. Brnobic, V. Kirincic, "Croatian Academic Research Wide Area Monitoring System - CARWAMS", The International Journal on Communications Antenna and Propagation - IRECAP, vol. 1, no. 4, pp. 72-78, Aug. 2011.

[8] J. McCalley, „System Protection Schemes: Limitations, Risks and Management“, PSERC Publication 10-19, 2010.

[9] V. Madani, D. Novosel, R. King, „Technological breakthroughs in System Integrity Protection Schemes“, 16th PSCC, Glasgow, Scotland, 2008.

[10] SIEMENS: „PSS/E 32.0.5 Program Operation Manual“

Stochastic Simulation of the Smart Grid and Demand Response Implementations on a City-wide Scale

B. Morvaj*, B. Jurisic* and N. Holjevac *

* Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia

Abstract - **The paper focuses on modeling and simulation of the stochastic nature of a city that has implemented the smart grid and demand response paradigm. A stochastic model has been developed in Anylogic using agent based and system dynamics modeling. The city model consists of over 1000 smart building agents with controllable loads, local energy storage and local renewable energy source (photovoltaic). After elaboration of modeling principles, two simulations were carried out. In the first one, the effect of the demand response implementation on a city-wide level from an energy perspective is examined. In the second one it is analyzed whether the demand response increases security in the power supply. Also, terms demand response, smart grid and smart building are defined and explained.**

KEYWORDS: stochastic simulation, demand response, smart building, smart grid, distributed renewable energy sources

I. INTRODUCTION

World primary energy consumption is projected to grow by 1.6% per year from 2011 to 2030, adding 36% to global consumption by 2030 [1]. By 2030 world population is projected to reach 8.3 billion. Even now more than half of the world's population lives in urban cities [2] and it is projected that this trend will continue for the next couple of decades [3].

With the increase of population and sustainable development in the spotlight, cities now face the challenge of sustainable local urban development. Difficulty in energy supply, waste management, limited resources, air pollution, human health concerns, traffic congestions, deteriorating and aging infrastructures are among the more basic technical, physical, and material problems [4, 5].

Energy supply is what underlays all mentioned problems. Without energy there is no development. Today building new conventional power plants is not anymore an option. Energy supply needs to be sustainable and renewable energy sources are the backbone of the sustainable development. To enable optimal use of renewables, the grid must be upgraded into smart grid. Also, generation must shift from central point(s) to distributed generation in the actual cities. Therefore, cities must also implement smart grid and demand response and become smart cities. Demand response enable better integration of renewables and optimal use of limited resources.

This research is sponsored by NATO's Public Diplomacy Division in the framework of "Science for Peace", project SfP-983805: Emergent phenomena testbed simulator for improving SCADA performance in power system security management.

II. DEFINITIONS

A. Smart city

There are numerous definitions to the term smart city. We define a smart city as a city that continually increases its performance in satisfying all needs of its citizens. This is based on literature where smart city if defined as a city well performing in a forward-looking way in economy, people, governance, mobility, environment, and living, built on the smart combination of endowments and activities of self-decisive, independent and aware citizens [6].

B. Smart grid

A smart grid is a concept that combines ICT and electrical power grids. Its main feature is establishing two-way communication in all nodes of the grid using advanced metering infrastructure. With customer involvement (passive and active) it can improve overall energy efficiency and reliability of the grid, and decrease energy consumption. It also enables control and optimization of renewable energy sources and demand side management programs by delivering real-time information to control nodes, enabling near-instantaneous balance of supply and demand at the level of individual consumers – buildings, even appliances [7].

C. Smart building

A smart building is an energy efficient building with additional subsystems for managing and controlling renewable energy sources, house appliances and energy consumption using most often a wireless communication technology. ICT enables smart buildings to communicate both with their inside devices and appliances, which they can also control, and with their surroundings. Furthermore, they can adapt to a grid's conditions and communicate with other buildings, hence creating active microgrids or virtual power plants.

D. Demand response

Demand response (DR) is an alternative solution to reduce peak power periods and decrease the demand in peak times. Also, in regions with high penetration of

renewable energy sources, DR can trigger the change of demand to follow the change of supply. DR programs enable customers to manage their consumption of electricity in response to supply conditions. For example, many programs have electricity customers reduce their consumption at critical peak load hours or in response to market prices. To achieve this goal, incentive-based, rate-based and demand reduction bids DR programs are developed [8].

III. MODEL

This model is a continuation of the previous work presented at IYCE in July 2011 [7]. In the smart grid environment, buildings act as active members of the grid. This is made possible by installing energy storage and a small renewable energy source – photovoltaic combined with controllable loads. The buildings can store or give into the grid electricity as part of the demand response program.

It is known that demand response has numerous benefits [9]. It is easy to calculate the impact of the demand response program on a small scale level consisting of few buildings. But it is very hard to show the results of implementing demand response on a large scale level such as a (small) city due to the stochastic nature of the buildings. Also, an energy management algorithm needs to be developed which would take into an account all possibilities (such as availability of the PV, controllable loads and energy storage capacity) to increase or decrease demand and/or supply. According to [10] this concept is called advanced smart grid. Currently, this cannot be fully implemented. But it is inevitable that such smart grid architecture will be developed that will integrated all the mentioned technologies and the paradigm because it is essential for a sustainable development of the grid [10]. Also, many utilities, as well as public and private initiatives have started to define and develop smart grid architecture even though currently it is not economically justified. The purpose of the model is to simulate the inevitable (advanced) smart grid and demand response paradigm implementation on a city level controlled by the developed energy management algorithm (Figure 2) and will be used for the project.

The city model is made in a software AnyLogic. AnyLogic is a simulation tool made in Java that supports system dynamics, process-centric and agent based modeling [11]. For this model, both agent based and system dynamics modeling, was used.

The city model consists of two levels:

- The lower level is a building modelled as an agent. It consists of controllable and uncontrollable loads, PV and an energy storage.
- The higher level (Figure 1) acts as an environment for the agent building where it is multiplied, each with its own unique behavior. Also, energy management algorithm is implemented on this level.

A. *Lower level – the agent building*

The building is modelled as an agent and it acts as a smart building which means that the automation system controls all loads and also the utility can control the building. The agent building's behavior is defined by the system dynamics modeling and a state chart. Building's consumption is based on average Europe home building consumption.

Agent building has 7 specific loads modeled using the system dynamics modeling. The most common loads are modeled. They are water heater, dishwasher, washing machine, refrigerator, television and PC, air condition and lighting. Based on research [12] the average consumption and working hours per day of the loads are defined but each load's consumption and the starting time is represented by uniform distribution to model the stochastic

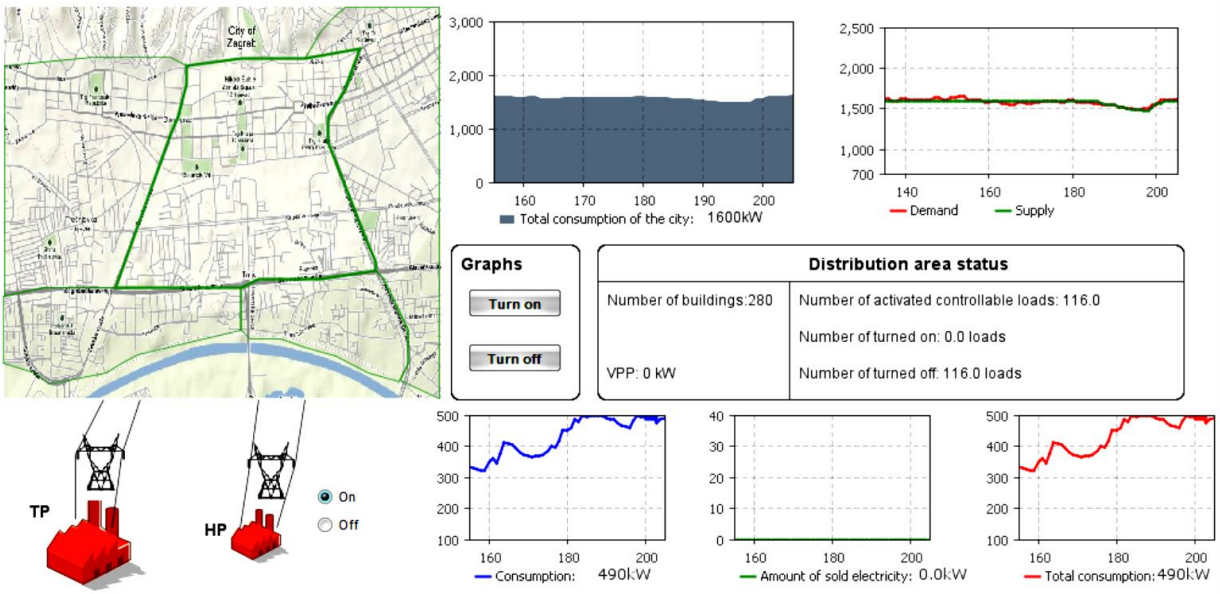


Figure 1. Overview of the main level

nature in the simulation since not all citizens will have the same kind of load that needs the same amount of electricity and will not turn it on at the same time. Uniform distribution is a probability distribution in which all intervals of the same length of the distribution's support are equally probable. The support is defined by the two parameters that define minimum and maximum values. Parameters of the loads are shown in Table 1.

TABLE 1. PARAMETERS OF THE LOADS IN THE BUILDING

Load	On time period per day (min)	Off time period per day (min)	Consumption (W)
Water heater	30 (cyclic)	70 (cyclic)	800-900
Dishwasher	60	1380	1300-1450
Washing machine	60	1380	1000-1200
Refrigerator	15 (cyclic)	35 (cyclic)	400-470
TV and PC	480	960	500-600
A/C	15 (cyclic)	35 (cyclic)	900-1100
Lightning	480	960	200-300

The building also has a photovoltaic system and an energy storage – battery. Photovoltaic is modeled using system dynamics and generates electricity depending the time of day and if there is any sun resource available. Battery storage is modeled using the state chart.

B. Higher level – the agent environment

The higher level acts as an environment for the agent building and simulates the stochastic behavior of a city. The building agent is multiplied 1030 times in the environment. Each building agent acts independently and uniquely, so there are 1030 buildings behaving differently which corresponds with the real city where human behavior is stochastic. Also, each modelled load starts at the different time in each building which makes 7210 modeled loads that are independent and stochastic.

There are two possibilities for the future – the city will be one big interconnected active grid controlled from a single main point or the city will be divided into smaller active distribution areas that are coordinated both individually and from the higher central point. We find that the latter option is more probable so the city is divided into five distribution areas.

Also, the city is supplied by two power plants – a thermal power plant that can change its output power by 5 percent total and a hydro power plant that can change its output power by 50 percent. The consumption, number of

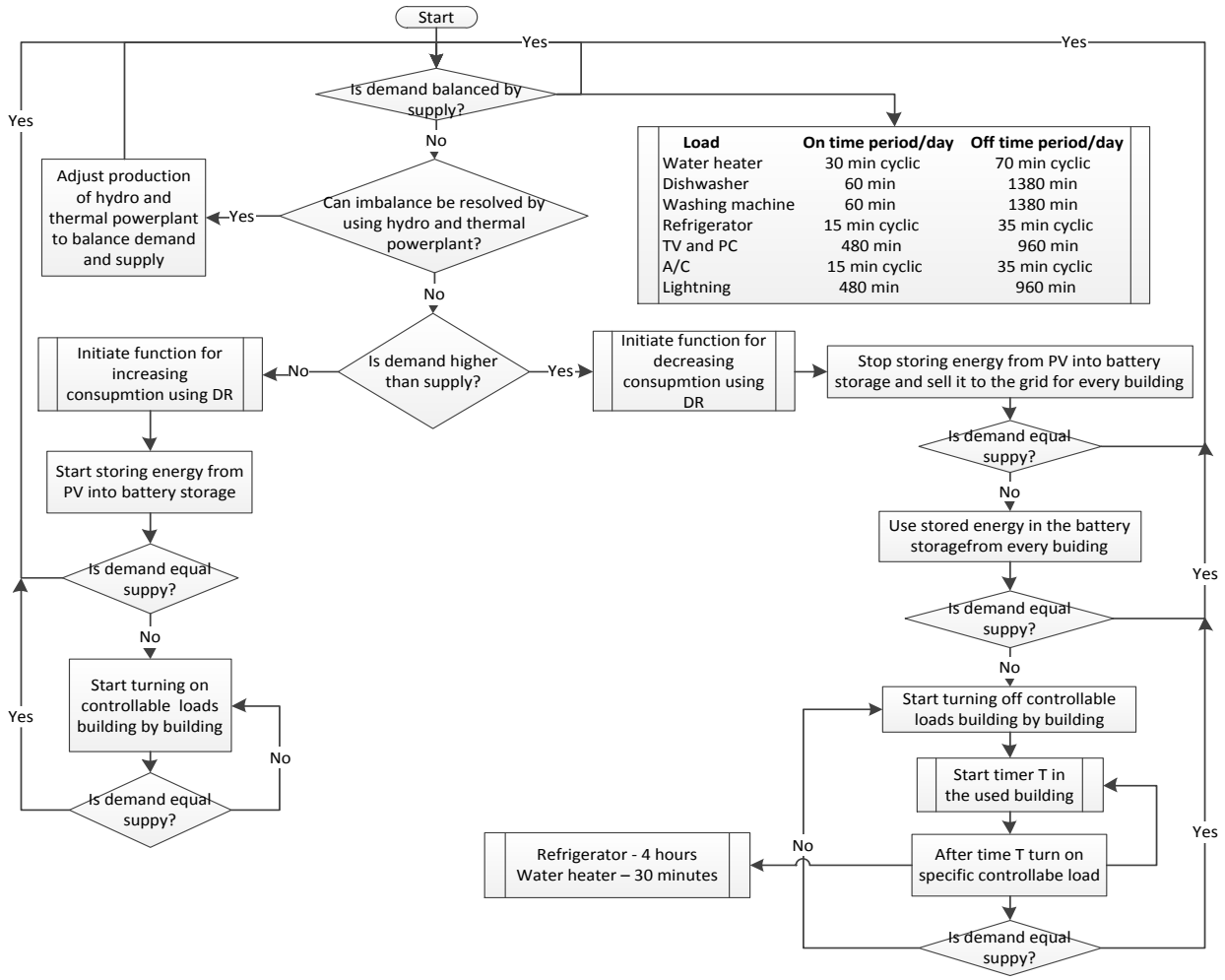


Figure 2. Overview of the algorithm

turned off loads, amount of generated electricity from the building, total consumption and generation of the each distribution area can be read on the simulation screen. The whole city is controlled by the developed energy management algorithm.

C. Energy management algorithm

The developed energy management algorithm controls in minute resolution and in the following manner (Figure 2). It can be divided into 3 cases. All are based on the balance of supply and demand. The first one occurs when the supply is higher than demand, the second one when the demand is higher than the supply and the third one when the supply is higher than the demand but only after the demand response was used.

- The first case: If the PV system produces electricity and is giving it to the grid then the energy surplus is stored into the energy storage to balance generation and consumption. This is done only after the thermo and hydro power plant decreased their generation by the maximum possible value;

- The second case: If the consumption is higher than generation due to a malfunction or peak power, first the thermo and hydro power plants increase their generation. If that is not enough then the extra needed energy is taken from the batteries. If the malfunction has not been repaired in the meantime and all batteries are empty, the four controllable loads (water heater, dishwasher, washing machine and refrigerator) start to be turned off in buildings until the balance is again achieved. During this, it is taken into account that if the refrigerator is turned off, it must be turned on after 4 hours or the food inside will start to turn bad. It is the same with the water heater, if it is turned off, it must be turned on after 45 minutes because tenants must not suffer any decrease in comfort if they want to take a shower.

- The third case: When the fault is repaired and the generation is higher than the consumption, one by one the loads are turned on while taking into account not to exceed the generation capacity. After a particular load has been turned on, its behavior resets and starts to have original on/off time periods.

IV. SIMULATION AND RESULTS

In the simulation, two scenarios were analyzed:

- Scenario 1: meeting the peak power using the demand response
- Scenario 2: minimizing the effect of bigger malfunction also using the demand response.

A. Scenario 1 results

During some period of the day (in this case from time t=146 to t=168 min) shown in Figure 3 there is a sudden increase in demand and hydro and thermal power plant cannot balance it out. This is without using demand response program. Additional 350 kW of power is needed to meet the peak period.

From period t=180 to t=198 min the demand decreases beyond balancing capabilities of the hydro and power plant and unbalance between demand and supply. Supply should decrease or demand increase by 150 kW. This is because PV are producing electricity and without demand response it cannot be controlled and stored into energy storage.

After simulating the same event but this time also using the demand response program, the obtained results are shown in Figure 4. Worst case scenario was simulated when there is no energy stored in the energy storage. This time it can be seen that the generation and consumption are

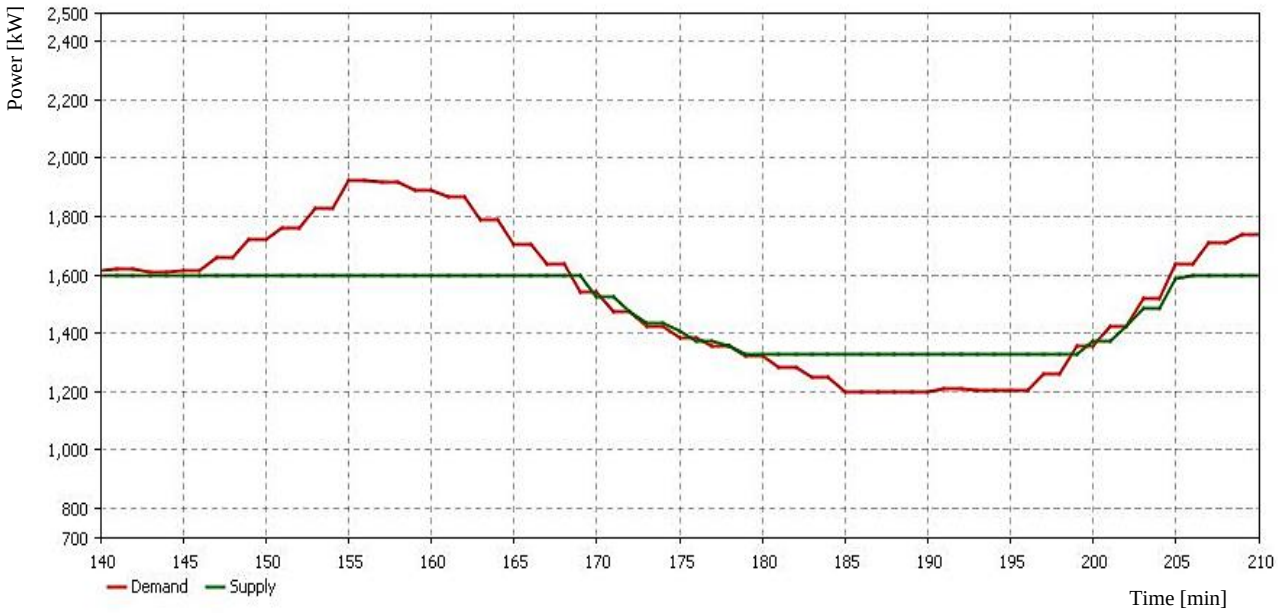


Figure 3. Results of scenario 1 simulation without demand response

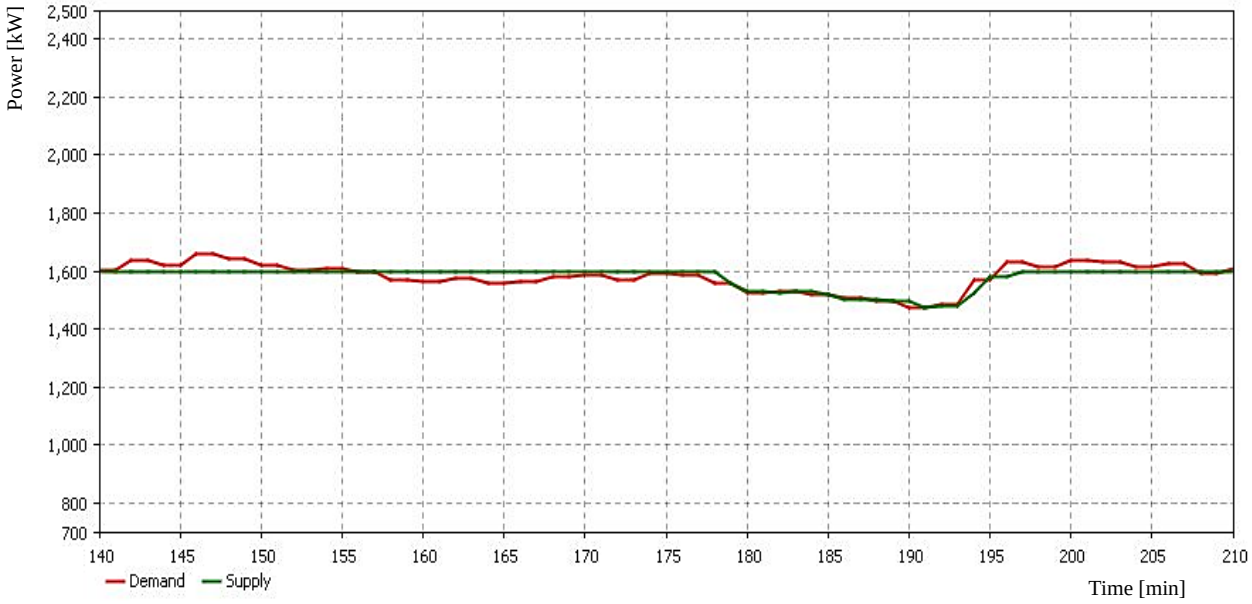


Figure 4. Results of scenario 1 simulation with demand response

always balanced due to the turning off controllable loads in period of higher demand and storing extra electricity from PV into energy storages in period of lower demand. There are minor unbalances because demand response cannot instantly and perfectly adjust demand. Actions taken by the demand response to balance out supply and demand at specific times can be seen in Table 2.

TABLE 2. OVERVIEW OF THE DEMAND RESPONSE ACTIONS

Time (min)	Amount of turned off controllable loads	Power decrease due DR
148	414	80 kW
153	779	190 kW
155	975	320 kW
160	932	300 kW
210	618	180 kW

B. Scenario 2 results

In this scenario, the following situation was simulated. Hydro power plant has stopped working due to a huge malfunction. This happened at the time t=224 min. It is not possible to import energy from anywhere else and all that is available is the demand response. The situation without the demand response is shown in Figure 5. There is an average need of 450 kW to balance demand and supply. The imbalance would probably cause the start of disconnecting customers. After simulating the same event but this time also using the demand response program, the obtained results are shown in Figure 6. It can be seen that due to the demand response, demand and supply are balanced for 25 min.

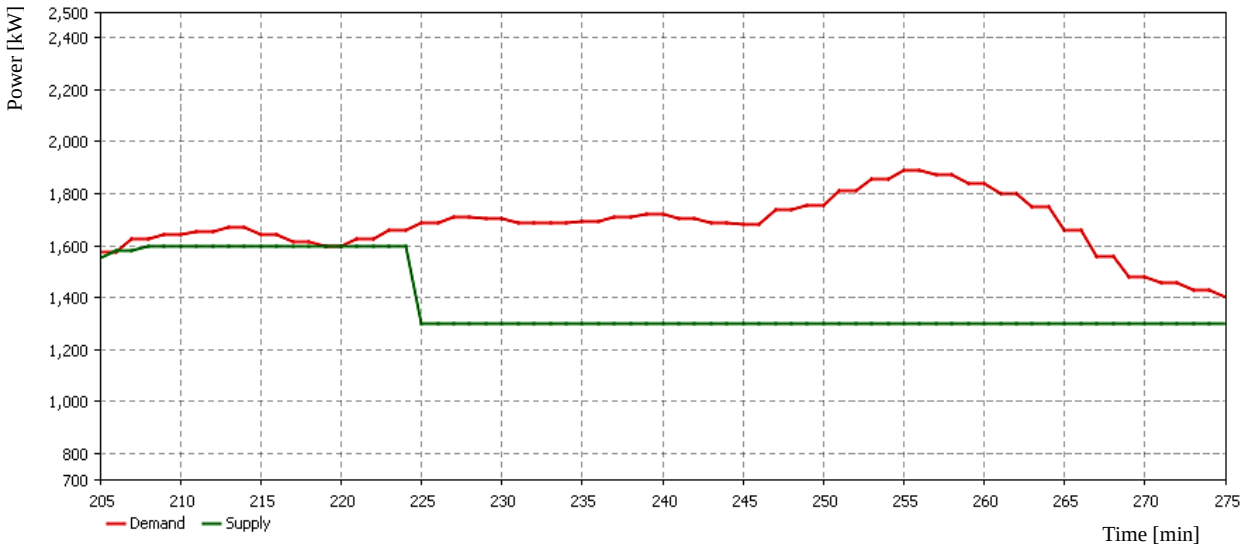


Figure 5. Results of scenario 2 simulation without demand response

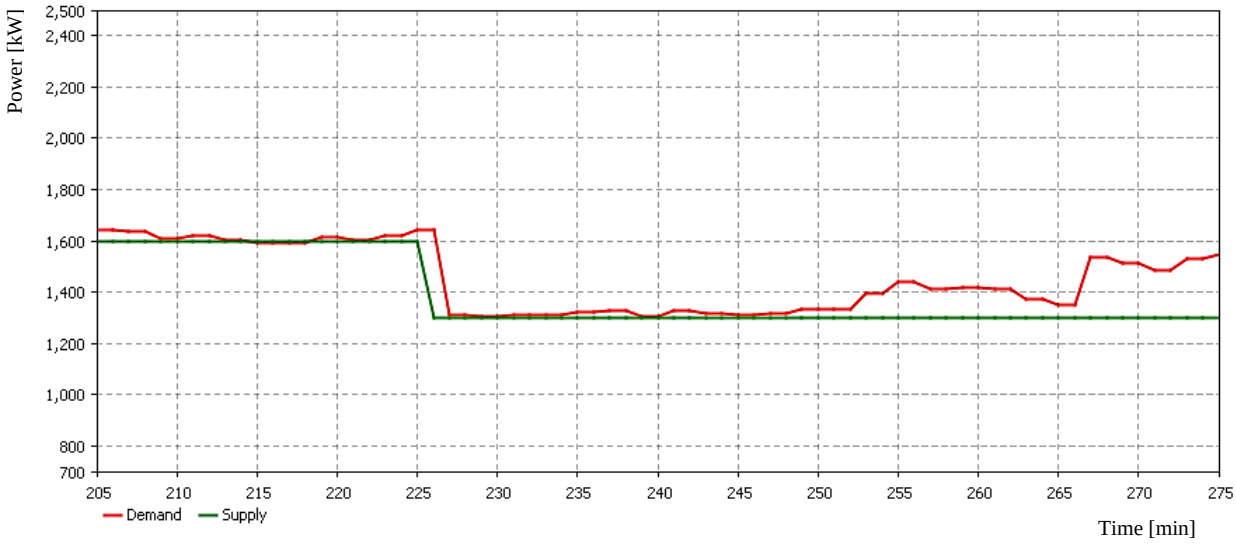


Figure 6. Results of scenario 2 simulation with demand response

That means there is enough time to repair the malfunction if it is a smaller one or turn on backup generators. None of the customers would be disconnected and would not even be aware of the problem. Actions taken by the demand response to balance out supply and demand at specific times can be seen in Table 3. Specific controllable loads that were turned off, need to be turned on after some time so that the comfort of customers is not disrupted. This is the reason why at the time t=253 min demand and supply cannot be balanced, since too many loads have to be turned on.

TABLE 3. OVERVIEW OF THE DEMAND RESPONSE ACTIONS

Time (min)	Amount of turned off controllable loads	Amount of automatically turned on controllable loads	Power decrease due DR
210	423	0	85 kW
228	1239	0	370 kW
238	1463	19	400 kW
246	1448	142	390 kW
255	1744	157	430 kW
265	1793	143	440 kW
275	1446	490	220 kW

V. CONCLUSION

The goal was to create a model for analysis of interactions within a smart city. Our aim was to develop an operable system for further development towards larger scale simulation between different systems vital to a smart city as part of the project. This model would be used for analyzing available current state of the city and possible options, and then for calculating and simulating the optimal usage of the local renewable energy sources, battery storages and controllable loads. The results would be passed to another module which would calculate if the solution is possible in terms of physical power flows.

For the purpose of this paper, the simulations were carried out in order to show the impact of the demand response on a city scale in terms of reducing peak powers and security of the supply. As it can be seen, demand response can decrease the need for building new power plants for the peak power periods. Also, it increase the security and availability of the supply in such a way that it gives more time for the repair of malfunctions or finding some alternative solution(s) until the malfunction can be dealt with.

REFERENCES

[1] BP Energy Outlook 2030, January 2013, Accessed in February 2013

[2] Dirks, S., Gurdgiev, C., & Keeling, M. "Smarter Cities for Smarter Growth: How Cities Can Optimize Their Systems for the Talent-Based Economy. Somers, NY: IBM Global Business Services.", 2010

[3] H. Chourabi , T. Nam, S. Walker and others, "Understanding Smart Cities: An Integrative Framework", 2012, 45th Hawaii International Conference on System Sciences

[4] Borja, J. "Counterpoint: Intelligent cities and innovative cities. " Universitat Oberta de Catalunya (UOC) Papers: E-Journal on the Knowledge Society, 2007

[5] Marceau, J. "Introduction: Innovation in the city and innovative cities. Innovation: Management", Policy & Practice, 10(2-3), 136-145 , 2008

[6] Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanović, N., & Meijers, E. "Smart Cities: Ranking of European Medium-Sized Cities." Centre of Regional Science (SRF), Vienna University of Technology, 2007

[7] B. Morvaj, L. Lugaric, S. Krajcar, "Demonstrating Smart Buildings and Smart Grid features in a Smart Energy City", 3rd International Youth Conference on Energetics. 2011

[8] S. Mohangheghi, J. Stoupis, Z. Wang, Z. Li, „Demand Response Architecture", IEEE, 2010

[9] Demand Response : a decisive breakthrough for Europe, How Europe could save Gigawatts, Billions of Euros and Millions of tons of CO2, Capgemini

[10] A. Carvalio, J. Cooper, "The Advanced Smart Grid Edge Power Driving Sustainability", Artech House, 2011

[11] AnyLogic overview, http://www.anylogic.com/overview, Accessed in February 2013.

[12] S. Karnouskos, T. Nass de Holanda, "Simulation of a Smart Grid City with Software Agents" , SAP Research

Framework for Designing a Smart Grid Testbed

B. Jurišić*, N. Holjevac* and B. Morvaj*
*Faculty of Electrical Engineering and Computing, Zagreb, University of Zagreb,
Unska 3, 1000 Zagreb, Croatia
bruno.jurisic@fer.hr ninoslav.holjevac@fer.hr boran.morvaj@fer.hr

Abstract - Traditional centralized electrical system is changing towards the distributed generation with additional smart control features based system - so called smart grid system. Current SCADA systems for the smart grid are not satisfactory enough and a new supervisory system is needed. The new system, a smart grid testbed, will take into account all of the smart grid components. In the paper a framework for such smart grid testbed is proposed. The framework comprises design of the testbed for testing power system requirements in a smart grid environment. It also provides a platform for further research and development of the grid. In the paper initial implementations steps of the testbed are defined. These steps explain possible ways in which simulation tools can be interconnected, development of the basic simulation of the smart grid using various software tools and the integration of the data acquired from the physical components installed such as hybrid system, PV system and measurements from the smart meters into the simulation.

I. INTRODUCTION

Economy, new technology and the need to preserve the environment are changing life habits of the modern human society and are changing the way electrical energy is produced and delivered to the consumers. Demands for ecological and clean production, higher efficiency and reliability, greater certainty and quality of the electrical energy allow distributed sources to take their part alongside the centralized, conventional sources. This transformation of electrical grid leads society towards a sustainable state [2]. In order to make transition possible there is a need to develop a range of simulations of the work of decentralized networks [2].

In the following work a concept of a testbed, which will help investigate the work of the grid with the connected distributed sources, will be described.

It starts with the chapter that deals with the description of the individual components. Here is the description of the installed components and the models developed in AnyLogic and PowerWorld given.

In the end the proposed concept how to integrate all the individual parts into the whole is described.

II. COMPONENTS

The system has all the components: Photovoltaic (PV) system, HybridC system, PoweWorld (PW) simulation, AnyLogic simulation and ANSYS models included and modeled separately. In the final concept all the components working together as a whole will provide valuable information about system performance under various conditions.

A. *PowerWorld simulation*

Configuration of the microgrid was chosen with the help of the software tool HOMER. Main factor in choosing configuration was not economy but the liability for microgrid to work both in the islanded and in on the grid operation [3]. The chosen microgrid is composed of (Figure 1):

- 2000 photovoltaic modules with power of 0,25 kW per unit.
- 4 wind turbines with power of 2 MW per unit,
- diesel aggregate with summed power of 4500 kW,
- 250 batteries with capacity of 7,6 kWh per unit and cumulative power of 0,51 MW,
- industry and households demand, scaled to daily consumption of 30 MWh/day and 51 MWh/day.

As it can be seen from Fig. 1 only the active power is observed in the simulation. The idea was to simulate the grid which consists of a few microgrids and observe what happens with the currents and voltages in grid. What the authors called microgrid in the simulation can also represent the Smart House or Smart Building if the power values of the components are reduced.

A generator and a load connected to the bus represent diesel aggregate and battery. The load is needed to simulate the battery discharging, that is to say, giving the power to the microgrid in the moment of a lack of energy from distributed sources. Generator

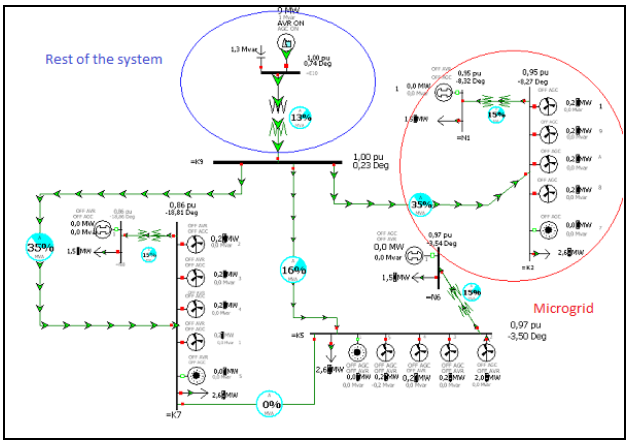


Figure 1. Overview of the PW simulation model level

This research is sponsored by NATO's Public Diplomacy Division in the framework of "Science for Peace", project SfP-983805: Emergent phenomena testbed simulator for improving SCADA performance in power system security management.

on the bus has a range of output power from -0,5 MW (maximum power which can be charging the battery) to 5 MW (maximum power which can both the battery and the diesel aggregate send to the grid).

Every microgrid contains battery, which differs from convencial distribution grid which usually does not contains battery storage. PW does not have the battery model inbuilt. Solution for that issue can be found in simultaneous control of the generator and load connected to the same bus. That cannot completely solve the problem because battery inherently has a parameter, its capacity, which cannot be modeled with the presented solution. Only way to limit the energy of the battery is to connect PW with software tool which has that ability, such as suggested AnyLogic.

PW can be used to easily simulate other physical components: photovoltaic microgrid on the D building and the hybrid microgrid on the C building (Fig. 6). Also, the load can be simulated from the real measurement from the building of the Faculty of Electrical Engineering and Computing (FER) where microgrids are located. That will allow testing and observation of different scenarios using the real, measured data.

The problem in the implementation of the mentioned case is how to efficiently connect PW to external software which can model the battery behavior. The issue can be solved using the EZJCOM which is a program interface which can connect both Java and PW [4],[5]. AnyLogic is the program which uses the Java programming language and therefore the connection can be made between PW and AnyLogic.

All the components of the simulation can be added through an integrated command line in the PW. PowerWorld uses its own script language and has its own syntax. Different macros and scripts can be coded and run from this integrated code editor. It can be used to connect PW with other programs through above mentioned EZJCOM environment.

The script language allows easier connection of models. The idea is to send results of AnyLogic simulation to the PW. That way dynamic altering of the model grid could be establised. AnyLogic would be responsible for running the multi-agent dynamic simulation of the representative microgrid and find possible problematic scenarios. Data formed to fit the PW would be sent and the microgrids configuration would be changed to check the static stability of the given network configuration under given conditions in the certain scenario.

B. Power Quality measurment

The central point of currently installed system for Power Quality monitoring and analyzing on the "Department of Energy and Power Systems", FER is the PQube device. PQube is an instrument for monitoring electric power systems (power disturbance monitor, a power meter, a power recorder, and a digital camera).

The communication with PQube is established over Web Server. Access to the measured data is done through the website with a fixed IP address http://161.53.66.233.

The "smart" monitoring of consumption on FER is being developed even further. For consumption measurement there are currently in use two SENTRON 3200 multimeters that are embedded in the main conductive fields of the low voltage distribution network. The control of multimeters is achieved with PLC (Programmable Logic Controller) and most recent data is displayed on the 12" TOUCHscreen.

The future system for monitoring the power quality will consist of three fixed and six moveable measurement points, depicted in Fig. 2. The fixed measurement points include installation of another two PQubes in a control cabinet of the main supply installations in the „Department of energy and power systems" and will serve for three-phase measuring of consumption and power quality. The moveable measurement points are planned for connection in order to observe specific load impacts on the grid.

Furthermore additional data about energy quality with distributed renewable energy sources will be gathered both from the hybrid system located on the top of the FER building C and photovoltaic plant, which has recently been installed on the roof of D-building of FER. All the data will be relayed to one central data gathering point. The server will manage the communication between all the measurement equipment and energy sources. The data gathered will be used as an input for the design of other modules in the testbed platform to study the impact of the renewable resources on the power quality.

C. Photovoltaic system

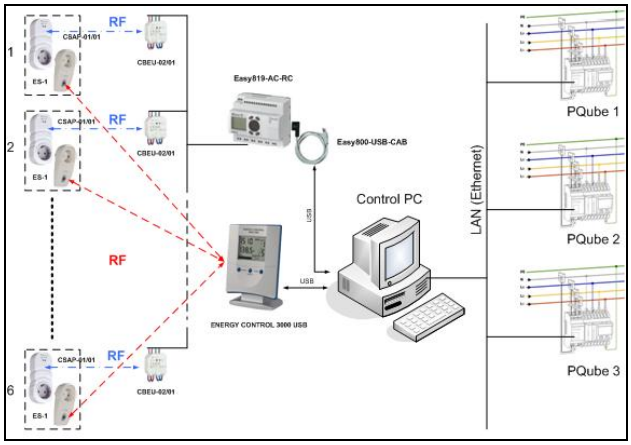


Figure 2. Overview of the PW simulation mod Proposed scheme of future system responsible for the monitoring of power quality

Before hardware implementation software predictions and calculations were made using the HOMER software. After the simulation of the PV system in HOMER all necessary hardware was bought (PV modules, inverter, communication and monitoring gear, protection equipment). For the installation 5 PV

modules with 200 W peak power per unit were chosen. Momentarily PV system is installed and ready to use.

This system has important place in the study of the smart grid and will take an important role in the scheme of the proposed testbed in this work.

D. *Hybrid system*

In order to have a complete data list from all the contributing components that are to be included in the testbed simulator, real data from the hybrid system must be included.

First the software predictions and calculations were made using the HOMER software. After the establishment of certain scenarios which are most likely to happen the search for all the necessary hardware had begun and a decision was made to buy certain hardware (PV modules, wind turbine, inverter, controller, data communicator, battery bank).

The software ANSYS was used to further determine the optimal position of the wind turbine depending on the available wind. ANSYS can give distributions of wind speeds and simulate the wind flow based on the input data of the developed model. Since the predefined place for this hybrid system is the roof of the FER building C, model in ANSYS was used to check if there is an acceptable placing available on the rooftop in regard to available wind resource.

Momentarily, due to some specific design problems of the wind aggregate pole and lack of space on the site that was predicted to be used for this hybrid system, the system has not yet been installed.

E. *Any Logic model*

The purpose of the model is to simulate a city that has implemented smart grid and demand response paradigm. There are two scopes of this model. First one is to find out what effect does implementing demand response on city level has and does it reduce peak power values. Second one is to find out if it increases security of power supply.

The city model is developed in AnyLogic software. AnyLogic is, as mentioned before, simulation tool based on Java programming language that supports system dynamics, process-centric and agent based modeling. For this model, both agent based and system dynamics modeling, were used.

The developed simulation and exemplary scenarios show that AnyLogic software can be used and fit in into the wider perspective of the emergency phenomena simulator. More can be found in [6].

Its main value lies in the possibility to easily model and generate different disturbances. Condition for that is the addition of a smart agent which will have the task to dynamically create and alter minor or major disturbances.

Once all the prerequisites are satisfied model developed in Anylogic can be in the center of the whole testbed simulator chain. Its main goal will be to provide the variety of possible problems and possible solution while not taking into account physical characteristics of a certain grid configuration but only considering the ideally modeled actions and responses.

III. INTEGRATION INTO TESTBED

This project consists of previously described parts that are all supposed to be fit together to make an integral testbed platform which will be capable of analyzing various grid configurations. The goal is to have a data acquisition system which will be constantly gathering data from all the hardware components installed. This system has a task to organize all the input data into frames and then forward the organized and converted data to the various simulation programs described previously. There various calculations will be done; Risk based analysis, physical constraints analysis, optimal power flow, agent based and dynamical modeling.

The value of the measured solar and wind resource and the corresponding production from PV and wind turbine will enable the simulation of real scenarios. Furthermore some trends and statistics can be obtained. This will not be the only task of the testbed since it will also be used to determine system response to unpredictable circumstances. Furthermore the best possible configuration will be searched for as well as the possible problematic cases through agent based modelling in AnyLogic.

The database on server in charge of data storage is proposed to be developed in Microsoft SQL and the connection between the data acquisition and data management and storage will be created using LabVIEW.

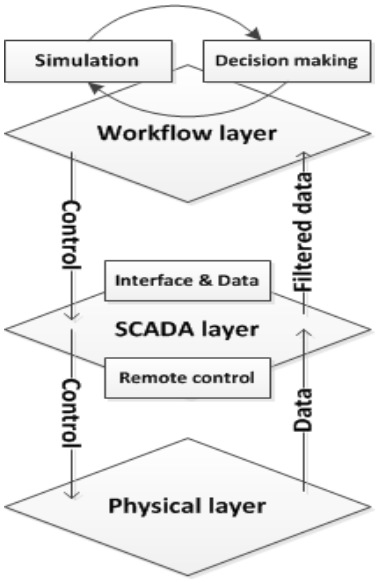


Figure 3. Conceptual layering of the proposed project

IV. LAYER APPROACH

The testbed platform consists of many different components. To simplify the development the proposed concept can be divided into 3 layers. As it can be seen from the Fig. 3 the layers are: Physical, SCADA and Workflow.

A. Physical layer

All the physical components installed come under the physical layer which is directly connected to the server and a database. The operating status of all the components will be visible in LabVIEW developed interface and on web. From web site depending of the simulation results and user requirements remote control can also be established.

LabVIEW has the possibility to collect the data from installed hardware and instruments. In the same time through predefined set of its VI-s (Virtual Instruments) in the Database Connectivity Toolkit (DCT kit) it can communicate with the database. This part of the testbed server will be gathering the data from variables that represent the measured values. After that it will store the data in created database where it will be accessible for further data mining and revision. The advantage of this approach is the possibility to monitor all the processes over the distance whether using local intranet which all the components will be connected to or global internet. The combination can enable the inclusion of remote alarming and command via the smart mobile phones or personal computers.

A LabVIEW Web Server web service which will be used to easily read the status of all the components of the system. With LabVIEW Web server, VI-s can be deployed as individual web services which can be called upon using HTTP protocol. These protocols do not require installation of any standalone run-time mechanisms what means they can be run using just HTML or JavaScript. The interface can be simply designed through usage of the LabVIEW Web UI Builder. It uses the similar graphical programming approach as LabVIEW and can be used to develop a front panel of the system which will be accessible through internet using HTTP protocol.

B. SCADA Layer

The SCADA layer can be represented by the LabVIEW interface. On the server the proposed LabVIEW interface will gather all the data and store it into the database. From there the information about the status of all the components can be sent to personal notebooks or phones via email notification or through access to the web-service of the LabVIEW interface (Fig. 4). Furthermore the data collected can be sent to various developed models and there used to generate different scenarios and results. The calculations will be done in the background in the workflow layer, but the statuses and reports will also be visible in LabVIEW interface.

C. Workflow Layer

In the workflow layer the most important parts are simulation software PowerWorld, AnyLogic and ANYSIS.

The data gathered by the server is used as input for the simulations in PW. PW can use the provided data to conduct more precise calculation. Of course certain assumptions have to be made since the developed simulation is more complex and on wider scale which means it includes more elements than are physically installed and included in data gathering process. The simulation has all the relevant elements modeled. As already mentioned, for some of them the real data from the real-time measurements is used. The rest are taken as characteristic values of certain grid elements. The similarity between the hardware system and parts of the developed simulation is high.

As stated above PW will be used to conduct calculations on a wider scale. The built low-voltage grid on FER that includes renewables and smart metering will become one part of simulated grid. Using AnyLogic, this part of the grid can be changed stochastically and multiplied through different developed algorithms similarly like it has been done

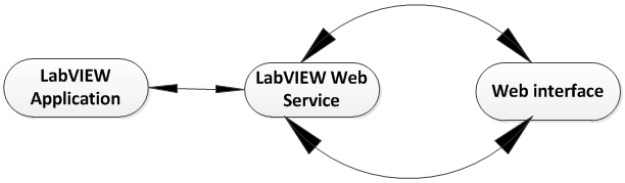


Figure 4. Web Service can communicate with the Application through common variables and shared memory

with the model of the house in the present AnyLogic model [7].

The great advantage of PW is the inbuilt script language which enables a uniform and simple change of components. This can be used, for example, to determine a new grid configuration under some specific conditions and add new parts or change already present ones. The whole graphical scheme is translated into script language. Therefore with the change of some of the code lines new elements can be added or the state of present ones can be changed. This is to be done using the Java based AnyLogic and EZ Jcom software interface as mentioned before. With cross software communication certain code fragments can be changed and by that the simulation configuration also. Certainly AnyLogic results would have to be interpreted in a certain way so that they can be efficiently used in PW. But the basic idea is feasible and can be done with the proposed tools. There are several examples of connecting JADE (Java Agent Development Framework) with PW through EZ Jcom interface [4]. Instead of using the JADE interface, AnyLogic can be used. This approach is a little less standardized but offers more ease of access and is more simple to work with [5].

Anylogic features allow building of some very complex models of the grid with decentralized decision

making and interactions. As it is shown before, simulation of the smart city has been developed using the Anylogic simulation tool

The workflow layer can be also used to expand the measured results and observed part of the electrical grid. Implemented solar and hybrid system on FER represent future LV grid alongside with the installed smart measurement. Mentioned implemented systems together with the corresponding load represent the paradigm of prosumer, a proactive consumer that has the ability to control part of the grid. Main point of many smartgrids researches is to envision how the grid containing many prosumers would work. The solution is to use simulations and expand the collected data from a small part of the grid to the whole grid.

Another link to have even more detailed insight into the grid is to be established between the PW simulation and ANSYS (Fig. 5.). PW is used to determine physical constraints and limitations, optimal power flows and voltage distribution across the network. If any of the modeled elements comes to a critical point i.e. power line overload or wind aggregate emergency shutdown due to strong wind, the more detailed simulation and calculation in ANSYS can be conducted. ANSYS models take into consideration heat transfer and dimensional constraints regarding the element position. At present time models for different types of lines and model of the air flow for the wind turbine are available. The set of these ANSYS models can be expanded.

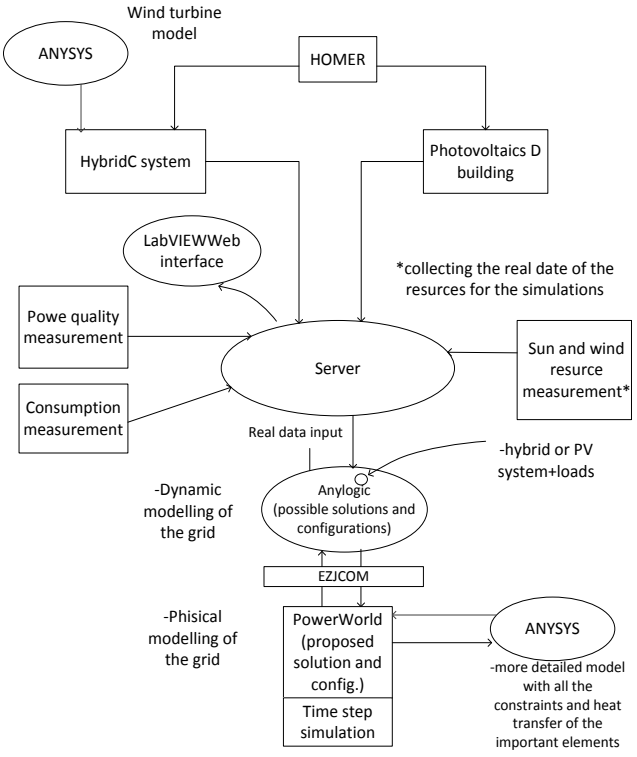


Figure 5. Detailed scheme of the testbed server

To sum up, the loop of PW and ANSYS would be able to measure the importance of certain elements and can calculate the consequences of certain actions. The PW would be getting data from the server application for some of its elements. AnyLogic would feed this loop of physical simulation tools with possible and suggested grid configurations. Also the data from measurements would be used.

D. Layers approach representation

On the Fig. 5 the detail connections between physical, SCADA and Workflow layers can be seen.

The physical layer contains PV system on the top of

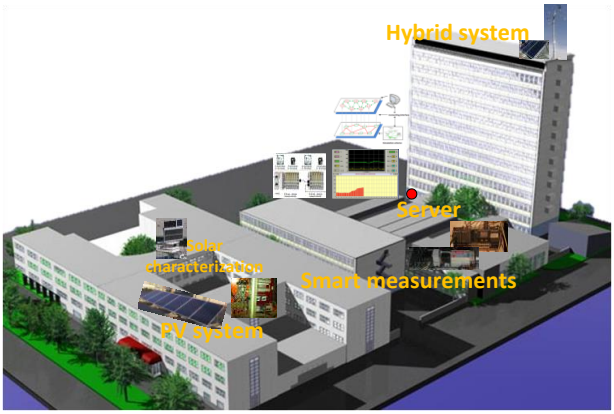


Figure 6. Physical disposition of the components

the D building and the hybrid system situated on the top of the C building (Fig. 6.). Parts of the physical layer are also the natural resources measurement, the consumption and the power quality measurement. All mentioned produces the information which is collected and processed on the server. Physical disposition of the mentioned components at FER shown on the Fig. 6.

Server can be both placed in physical and the SCADA layer. What makes the server belong to the SCADA layer is that LabVIEW will be installed on it. Through the intranet or global internet users will be able to control their production and consumption, see system status, alarms etc. from a mobile phone or a personal computer.

Furthermore, the workflow layer has been represented by the AnyLogic, PowerWorld and ANYSIS. Communication between different software tools can be established by using the EZJCOM environment as mentioned before.

V. CONLUSION

The current state of the grid is not sustainable due the explotation of the fossil fuels reserves and it needs to be changed in the sustainable smart grid. That can be done by increasing the share of the renewable resources. The main issues of smart grid which contains many renewable resources and many prosumers is the stochastic nature of the renewables and the complexity of control of the grid composed of a large amount of prosumers. There are several opinions how the transformation can be made. Many simulations must be made before the transformation of the grid can be

completely done. The „Emergent phenomena testbed simulator for improving SCADA performance in power system security management" project is motivated by mentioned solution.

This paper is the preview of the framework proposed on the mentioned project. The framework can be divided in three layers: workflow, SCADA and physical layer. Several installations are or will be implemented at the Faculty of Electrical Engineering and Computing in Zagreb as it is shown on the Fig. 6.

Physical layer contains the hybrid system, the solar system, the power quality measuring system, the consumption and the server. SCADA layer will be represented through the LabVIEW interface which can be controllable by the intranet or global internet. The server as a central unit is used as a data collecting tool and to run the LabVIEW interface on it. The data will be collected from the units of the physical layer through MODBUS interface. The collected data will also allow studding the renewable sources effect on the power quality. The access to the data will be possible online through portable units.

In the workflow layer the most important parts are simulation software PowerWorld, AnyLogic and ANYSIS. They can be used to expand the measured results and observed part of the electrical grid. In the simulations real measured data can be used as an input. Connecting the PowerWorld and the Anylogic software together gives more features to the PowerWorld. The connection will be made through the EZJcom interface. By adding some physical constraint to the PowerWorld through that connection energy storage can be simulated in the PowerWorld. That gives many opportunities for the simulations of the smart grids. AnyLogic is used for complex simulations with the decentralized decision making and interactions. The simulation of the smart city has been developed using the AnyLogic simulation tool. ANYSIS software is used to simulate the air flow around the wind turbine and also can simulate the heat transfer. Adding the heat transfer simulation can expand the PowerWorld features and allows using the full capacity of the grid in every situation.

The authors suggest using the proposed framework as a tool for further researches and the simulations in the laboratory for renewable resources which will be established at the Faculty of Electrical Engineering and Computing in Zagreb.

ACKNOWLEDGMENT

This research is sponsored by NATO's Public Diplomacy Division in the framework of "Science for Peace", project SfP-983805: Emergent phenomena testbed simulator for improving SCADA performance in power system security management.

REFERENCES

[1] Z. Šimić, J. Havelka and L. Lugarić, "Checking Numerous Smart Grid Potentials", Smart City expo, World congress, Presentation, 19 Nov. -2 Dec. 2011, Barcelona, Spain

[2] L. Lugarić, S. Krajcar and Z. Šimić, "Smart city — Platform for emergent phenomena power system testbed simulator," Proceedings of Innovative Smart Grid Technologies Conference Europe (ISGT Europe), 2010 IEEE PES, 11-13 Oct. 2010. 1-7

[3] K. Schneider, J. Fuller, F. Tuffner, "Modern Grid Strategy: Enhanced GridLAB-D Capabilities Final Report," Richland : Pacific Northwest National Laboratory, US DOE, 2009. DE-AC05-76RL01830.

[4] T. Logenthiran and D. Srinivasan, "Multi-agent system for the operation of an integrated microgrid," J. of Renewable and Sustainable Energy 4, 013116 (2012)

[5] T. Logenthiran, D. Srinivasan, A.M. Khambadkone and H.N. Aung, "Multi-Agent System (MAS) for short-term generation scheduling of a microgrid," IEEE International Conference on Sustainable Energy Technologies (ICSET), 6-9 Dec. 2010.

[6] B. Morvaj, N. Holjevac and B. Jurisic, "Stochastic Simulation of the Smart Grid and Demand Response Implementations on a City-wide Scale", (Submitted for publishing at MIPRO 2013 Conference)

[7] B. Morvaj, L. Lugaric, and S. Krajcar, "Demonstrating smart buildings and smart grid features in a smart energy city." Energetics (IYCE), Proceedings of the 2011 3rd International Youth Conference on. IEEE, 2011.

Optimal scheduling of a CHP system with energy storage

L. Majić, I. Krželj, M. Delimar
University of Zagreb/Faculty of Electrical Engineering and Computing, Zagreb, Croatia

Abstract - Microgrid concept has been proven to have a positive impact on the operation of the power system. Distributed generation and energy storage enable control of microgrid power flow which can be optimized for participation in the electricity markets. This paper presents an optimization model for economical scheduling of a microgrid using linear programming. Two microgrid models are taken into account. The microgrid proposed in the first model consists of thermal and electrical loads and a CHP unit. The second model consists of the units considered in the first model with addition of thermal and electrical storage. Optimization results for the two cases are compared to determine the impact of energy storage on an optimal scheduling. Operational costs are calculated for both cases and compared to the benchmark case where thermal energy is supplied from a gas fired boiler and all electricity is acquired from the grid. Additionally, the savings in several cases of energy storage size are compared.

Index terms – energy storage, optimal scheduling, CHP

LIST OF SYMBOLS

Q_h	heat consumption of the hospital
P_h	electicity consumption of the hospital
Q_{gb}	heat generated by the gas fired boiler
$\mathrm{Q_{gb_{max}}}$	maximum heat output of the gas fired boiler
C_{ee}	electricity price
C_g	price of natural gas
η_{gb}	efficiency of the gas fired boiler
η_{CHP_e}	electrical efficiency of the CHP unit
η_{CHP_t}	thermal efficiency of the CHP unit
Q_{CHP}	heat generated by the CHP unit
P_{CHP}	electricity generated by the CHP unit
$Q_{CHP_{max}}$	maximum heat output of the CHP unit
$P_{CHP_{max}}$	maximum electricity output of the CHP unit
P_{bat}	battery energy flow (+ for charging)
$P_{bat_{max}}$	maximum battery charge/discharge rate
E_{bat}	amount of energy stored in the battery
$E_{bat_{min}}$	minimum of energy allowed in the battery
$E_{bat_{max}}$	maximum capacity of the battery
η_{bat}	battery charge/discharge efficiency
Q_{tss}	thermal storage system energy flow (+ for charging)
$Q_{tss_{max}}$	maximum thermal storage input/output
E_{tss}	amount of energy stored in thermal storage
$E_{tss_{max}}$	maximum capacity of thermal storage
TL	thermal storage system losses

I. INTRODUCTION

Combined heat and power (CHP) technology is used to simultaneously generate electricity and useful heat in one single unit. One of the key advantages of CHP is increased energy efficiency compared to the standard divided heat and electricity production. The overall efficiency of cogeneration is between 70 to 85 percent[1] (from 27 to 45 percent of electricity and 40 to 50 percent of heat energy), compared to 50% overall efficiency when heat and power are generated separately. Cogeneration technologies play an important role as a distributed energy source because of the smaller energy loss in the network, reducing congestion in the transmission, increasing the quality and reliability of the continuous voltage supply etc. Additionally, the detrimental effect on the environment is reduced, which is nowadays one of the key elements in terms of planning and operation of the power system.

During the last twenty years, the traditional vertically integrated electric utility structure has been deregulated and replaced by a competitive market. The liberalized market (spot market) is an auction market with market clearing prices, in which generators directly sell their electricity to the market thus achieving better efficiency and higher consumer benefits. CHP advantages in this kind of market are noticeable in various aspects; firstly, it can improve efficiency of the system and secondly, it can provide consumers with heat and electricity in the optimal way. Nowadays, the vast majority of CHP installations are designed to supply only local electrical and thermal loads, mainly working in a heat driven mode, in which electrical power is generated to meet base load or for peak shaving. When electricity is sold in price market, optimal operation and planning of cogeneration is very significant and it is necessary to determine an optimal operation schedule for a short time period, based on the forecasts of electricity spot prices and predictions of the local heat and electricity demands.

In this study an optimization of proposed model and its participation in the electricity spot market is accomplished. The proposed model consists of the local consumption of electricity and thermal energy, CHP unit and heat and electricity storage system. For the purposes of this study the object instance on which such a model could be applied to is a hospital. In a hospital there is a high demand for thermal energy which makes cogeneration more cost-effective; therefore, such a system would replace the standard solution with the UPS

and diesel generator, which would partially justify the investment.

In this study the gas CHP is used because of its numerous advantages such as high reliability, low emissions, high grade heat available, small dimensions, good power to heat ratio, etc. In this paper three different scenarios and their results are presented. Optimization was conducted in GAMS using linear programming.

II. Optimization Model

A. Case I

The system considered in the first case is similar to the existing situation in hospitals. Heat demand is covered by a gas fired boiler (1), while all the electricity used is imported from the grid.

$$Q_{gb}(t) = Q_h(t) \tag{1}$$

Electricity is bought and sold at multi tariff prices, for the purposes of this paper assumed to be equal to prices on the spot market. The cost function of this system is as following:

$$z = \sum_{t=1}^{n} \left(C_g * \frac{Q_{gb}(t)}{\eta_{gb}} + P_h(t) * C_{ee}(t) \right) \tag{2}$$

This scenario has no optimization variables, the costs are calculated and compared with the costs from cases II and III to determine the benefits of setups described in those scenarios.

B. Case II

A CHP unit is introduced in case II, as shown on figure 1. The CHP unit operates in a heat following regime, covering the thermal demand (3), and the electricity output is determined taking constraints into account (5-8) in order to minimize the overall cost of the system(4).

$$Q_{CHP}(t) = Q_h(t) \tag{3}$$

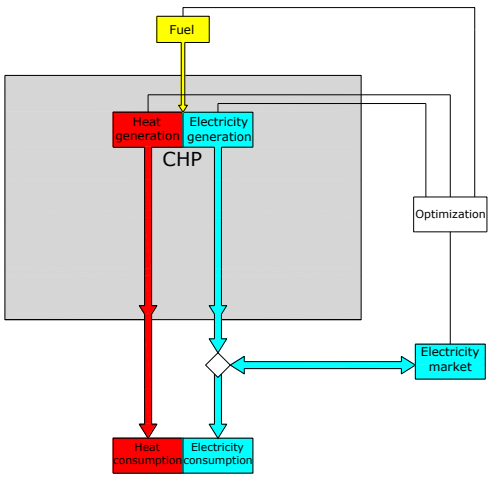


Fig. 1. Setup considered in case II

$$z = \sum_{t=1}^{n} \left(C_g * (Q_{chp}(t)/\eta_{CHP_t} + \right.$$
$$+P_{chp}(t) + 0.15 * \left(1.7P_{CHP_{max}} - P_{chp}(t)\right)\Big)$$
$$+(P_h(t) - P_{chp}(t)) * C_{ee}(t) \tag{4}$$

A feasible operation region of a typical CHP unit [1] is shown in figure 2:

$$P_{CHP}(t) < P_{CHP_{max}} \tag{5}$$

$$P_{CHP}(t) < 850 - 0.2 * Q_{CHP}(t) \tag{6}$$

$$P_{CHP}(t) > 265 - 0.0833 * Q_{CHP}(t) \tag{7}$$

$$P_{CHP}(t) < -635 + 0.4167 * Q_{CHP}(t) \tag{8}$$

The constraints consist of the maximum electrical output (5), the line of maximum (6) and minimum (7) fuel intake and the line of maximum production of heat (8). The efficiency of the CHP unit as a function of heat and electricity output is not calculated directly because it would introduce nonlinearities in the model. Efficiency is taken into account in the cost calculation (4), derived from a typical efficiency curve for a CHP unit [2]. Thermal efficiency η_{CHP_t} is assumed to be 0.85 regardless of thermal output. Hourly values for electricity generation $P_{chp}(t)$ are determined by finding the minimum value of the cost function(4).

C. Case III

In the third case thermal and battery storage are introduced in the system from the second case (fig. 3). Thermal capacity of the storage in this paper is 2500kWh, with charge/discharge rate of 1000kW. Thermal storage is defined by its capacity and rate of charge/discharge and it is constrained as follows:

$$0 < E_{tss}(t) < E_{tss_{max}} \tag{9}$$

$$-Q_{tss_{max}} < Q_{tss}(t) < Q_{tss_{max}} \tag{10}$$

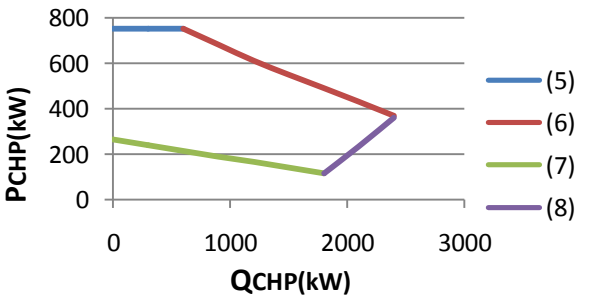


Fig. 2. CHP feasable operation region

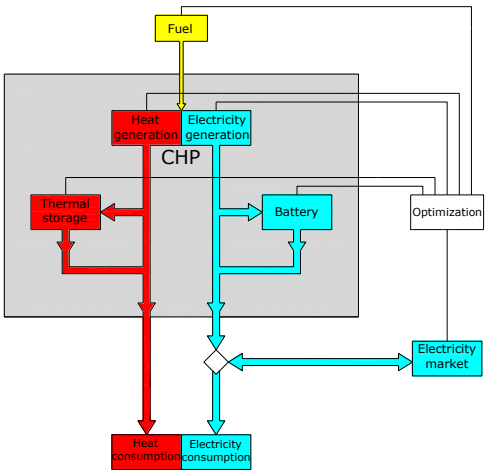


Fig. 3. Setup considered in case III

The energy balance in thermal storage system is ensured by equation (11).

$$E_{tss}(t) = (1 - TL) * (E_{tss}(t) + Q_{tss}(t)) \quad (11)$$

The heat demand is covered by the CHP unit and thermal storage (12).

$$Q_h(t) = Q_{CHP}(t) - Q_{tss}(t) \quad (12)$$

Battery is modeled similar to thermal storage and constrained by maximum and minimum capacity (13) and maximum charging/discharging rate(14,15)[3]. Minimum battery capacity is set due to the negative impact of deep discharge on battery life [4], as well as possible battery utilization in off grid operation.

$$E_{bat_{min}} < E_{bat}(t) < E_{bat_{max}} \quad (13)$$

$$-P_{bat_{max}} < P_{bat}(t) < P_{bat_{max}} \quad (14)$$

$$P_{bat}(t) < P_{bat_{max}} * (5 - 5 * E_{bat}(t)/E_{bat_{max}}) \quad (15)$$

Energy balance in the battery is modelled by equations (16, 17):

$$\forall P_{bat}(t) > 0, E_{bat}(t) = E_{bat}(t-1) + P_{bat}(t) * \eta_{bat} \quad (16)$$

$$\forall P_{bat}(t) < 0, E_{bat}(t) = E_{bat}(t-1) + P_{bat}(t)/\eta_{bat} \quad (17)$$

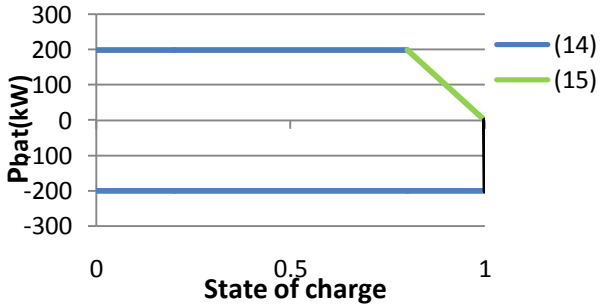


Fig. 4. Feasable operation region of the battery

Capacity of the battery system is taken to be 1000kWh, with charge/discharge rate of 200kW and efficiency of 90%, resulting in battery cycle efficiency of 81%. The cost function of this case is defined by equation (18) and its minimization results in optimal values for $P_{bat}, Q_{tss}, P_{CHP}$ and Q_{CHP} for every hour of the simulation.

$$z = \sum_{t=1}^{n} (C_g * (Q_{chp}(t)/\eta_{CHP_t} + P_{chp}(t) + 0.15 * \left(1.7P_{CHP_{max}} - P_{chp}(t)\right)) + (P_h(t) + P_{bat}(t) - P_{chp}(t)) * C_{ee}(t) \quad (18)$$

III. RESULTS

Input parameters used in this paper are electricity prices from European Electricity Exchange (EEX) and energy demand profile of a hospital in a 48 hour period. The price and demand curves refer to the same period (00:00h, 1.12.2012 – 24:00h, 2.12.2012.)

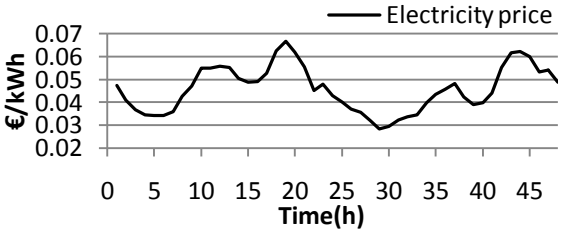


Fig. 5. Price curve

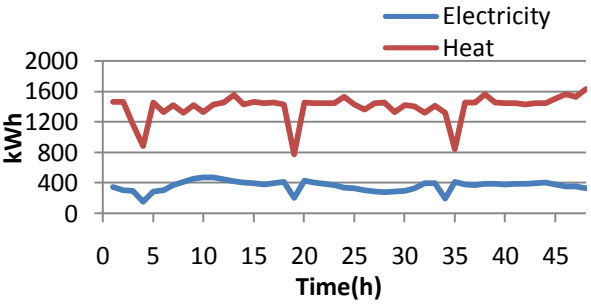


Fig. 6. Demand curves

A. Case I

The system from case I does not have any decision variables. Costs were calculated in order to determine benefits of setups from cases II and III (Table I).

B. Case II

In the second case considered in this study, optimization was done in order to determine optimal hourly scheduling of the CHP unit. As expected, electricity output was at its maximum for most of the time, except for hours where the price was lower than the CHP fuel costs (figure 7). During the low price period, it is cheaper to import the electricity than to produce it. However, the electricity generation was not entirely shut down because of one of the constraints (7).

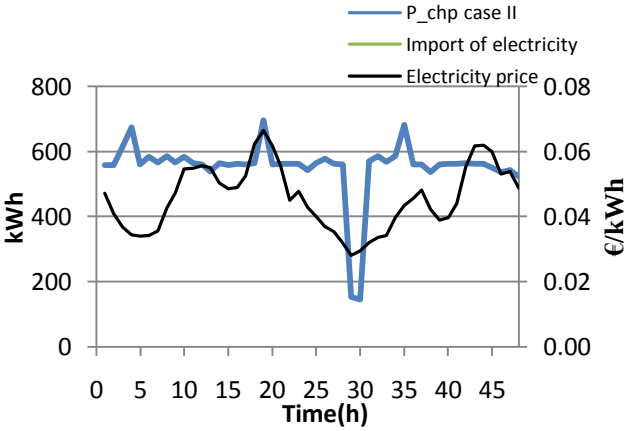


Fig. 7. Optimization result for case II

C. Case III

The final case simulated is the most complex one, with four decision variables. Since the heat demand can partially be covered by thermal storage, CHP heat generation can also be controlled to achieve better results. Thermal storage charges before the peak prices occur and discharge during peak periods, reduction the need for heat generation of the CHP, because lower heat production allows more electricity production due to constraint (6), which means selling more electricity at peak price.

Batteries also react to price volatility by absorbing energy during the hours with cheaper electricity and discharging when the prices are high (fig. 8). Heat and electricity storage behavior is similar in terms that both of them are charged before the peak prices occur.

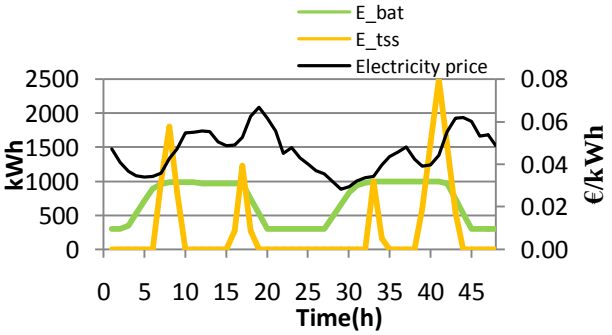


Fig. 8. Storage system behaviour

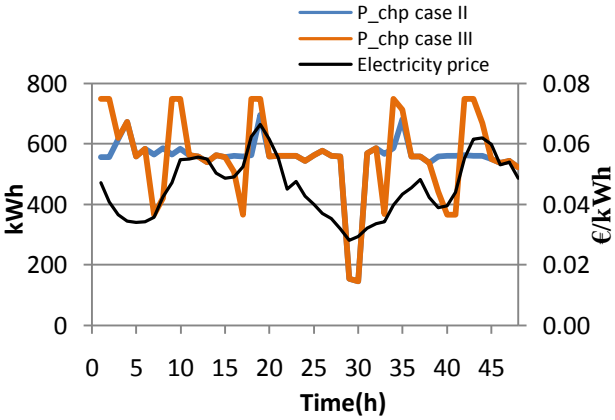


Fig. 9. Comparison of CHP electrical outputs

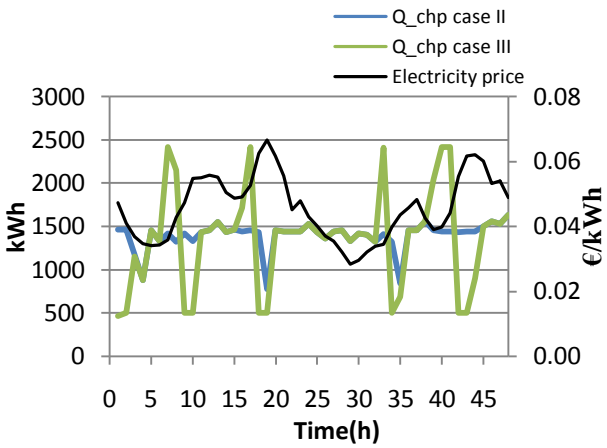


Fig. 10. Comparison of CHP thermal outputs

TABLE I COMPARISON OF SAVINGS

Scenario	Energy cost(€)	Savings compared to case I (%)
Case I	3734	-
Case II	3485	6.67
Case III	3413	8.60

However, thermal storage charges/discharges more often, since storing large amounts of energy for longer periods is discouraged due do 2% of hourly losses (11).

Comparison of CHP electrical and thermal outputs (fig. 9, 10) reveals that the addition of thermal storage in case III enabled generation of larger amounts of electricity during the peak hours and reduced the heat load of the CHP unit. However, the mode of operation presented that has proved to reduce the costs might reduce the lifespan of the CHP unit because of the frequent alternating between high and low outputs.

The financial benefits of the setups analyzed in this study are shown in Table I. Using a CHP unit has resulted in costs lower by 6.67%. Thermal efficiency of the boiler and the CHP unit are the same, which means the savings are the result of the electricity generation. Introduction of energy storage has further reduced the energy cost of the system, resulting in total savings of 8.6%. Thermal storage reduced the impact of CHP constraints and the battery storage contributed by charging in hours with low prices and discharging when the prices were high.

To further investigate the influence of energy storage, several modified versions of case III have been simulated. Storage capacities and charge/discharge rates were varied from 50% to 200% of the initial values (fig. 11). Savings were calculated with respect to case I for three variants: battery storage only, thermal storage only and the usage of both electricity and heat storage. Batteries have shown to increase savings proportionally with their size, since the energy demand they are able to transfer from high price periods to low price periods is linearly related to their capacity and charge/discharge rate. On the other hand, thermal storage increase has lower impact on the costs, because after a certain point, they allow the CHP

unit to generate electricity in the optimal operating point, and further increase does not benefit the overall system performance.

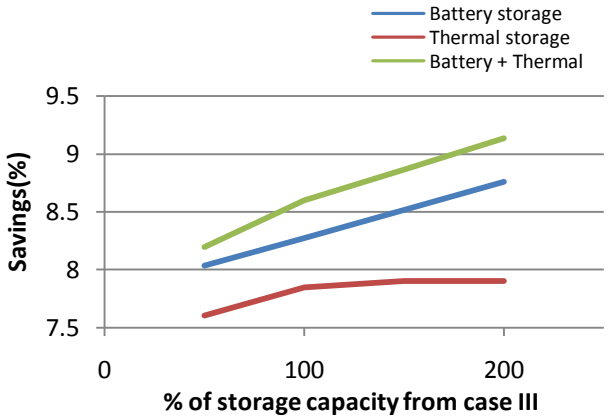


Fig. 11. Impact of storage sizing on savings

IV. CONCLUSION

This paper is an analysis of the possibility of reducing costs by using a CHP unit in a building with a large thermal demand, for example a hospital. Running a CHP unit in a heat following mode can result in inefficient scheduling of electrical output if electricity is bought and sold at variable prices. By introducing heat storage it is possible to transfer thermal demand in a way to improve the performance of a CHP unit and reduce the costs. Additionally, using battery storage enables similar manipulation with electrical demand by charging and discharging according to electricity prices.

Onsite energy generation combined with energy storage reduces energy dependence, which is an important feature for high priority institutions such as hospitals. The system described in case III of this paper could be configured to provide power if the supply from grid is interrupted.

The purpose of this paper was to illustrate the possible benefits of the proposed system, but for more conclusive results a much larger period of time should be considered, taking into account the different energy demand profiles throughout the year. Additionally, this study has not addressed the capital investment needed to install the system, which is crucial when analyzing the potential benefits.

REFERENCES

[1] Fabricio Salgado, Pedro Pedrero, “Short-term operation planning on cogeneration systems: A survey”, Electric Power Systems Research 78 (2008) 835–848

[2] U.S. Environmental Protection Agency Combined Heat and Power Partnership, “Catalog of CHP technologies”, 2008

[3] S. Dearbon, “Charging Li-ion batteries for maximum run times," Principal Applications Engineer, MicrochipTechnology, Chandler, Ariz.

[4] T. Guena and P. Leblanc, “How depth of discharge affects the cycle life of lithium-metal-polymer batteries,"Telecommunications Energy Conference, 2006. INTELEC '06. 28th Annual International

[5] M. Pirouti, J. Wu, A. Bagadanavicius, J. Ekanayake and N. Jenkins, “Optimal operation of biomass combined heat and power in a spot market," PowerTech, 2011 IEEE Trondheim

[6] D. Brujic, M. Ristic and K. Thoma, “Optimal operation of distributed CHP systems for participation in electricity spot markets," EUROCON, 2007. The International Conference on "Computer as a Tool"

Poboljšanje performansi bežičnih mreža korištenjem relejnih postaja

Andrej Caplić, Željko Ilić
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
E-mail: {andrej.caplic, zeljko.ilic}@fer.hr

***Sažetak* – Tijekom posljednjih desetak godina veliki interes je usmjeren na uređaje koji za komunikaciju koriste bežični medij. Problem koji proizlazi iz te činjenice je omogućavanje pristupa bežičnoj mreži što većem broju korisnika. Bežični uređaji, tj. pokretna korisnička oprema u mnogočemu je ograničena, ta ograničenja uključuju prijenosnu brzinu (kapacitet), domet prijenosa, kapacitet baterije, itd.**

U situacijama kada izravna komunikacija između bazne postaje i pokretne korisničke opreme nije moguća signal se šalje preko niza relejnih postaja. Takva telekomunikacijska mreža naziva se višeskočna relejna mreža (engl. *Multihop Relay Network*) i predstavlja jedno od rješenja povećanja dometa prijenosa signala u bežičnim mrežama četvrte generacije (4G). Gledano na način na koji se signal obrađuje unutar pojedinog releja dana mreža uključuje dva principa rada, i to: pojačaj i proslijedi (engl. *Amplify and Forward*, AF) te dekodiraj i proslijedi (engl. *Decode and Forward*, DF).

U ovome članku istaknute su karakteristike navedenih tehnika obrade signala te je opisan model kanala na temelju kojeg je napravljena usporedba navedenih tehnika u pogledu prijenosnog kapaciteta uzimajući pri tome u razmatranje broj relejnih postaja. Također, autori iznose vlastitu formulu ovisnosti promjene kapaciteta AF mreže o broju skokova.

Proračunski rezultati pokazuju da tehnika DF pruža veću spektralnu iskoristivost, odnosno veći prijenosni kapacitet ali uz unos dodatnog kašnjenja što predstavlja kritičan faktor kod usluga u realnom vremenu.

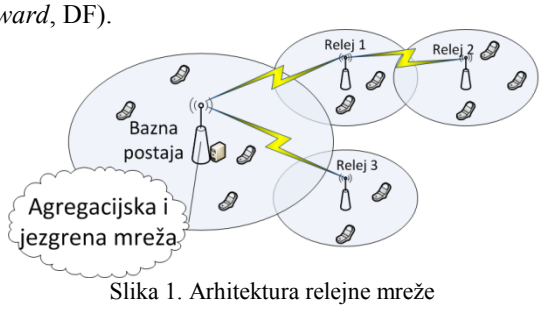
***Ključne riječi* – bežične komunikacije, relejna postaja, tehnika pojačaj i proslijedi, tehnika dekodiraj i proslijedi.**

I. UVOD

Višeskočne relejne mreže (engl. *Multihop Relay Networks*) u posljednje vrijeme privlače veliku pažnju znanstvene zajednice. To je iz razloga što predstavljaju obećavajuće rješenje za problem koji je trenutno vrlo aktualan u svijetu bežičnih komunikacija četvrte generacije (4G), a to je domet prijenosa. Iz razloga što su centralna frekvencija, kao i prijenosni pojas, relativno visoko postavljeni, 5 GHz te 100 MHz, slijedno gledano, takvi su sustavi vrlo osjetljivi na prigušenje signala te na taj način ostvaruju manje područje pokrivanja [1]. Naime, opće poznata stvar je da u bežičnim mrežama porastom frekvencije prigušenje signala raste, tj. snaga signala opada. Korištenjem relejnih postaja koje će prosljeđivati signal (poslan od bazne postaje do korisničke opreme) navedeni problem može biti značajno otklonjen.

Relejem možemo smatrati baznu postaju koja se najčešće nalazi u liniji vidljivosti s ostalim relejnim postajama ili glavnom baznom postajom ali za razliku od glavne bazne postaje ni na koji način nije povezana s agregacijskom, tj. jezgrenom mrežom, već prvenstveno služi kao među postaja, (Slika 1).

Tehnike relejnog prijenosa, obzirom na način na koji se signal obrađuje unutar pojedinog releja, se dijele u dvije osnovne grupe: pojačaj i proslijedi (engl. *Amplify and Forward*, AF) te dekodiraj i proslijedi (engl. *Decode and Forward*, DF).



Slika 1. Arhitektura relejne mreže

Postoji velik broj znanstvenih radova koji se bave problematikom relejnih mreža. U [2] i [3] napravljena je analiza performansi (utjecaj povećanja broja antena i skokova na prijenosni kapacitet) AF i DF relejnih mreža. Osnovni pojmovi vezani uz komunikacije (kapacitet kanala, prijenosna funkcija sustava, fenomeni širenja elektromagnetskog vala kroz prostor) dani su u [4] – [6]. Nadalje, u [7] - [11] obrađena tematika vezana je uz kapacitet relejnih mreža, dok je u [12] i [13] opisana implementacija relejnih mreža u tehnologiji bežičnog prijenosa WiMAX. Konačno, u [14] fokus predstavljaju performanse AF relejnih mreža s pojačanjem fiksnog iznosa.

Dani članak je organiziran na sljedeći način. U II. poglavlju prikazane i analizirane su dvije najpoznatije tehnike relejnog prijenosa (AF i DF). Zatim se u III. poglavlju definira model relejnog kanala pomoću kojeg je u IV. poglavlju iskazan kapacitet DF i AF relejnih tehnika u SISO (engl. *Single Input Single Output*) kanalu pri čemu je svakoj relejnoj postaji poznato stanje bežičnog kanala (engl. *Channel-State Information*, CSI). Kapacitet je predstavljen kao funkcija broja skokova i omjera srednje snage signala prema srednjoj snazi šuma (engl. *Signal to Noise Ratio*, SNR). Također je napravljena i usporedba kapaciteta za DF i AF relejne mreže. U posljednjem, V. poglavlju, iskazan je utjecaj broja skokova na kapacitete DF i AF relejnih mreža pri čemu je pokazano da promjenom broja skokova utječemo na kapacitet relejne mreže, poglavito u području visokog omjera SNR i malog broja relejnih postaja. Na kraju članka nalazi se zaključak kao i popis korištene literature.

II. TEHNIKE RELEJNOG PRIJENOSA

U danom poglavlju navedene su osnovne značajke prethodno spomenutih tehnika relejnog prijenosa – AF i DF.

A. Pojačaj i proslijedi (AF)

Ovu vrstu prosljeđivanja možemo podijeliti u dvije podvrste, pojačanje fiksnog iznosa (engl. *Fixed Gain Amplification*) i tehnika adaptivnog prosljeđivanja (engl. *Adaptive Forwarding*) [2] [3]. Iako je dan opis obje tehnike u nastavku članka pod tehnikom AF se podrazumijeva tehnika pojačanja fiksnog iznosa.

Kod pojačanja fiksnog iznosa releji prvo pojačavaju snagu primljenog signala te ga zatim odašilju do sljedećeg releja ili pokretne korisničke opreme [2]. Iznos pojačanja snage signala unaprijed je određen. Takva tehnika predstavlja najjednostavniji pristup koji uključuje najmanji stupanj obrade na svakom pojedinom releju (unosi minimalno vrijeme obrade - kašnjenje). Važno je napomenuti da se kod ove tehnike uz pojačavanje izvornog signala pojačava i šum što uvelike smanjuje performanse. Naime, svaki je relej (možemo ga smatrati pojačalom) obilježen dvama faktorima pojačanjem (u odnosu na izotropni radijator) i faktorom šuma (engl. *Noise Figure*, NF). Faktor šuma, je omjer između SNR-a na ulazu u relej i SNR-a na izlazu iz releja.

$$NF = \frac{SNR_{ulaz}}{SNR_{izlaz}} = \frac{\frac{P_{signal_{ulaz}}}{P_{šum_{ulaz}}}}{\frac{P_{signal_{izlaz}}}{P_{šum_{izlaz}}}} \qquad NF\ [dB] = SNR_{ulaz}[dB] - SNR_{izlaz}[dB]$$

U idealnim uvjetima prijenosa SNR na ulazu i izlazu releja trebao bi biti isti, no zbog same prirode pojačanja signala proizlazi činjenica da se pri pojačanju generira dodatan šum koji se zbraja na pridošli šum. Iz tog razloga SNR na izlazu i ulazu nije jednak, štoviše SNR na izlazu je manji od SNR-a na ulazu. Ova tehnika se koristi u situacijama u kojima je potrebno povećati područje pokrivanja i pokrivanje područja u kojima problem stvara sjenjenje.

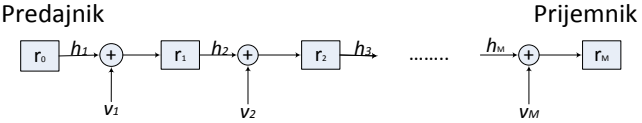
Tehnika adaptivnog prosljeđivanja (engl. *Adaptive Forwarding*) u pravilu predstavlja mnogo napredniju tehniku od pojačanja fiksnog iznosa [3]. Naime, pojačanje signala na svakome releju ovisi o trenutnom stanju kanala (CSI), stoga je potrebno implementirati posebne uređaje koji procjenjuju kvalitetu kanala u svakome trenutku iz čega proizlazi kompleksnost ove tehnike. Procjena kanala se najčešće ostvaruje slanjem slijeda poznatih bitova na različitim frekvencijama [3]. Usporedbom (na odredištu) primljenog i onoga što je trebalo biti primljeno na jednostavan se način može odrediti prijenosna funkcija sustava te se pomoću lako može prilagoditi na uvjete u komunikacijskom kanalu.

B. Dekodiraj i proslijedi (DF)

U relejnim mrežama u kojima je implementiran ovaj način rada releji prvo dekodiraju primljeni sadržaj bitova zaprimljen u prvoj fazi slanja te provode korekciju pogrešaka prije nego što sadržaj proslijede do sljedećeg releja ili pokretne korisničke opreme. Podaci se prosljeđuju jedino u slučaju ako zaprimljeni bitovi ne sadrže pogrešku. Pogreška na bitovima provjerava se pomoću cikličke provjere redundancije (engl. *Cyclic Redundancy Check*, CRC) [2]. U slučaju pogreške šalje se zahtjev za retransmisijom. Kako bi se broj retransmisija neispravno primljenih paketa minimizirao kod ove se tehnike koristi hibridni automatski zahtjev za ponovljenim slanjem podataka (engl. *hybrid Automatic Repeat reQuest*) [4]. DF tehnika pokazuje najbolje performanse u uvjetima kada je kanal relativno stabilan te nije pod utjecajem degradirajućih faktora (kao što su višestazna propagacija, sjenjenje, ogib). Nedostatak ove tehnike unošenje je dodatnog kašnjenja uslijed obrade signala.

III. MODEL KANALA

Kao što je već otprije rečeno, komunikacija se u višeskočnoj relejnoj mreži odvija preko niza releja. Obzirom na tu činjenicu, kanal definiramo na način predočen na slici (Slika 2).



Slika 2. Model kanala višeskočne relejne mreže

Primljeni signal na *i*-tom releju dan je sljedećom formulom [2][3]:

$$y_i = \sqrt{P_s} * h_i * x_{i-1} + v_i, \quad i = 1, 2, \dots, M;$$

pri čemu P_s, h_i i v_i označavaju odaslanu snagu signala s releja *i*-1 (pretpostavljamo da je odaslana snaga konstantna, tj. da svi releji odašilju istom snagom), kvazi-statičan frekvencijski selektivan kanal s fedingom modeliranim pomoću kompleksnih, neovisnih, jednoliko raspoređenih Gaussovih varijabli srednje vrijednosti nula (engl. *Zero-Mean, Circulary Symmetric, Complex Gaussian*, ZMCSCG) te aditivni bijeli Gaussov šum, slijedno gledano. Također, pretpostavljamo da je varijanca šuma (σ_v^2) jednaka za sve releje. Iz razloga jednake udaljenosti među svim parovima susjednih čvorova (koji se nalaze na zamišljenoj ravnoj liniji) varijanca kanala je konstantna vrijednost. Iz razloga što je statistika svakog kanala među svim parovima susjednih releja jednaka izbacujemo oznaku *i* te zapisujemo [2]:

$$E_h\{|h_i|^2\}_{i=1}^{M} = E_h\{|h|^2\} = \xi_M = \xi_1 * M^{\alpha}$$

pri čemu *E*, ξ_1, ξ_M i α označavaju statističko očekivanje varijable h (dobitak kanala), varijancu kanala koji predstavlja direktnu vezu između izvorišta i odredišta, varijancu kanala sačinjenog od *M* skokova (broj skokova jednak je broju relejnih postaja uvećanim za jedan) te faktor prigušenja staze, slijedno gledano. Iz dane formule vidljivo je da se povećanjem broja skokova (parametar *M*), odnosno smanjenjem međusobne udaljenosti među postajama, varijanca kanala povećava. Vrijednosti α razlikuju se prema okolini što je dano u [5].

Ukoliko uvedemo veličinu koja opisuje maksimalan broj skokova (M_{max}), tada iz prethodne jednakosti dobivamo [2]:

$$\frac{\xi_M}{\xi_{M_{max}}} = \frac{M^{\alpha} * \xi_1}{M_{max}^{\alpha} * \xi_1} = \left(\frac{M}{M_{max}}\right)^{\alpha}$$

Korištenjem maksimalnog broja skokova dobitak kanala iznosi jedan ($\xi_{Mmax} = 1$). Drugim riječima to znači da je postavljeno toliko relejnih postaja da nema pogoršanja signala. Uzevši to u obzir prethodnu formulu možemo zapisati kao [2]:

$$\xi_M = \left(\frac{M}{M_{max}}\right)^{\alpha} \triangleq V$$

Uzimajući u obzir činjenicu da su svi kanali modelirani kao ZMCSCG varijable proizlazi da razdioba amplitudnog

spektra kanala prati χ^2 razdiobu, s dva stupnja slobode [2]. Ukoliko q označava $|h|^2$, funkcija gustoće vjerojatnosti (engl. *Probability Density Function*, PDF) i kumulativna funkcija razdiobe (engl. *Cumulative Distribution Function*, CDF) slučajne varijable $q \sim \chi_2^2$ se zapisuju kao [2]:

$$f_Q(q) = \frac{1}{V} * \exp\left(-\frac{q}{V}\right)$$
$$F_Q(q) = 1 - \exp\left(-\frac{q}{V}\right)$$

Na ovaj način trenutni CSI, varijable h_i, poznat je svakoj relejnoj postaji te na taj način kompenziramo promjenu faze odaslanog signala uslijed propagacije. Ovisno o tehnici relejnog prijenosa, relejne postaje na određeni način obrađuju dobiveni signal te ga prosljeđuju prema sljedećoj relejnoj postaji ili korisničkoj opremi.

IV. KAPACITET KANALA

Kapacitet kanala predstavlja matematičku veličinu kojom iskazujemo maksimalnu količinu informacije po simbolu koja se u prosjeku može prenijeti komunikacijskim kanalom [6]. Izraz kojim definiramo kapacitet kanala glasi:

$$C = Blog_2\left(1 + \frac{S}{N}\right) = Blog_2\left(1 + \frac{S}{N_0 B}\right)\ [bit/s],$$

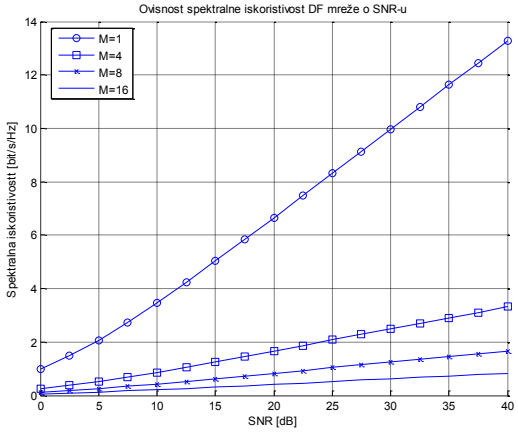
pri čemu *B, S* i *N* označavaju širinu frekvencijskog pojasa, srednju snagu signala te srednju snagu šuma [5]. Kako bi pojednostavili proračun, u nastavku ćemo računati normalizirani kapacitet (u obzir nećemo uzimati širinu frekvencijskog pojasa), tako da ćemo zapravo računati iskoristivost prijenosa izraženu u jedinici [*bit/s/Hz*]. Također, razmatrati će se kapacitet SISO relejnih mreža [11].

A. Kapacitet DF relejne mreže

Ukupan kapacitet DF sustava dan je formulom [2]:

$$C_{DF} = \frac{1}{M} min\{c_1, c_2, c_3, \dots, c_M\}$$
$$= \frac{1}{M} log_2(1 + min\{\gamma_1, \gamma_2, \gamma_3, \dots, \gamma_M\}),$$

pri čemu *M,* γ_i i c_i predstavljaju broj skokova, SNR za *i*-ti skok te kapacitet *i*-tog kanala, slijedno gledano.



Slika 3. Spektralna iskoristivost (kapacitet) DF mreže u ovisnosti o SNR-u i broju skokova

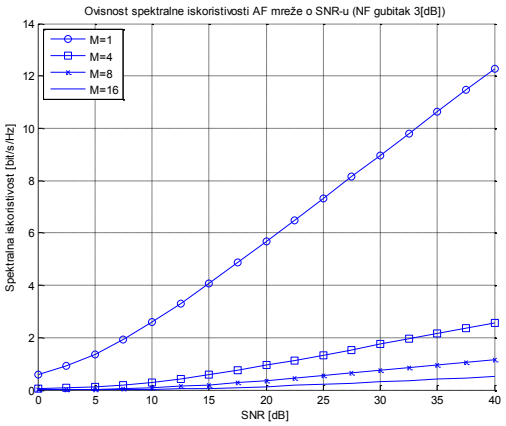
Na slici (Slika 3) može se primijetiti povećanje spektralne iskoristivost DF mreže uslijed povećanja omjera SNR. No, povećanjem broja relejnih postaja spektralna iskoristivost ulazi u zasićenje, iz čega se zaključuje da kod projektiranja mreža treba paziti na broj relejnih postaja kako ne bismo ozbiljno degradirali performanse mreže.

B. Kapacitet AF relejne mreže

Ukupan kapacitet AF sustava dan je formulom [2]:

$$C_{AF} = \frac{1}{M} log_2(1 + \gamma_{total}),$$

pri čemu *M* i γ_{total} predstavljaju broj skokova, odnosno omjer SNR na odredištu.



Slika 4. Spektralna iskoristivost (kapacitet) AF mreže u ovisnosti o SNR-u, broju skokova i uz *NF* od 3 dB

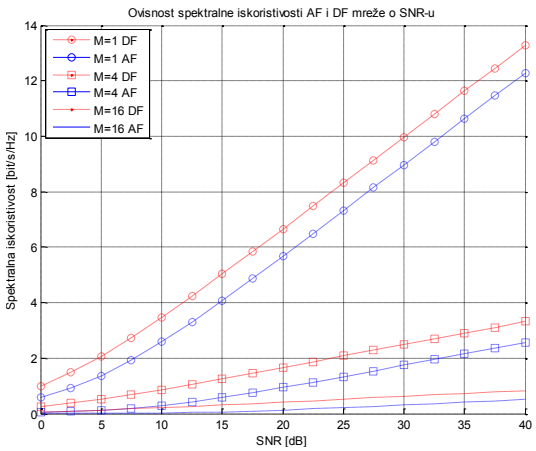
Na slici (Slika 4) primjećujemo identičnu situaciju kao i na slici (Slika 3), tj. povećanjem omjera SNR povećavamo spektralnu iskoristivost sustava. No, isto tako, povećanjem broja relejnih postaja spektralna iskoristivost ulazi u zasićenje.

C. Usporedba kapaciteta AF i DF mreže

Spomenimo još jedanput način rada AF relejnih postaja; dolaskom na relej, korisnički signal (koji je već pod utjecajem određene količine šuma) zbraja se sa šumom prijemnika te se zatim (to sve) zajedno pojačava. Jasno je da će dolaskom do odredišnog čvora korisni signal biti višestruko degradiran. Iz toga proizlazi činjenica da je:

$$\gamma_{total} \leq \gamma_k = min\{\gamma_1, \gamma_2, \gamma_3, \dots, \gamma_M\},$$

odnosno: $C_{AF} \leq C_{DF}$.



Slika 5. Usporedba kapaciteta AF i DF mreže u ovisnosti o SNR-u i broju skokova

Na slici (Slika 5) je vidljivo da DF mreža pokazuje bolje performanse (spektralnu iskoristivost) u odnosu na AF mrežu uz isti omjer SNR. Iz toga zaključujemo da bi DF mreža uz jednake uvjete u kanalu i jednak prijenosi pojas ostvarivala veći prijenosni kapacitet.

V. UTJECAJ BROJA SKOKOVA NA KAPACITET I PRIJENOSNU FUNKCIJU RELEJNIH MREŽA

A. Utjecaj broja skokova na kapacitet DF relejne mreže

Kapacitet svakog pojedinačnog skoka u DF mreži dan je formulom [2]:

$$c_i = log_2\left(1 + \frac{P_S}{\sigma_v^2}|h_i|^2\right),\ i=1, 2, \dots M.$$

Pod pretpostavkom da su sve relejne postaje jednako razmaknute i da u svim kanalima vladaju jednaki uvjeti proizlazi činjenica da je funkcija gustoće za sve kanale jednaka i iznosi:

$$f_{C'}(c') = \exp\left(-\frac{2^{c'}-1}{\rho * V}\right)\frac{2^{c'} * \ln(2)}{\rho * V}$$

pri čemu je $c' = log_2\left(1 + \frac{P_S}{\sigma_v^2}|h_i|^2\right)$ i $\rho = \frac{P_S}{\sigma_v^2}$.

Nadalje, potrebno je odrediti očekivanje kapaciteta DF mreže. Od prije je poznato da kapacitet DF mreže iznosi $C_{DF} = \frac{1}{M} min\{c_1, c_2, c_3, \dots, c_M\}$. Stoga, potrebno je odrediti koji je kapacitet najmanji po iznosu. Neka c_j predstavlja najmanji od svih kapaciteta u mreži. Matematički zapisano: $C_{DF} * M = c_j \quad za \quad c_j \le c_k, k = 1, 2, 3, \dots, j-1, j+1, \dots, M$. Obzirom na prethodne izraze za očekivanje kapaciteta DF mreže dobiva se sljedeća jednakost:

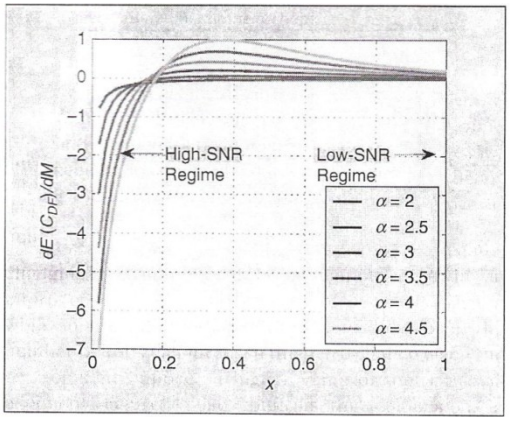
$$E\{C_{DF}\} = \frac{1}{M}\int_0^\infty c'' f_{C''}(c'')dc'' = \frac{Ei\left(\frac{-M}{\rho * V}\right) * \exp\left(\frac{M}{\rho * V}\right)}{M * \ln(2)},$$

pri čemu $Ei(x)$ označava eksponencijalni integral [2].

Kako bi uvidjeli kako promjena broja skokova utječe na promjenu kapaciteta relejne mreže, parcijalno ćemo derivirati očekivani kapacitet po broju skokova, *M*. Iako je broj skokova u stvarnosti cijeli broj a parcijalna se derivacija odvija na kontinuiranom intervalu, ovim postupkom dobivamo „grubi“ pogled u prirodu stvari.

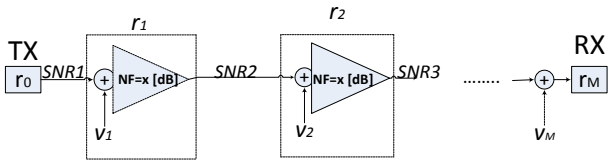
$$\frac{\partial E\{C_{DF}\}}{\partial M} = \frac{exp\left(\frac{{M_{max}}^\alpha}{\rho M^{\alpha-1}}\right) Ei\left(-\frac{{M_{max}}^\alpha}{\rho M^{\alpha-1}}\right)}{\frac{\left[\frac{{M_{max}}^\alpha}{\rho M^{\alpha-1}} * (\alpha-1) + 1\right] + (\alpha-1)}{(M^2 * \ln(2))}}$$

Na slici (Slika 6) se može primijetiti kako u područjima malog SNR-a promjena broja skokova ne može rezultirati poboljšanjem performansi prijenosa, dok u području velikog SNR-a broj skokova kao i faktor prigušenja okoline (α) imaju utjecaj na performanse mreže.



Slika 6. Ovisnost kapaciteta DF mreže o parametru *M* (broj skokova) uz $x = \left(\frac{M_{max}}{M}\right)\rho^{-\frac{1}{\alpha-1}}$ i $M_{max} = 10$, [2]

B. Utjecaj broja skokova na kapacitet AF relejne mreže



Slika 7. Model arhitekture AF mreže

Kao što je već navedeno, kapacitet AF mreža dan je sljedećom formulom [2]:

$$C_{AF} = \frac{1}{M} log_2(1 + SNR_{total}).$$

Pri čemu SNR_{total} predstavlja omjer signal-šum na odredištu. Na slici (Slika 7) može se primijetiti sljedeći odnos između SNR-ova susjednih kanala:

$$\begin{aligned} SNR_2[dB] &= SNR_1[dB] - NF[dB], \\ SNR_3[dB] &= SNR_2[dB] - NF[dB] \\ &= SNR_1[dB] - NF[dB] - NF[dB] \\ &= SNR_1[dB] - 2 * NF[dB] \end{aligned}$$

$$\dots$$

$$SNR_{total}[dB] = SNR_1[dB] - M * NF[dB],$$

$$SNR_{total} = \frac{SNR_1}{M * NF} = \frac{\frac{P_S}{\sigma_v^2}\left(\frac{M}{M_{max}}\right)^\alpha}{M * NF},$$

pri čemu SNR_1, *M*, *NF* označavaju omjer signal-šum na početnom releju, broj skokova te faktor šuma svakog releja. Pretpostavka je da su svi releji jednako razmaknuti te da su faktori šuma svih releja iznosom jednaki: $NF_1 = NF_2 = \dots = NF_{M-1} = NF\ [dB]$

Koristeći gore navedene jednakosti za kapacitet AF mreže dobivamo:

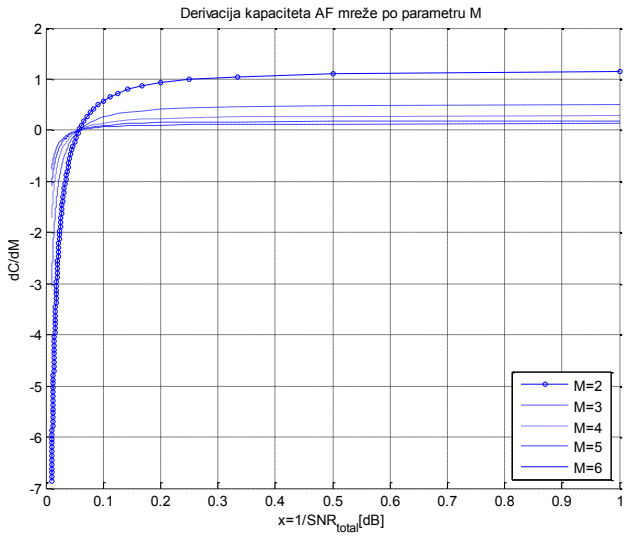
$$C_{AF} = \frac{1}{M} log_2\left(1 + \frac{\frac{P_S}{\sigma_v^2}\left(\frac{M}{M_{max}}\right)^\alpha}{M * NF}\right) = \frac{1}{M} log_2(1 + SNR_{total})$$

Kako bi prikazali ovisnost promjene ukupnog kapaciteta o broju skokova potrebno je derivirati ukupni kapacitet po varijabli M, koja predstavlja broj skokova:

$$\frac{\partial C_{AF}}{\partial M} = \frac{M * \frac{1}{\left(1 + \frac{\frac{P_S}{\sigma_v^2}\left(\frac{M}{M_{max}}\right)^\alpha}{M * NF}\right) * \ln(2)} * \frac{(\alpha-1)}{M^2 * NF} * \frac{P_S}{\sigma_v^2}\left(\frac{M}{M_{max}}\right)^\alpha - log_2\left(1 + \frac{\frac{P_S}{\sigma_v^2}\left(\frac{M}{M_{max}}\right)^\alpha}{M * NF}\right)}{M^2}$$

Ukoliko se prethodni izraz sredi, konačno dobivamo:

$$\frac{\partial C_{AF}}{\partial M} = \frac{\frac{(\alpha-1)*SNR_{total}}{(1+SNR_{total})*ln(2)} - log_2(1+SNR_{total})}{M^2}$$



Slika 8. Ovisnost kapaciteta AF mreže o parametru *M* (broj skokova)

Na slici (Slika 8) je moguće zamijetiti da u prisutnosti malog SNR-a (desna strana slike) promjena broja skokova ne utječe značajno na promjenu kapaciteta sustava. U situacijama kada je omjer SNR relativno velik (govorimo o vrijednosti SNR-a većim od 20 dB) broj skokova značajno utječe na promjenu kapaciteta relejne mreže. Također se može primijetiti da će u mreži s malim brojem releja, promjena broja releja puno više utjecati na promjenu kapaciteta nego što će to biti slučaj u sustavima s većim brojem releja. Pojednostavljeno rečeno, u sustavu s jednim relejom dodavanjem još jednog releja puno više utječemo na promjenu kapaciteta nego što bi to učinili u sustavu s pet releja dodavanjem šestog.

C. Utjecaj broja skokova na prijenosnu funkciju AF relejnih mreža

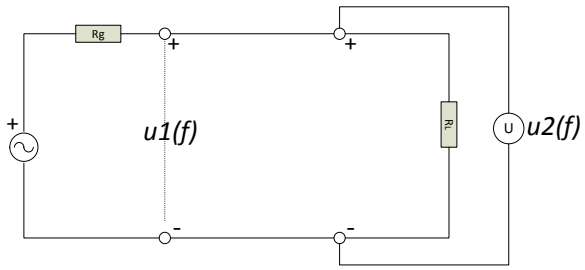
Uz prijenosne kanale vežu se dva bitna pojma, a to su impulsni odziv – $h(t)$ i prijenosna funkcija – $H(f)$. Može se reći da je u impulsnom odzivu, odnosno prijenosnoj funkciji zapisana „priroda“ sustava. Impulsni odziv i prijenosna funkcija predstavljaju Fourierov transformacijski par, odnosno:

$$H(f) = \int_{-\infty}^{\infty} h(t)e^{-j2\pi ft}dt$$

$$h(t) = \int_{-\infty}^{\infty} H(f)e^{j2\pi ft}df$$

Impulsni odziv predstavlja odziv sustava (kanala, filtera, pojačala, ...) na impuls, najčešće Diracov delta impuls jedinične površine: $\int_{-\infty}^{\infty}\delta(t)dt = 1$, $\delta(t)\begin{cases} = 0 \; za \; t \neq 0 \\ \neq 0 \; za \; t = 0 \end{cases}$

U linearnim invarijantnim sustavima (engl. *Linear Time-Invariant System*, LTI) odziv na Diracov delta impuls predstavljen je funkcijom oblika $\sin(x)/x$. U realnim sustavima impulsni odziv je gotovo nemoguće dobiti iz razloga što je pomoću suvremene opreme nemoguće generirati Diracov delta impuls. Iz tog razloga prijenosna funkcija predstavlja primarni parametar prijenosnih sustava. Prijenosna funkcija prikazuje kako sustav utječe na pojedine frekvencijske komponente, tj. pojačava li ih ili ih prigušuje. Jedan od načina dobivanja prijenosne funkcije sustava prikazan je na slici (Slika 9).



Slika 9. Model za mjerenje prijenosne funkcije LTI sustava

$$|z_0| = R_g = R_L$$

$$|H(f)| = \left|\frac{u_2(f)}{u_1(f)}\right|$$

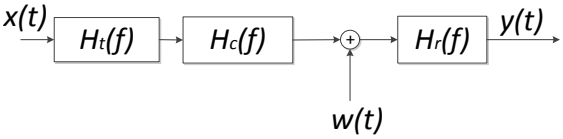
Prijenosna funkcija predstavlja omjera napona (u frekvencijskoj domeni) na trošilu i na izvoru pri čemu je amplituda ulaznog signala kontinuirana dok se frekvencija mijenja. Ukoliko se radi o izravnoj povezanosti bazne postaje i korisničke opreme (Slika 10), tada je prijenosna funkcija sustava jednaka:

$$H(f) = H_t(f)H_c(f)H_r(f),$$

odnosno u vremenskoj domeni:

$$h(t) = h_t(t) * h_c(t) * h_r(t),$$

pri čemu operator * predstavlja konvoluciju.



Slika 10. Model za opis prijenosne funkcije

$H_t(f)$ predstavlja prijenosnu funkciju predajnika, $H_c(f)$ predstavlja prijenosnu funkciju kanala, $H_r(f)$ predstavlja prijenosnu funkciju prijamnika, x(t) je signal na ulazu u sustav dok je w(t) aditivni šum – pretpostavimo da se radi o bijelom Gaussovom šumu (engl. *Additive White Gaussian Noise*, AWGN) spektralne gustoće $N_0/2$ i da ja pojasno ograničen na područje frekvencija od 0 do B [Hz]. Za šum vrijedi još jedna važna činjenica, a to je da svaki uzorak šuma ima Gaussovu funkciju gustoće vjerojatnosti, srednju vrijednost jednaku nula i varijancu određenu izrazom $\sigma^2 = N_0B$. Ukoliko se broj skokova povećava, tj. ukoliko je u komunikaciji uposleno više releja, što prikazuje Slika 11, prijenosna funkcija sustava (ukoliko se radi o AF mreži) je:

$$H(f) = H_{t1}(f)H_{c1}(f)H_{r1}(f)H_{t2}(f)H_{c2}(f)H_{r2}(f)\ldots H_{tM}(f)H_{cM}(f)H_{rM}(f)$$

Nadalje, ukoliko pretpostavimo da se radi o identičnim predajnicima:

$$H_{t1}(f) = H_{t2}(f) = H_{tM}(f) = H_t(f),$$

prijemnicima:

$$H_{r1}(f) = H_{r2}(f) = H_{rM}(f) = H_r(f)$$

i kanalima:

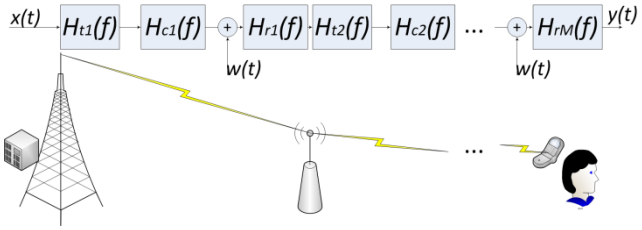
$$H_{c1}(f) = H_{c2}(f) = H_{cM}(f) = H_c(f),$$

tada gornju jednadžbu zapisujemo kao:

$$H_{AF}(f) = H_t(f)^M H_c(f)^M H_r(f)^M = (H_t(f)H_c(f)H_r(f))^M,$$

pri čemu parametar M označava broj skokova. Korištenjem Fourierove transformacije, za impulsni odziv dobivamo:

$$h_{AF}(t) = \int_{-\infty}^{\infty} H_{AF}(f) e^{j2\pi f t} df$$



Slika 11. Prikaz prijenosnih funkcija podsustava

VI. ZAKLJUČAK I DALJNJI RAD

U ovom članku analizirane su višeskočne relejne mreže (engl. *Multihop Relay Networks*). Pozornost je bila usmjerena ka usporedbi prijenosnog kapaciteta dviju osnovnih, tj. najpoznatijih prijenosnih tehnika, a to su pojačaj i proslijedi (engl. *Amplify and Forward*, AF) te dekodiraj i proslijedi (engl. *Decode and Forward,* DF). Pokazano je da, što se kapaciteta tiče, DF mreže predstavljaju bolje rješenje, jer uz isti iznos omjera signal-šum (SNR) pružaju veću spektralnu iskoristivost, tj. veći prijenosni kapacitet. No, sklopovska implementacija DF mreže daleko je složenija te samim time povlači veći financijski trošak implementacije. Uz to, DF relejne mreže unose dodatno kašnjenje te nisu u potpunosti prihvatljive za usluge kojima je kašnjenje kritičan faktor (primjerice govorni promet).

Nadalje, izvođenjem formula koje prikazuju ovisnost kapaciteta mreže o broju skokova, te grafičkim prikazom tih izraza, vidljivo je da je kod obiju relejnih tehnika prisutan fenomen „niske“ i „visoke“ razine SNR-a. Naime, ukoliko je omjer SNR mali, uz povećanje broja skokova, tj. smanjenjem međusobne udaljenosti među postajama kapacitet sustava se neće značajno promijeniti. Za razliku od toga, kod velikog omjera SNR vidljivo da je broj skokova kao i faktor prigušenja signala (koji je ovisan o okolini) značajno utječu na promjenu kapaciteta mreže.

Trenutno najperspektivniju, a time i najaktualniju temu u svijetu bežičnih komunikacija predstavlja javna pokretna mreža četvrta generacije, odnosno LTE (engl. *Long Term Evolution*). Očekuje se da će kroz koje vrijeme releji zauzeti važno mjestu unutar te tehnologije [15]. Stoga bi kao neku od budućih tema bilo zanimljivo istražiti ulogu releja u takvoj mreži, objasniti način implementacije releja, ili pak definirati protokole usmjeravanja, tj. odabira relejnih postaja koje sudjeluju u komunikaciji bazna postaja - korisnik.

Literatura

[1] W. MOHR, R. LUDER, K. H. MOHRMANN, „Data Rate Estimates, Range Calculations and Spectrum Demand for New Elements of System Beyond IMT-2000“,*5th International Symposium on Wireless Personal Multimedia Communications*, pp. 37-46, Listopad 2002.

[2] B. ZAFAR, S. GHEREKHLOO, M. HAARDT, „Analysis of multihop relaying networks“, *IEEE vehicular technology magazine*, pp. 40-47, Rujan 2012.

[3] G. HUANG, Y. WANG, J. COON, “Performance of Multihop Decode-and-Forward and Amplify-and-Forward Relay Networks with Channel Estimation”, *IEEE Pacific Rim Conference*, pp. 352-357, Kolovoz 2011

[4] E. SOLJANIN, „Hybrid ARQ In Wireless Networks“, *Mathematical Sciences Research Center*, Bell Labs, Ožujak 2003.

[5] R. NAĐ, *Planiranje mobilnih sustava*, Predavanja, Zagreb, 2011.

[6] I. S. PANDŽIĆ, Ž. ILIĆ, M. KOS, A. BAŽANT, Z. VRDOLJAK, V. SINKOVIĆ, *Uvod u teoriju informacije*, Zagreb, 2007.

[7] T. M. COVER, A. A. EL GAMAL, „Capacity Theorems for the Relay Channel“, *IEEE Transactions on Information Theory*, pp. 572-584, Rujan 1979.

[8] M. JI, Z. WANG, H. R. SADJADPOUR, J.J. GARCIA-LUNA-ACEVES, „The Capacity of Multihop Relay Wireless Networks“, *IEEE – European Wireless Conference*, 2012.

[9] M. JI, Z. WANG, H. R. SADJADPOUR, J.J. GARCIA-LUNA-ACEVES, “A unifying perspective on the capacity of wireless ad hoc networks”, *Infocom*, pp. 753–761., Travanj 2008.

[10] P. GUPTA, P.R.KUMAR, “The capacity of wireless networks”, *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, Ožujak 2000.

[11] Y. T. SAGAR, J. YANG, H.M. KWON, “Capacity of a Modulo-Sum Arbitary SISO Relay Network”, *IEEE Vehicular Technology Conference*, pp. 1-5, Svibanj 2012.

[12] S. W. PETERS, R. W. HEATH, „The Future of WiMAX: Multihop Relaying with IEEE 802.16j“, *IEEE Communications Magazine*, pp .104-111, Siječanj 2009.

[13] A. C. KURVE, *Multihop relaying in WiMAX – Cooperative Principles and Relay Routing*, Travanj 2010.

[14] G. FARHADI, N. C. BEAULIEU, “On the performance of amplify and forward cooperative links with fixed gain relays”, *IEEE Trans. Wireless Commun*., vol. 7, no. 5, pp. 1851-1856, Svibanj 2008.

[15] K. FITCHARD, “The Future of Mobile Networks: Beyond 4G”, *Bloomberg Business Week Technology*, GigaOM, Prosinac 2012.

Integration of Advanced Additional Capabilites to a Modern IPTV Platform: a Technical Study

Goran Macut, Neven Drljević
University of Zagreb Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
goran.macut@gmail.com

Abstract - This paper explores efforts to create 3rd party add-on components for a modern IPTV platform. Two scenarios are studied: creation of an IPTV component with two way communication to a 3rd party service and creation of a natural (motion) interaction interface for the IPTV platform. Extensibility offered by modern IPTV platforms is explored, as well as the challenges faced.

I. INTRODUCTION

Internet Protocol television (IPTV) is becoming a more common way of distributing television content each year, especially in the context of Triple-play services [1]. IPTV is a distribution method with backing infrastructure through which television content is delivered using the Internet Protocol (IP) over a packet-switched network (local Internet Service Provider network or the global Internet). To deliver the actual content to the television set and display the IPTV interface it usually uses a specialized device called a set-top box (STB). However, unlike traditional distribution mechanisms (over-the-air, cable, satellite), and due to its bidirectional nature and general computing advances, IPTV allows for enhanced interactivity and value-added services, if the IPTV platform supports those capabilities [2].

During the summer of 2012, the authors had the opportunity to participate in Ericsson Nikola Tesla Summer Camp [3] program. The goal of their Summer Camp project was the creation of two new prototype add-on components for Ericsson IPTV platform [4], both of which would showcase different features and challenges faced in modern development for embedded IPTV platforms. The Ericsson IPTV platform is a modern platform based on mature high-level technologies and as such an excellent starting point for the creation of advanced services.

The first component developed and studied in this paper is the EMH IPTV Viewer, an add-on that enables viewing patient data from the Ericsson Mobile Health (EMH) [5] system on the television set using the Ericsson IPTV. Due to the necessity of communicating with a server external to the IPTV ecosystem, the development of this component allowed for an opportunity to study communication methods and the challenges of their implementation. Communication from an embedded IPTV platform client to a 3rd party Internet-based service is an extremely common scenario for add-ons for an IPTV platform [6].

Work which forms the basis for this paper was done as part of the Ericsson Nikola Tesla Summer Camp program, sponsored by Ericsson.

The second developed and studied component is a standalone natural interaction (NI) interface with bindings for the Ericsson IPTV platform. NI interfaces – the detection of the motions, voice and actions of the user and meaningful interaction based on them [7], are a new and growing field in user interaction, expected to be a desired feature in set-top boxes in the near future [8].As such, these interfaces are an excellent real-world scenario to explore, especially in regard of their applicability to the study of challenges faced in integration of additional hardware add-ons to an IPTV STB in general.

Although this paper focuses on the specific components developed for the Ericsson IPTV platform, the challenges faced and the methods used have broad applicability with a wide array of modern IPTV platforms.

II. DEVELOPMENT ENVIRONMENT

The Ericsson IPTV platform is comprised of multiple middleware layers. The layer used for development of components described in this study is the end-user layer called the portal. It is a distributed system based on a server application and thin clients that IPTV users have in their homes. Thin clients (STBs) are basically low performance computing devices that run a lightweight client environment. On startup, the client environment connects to the IPTV portal server and loads a complete IPTV client interface.

The distributed nature of this solution offers substantial advantages over monolithic solutions. The main advantage is that thin clients always get the up-to-date version of the portal and additional functionality can be remotely distributed to the users. On the other hand, development is often more demanding since both server and client have to be implemented, including their communication.

The Ericsson IPTV portal is an example of a modern IPTV interface system with a very flexible and extensible architecture. The main design philosophy of such systems is extension without removing existing functionality. Every part of the system can be dynamically replaced with one that offers more functionality, but the replaced parts stay in the system.

The system is based on a common core and a set of additional modules called applications. Applications can be deployed independently of each other, each with its own functionality. The idea is that the core portal provides only minimal functionality, and that the applications bring additional functionality. All additional components

developed during this study are different applications that can be deployed at will. They don't interfere with the existing system in any way, but rather extend it.

III. Scenario: Connecting to a 3rd Party Service

EMH IPTV Viewer is an application that, when deployed to the portal, offers an additional user interface to the EMH system. Users can use the application to view data and interact with the EMH system from their home, using their television sets and the Ericsson IPTV solution.

This is an example of a very common need, a vendor application that enables two way communication with a 3rd party service, from an environment that users know and consider easy to use. EMH IPTV Viewer is additionally interesting because of its target demographics. Older users have a greater need for such health related solutions, and they still consider their television sets the main electronic device in their households and use them more frequently than, for example, personal computers or smartphones.

A. *Ericsson Mobile Health System Overview*

The EMH system consists of sensors, gateways, the EMH server system and presentation devices and applications (colloquially called screens), as can be seen in Fig. 1.

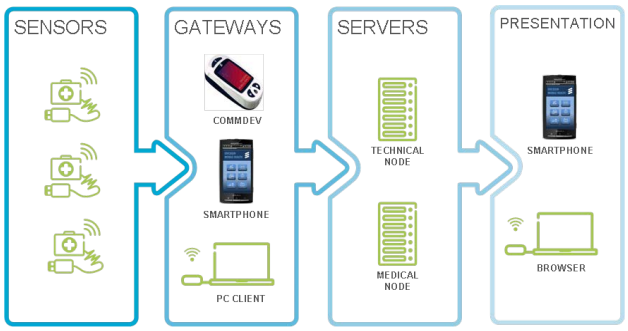


Figure 1. Ericsson Mobile Health overview

Sensors are used to take the actual measurements, gateways are physically collocated with sensors and serve to collect measurements and forward them over IP networks to the server. Servers primarily store the measurements and other patient, organization and process information. The presentation devices and applications (screens) present the stored information to the users. Currently deployed screens include a web interface and an Android mobile application. The developed EMH IPTV Viewer represents an additional screen for the EMH system.

B. *Component Features*

The EMH IPTV application is designed to be used by users with basic knowledge of technology. The application is operated using the remote control for the STB. The application interface uses symbols and language consistent with the rest of the IPTV portal, which should make the use of the application seamless. It provides access to Personal Healthcare Record stored on the EMH server system through four views: Profile, Measurements, Calendar and Questionnaires (Fig. 2).

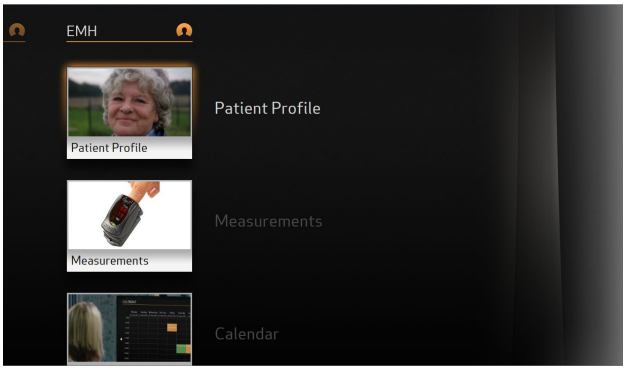


Figure 2. EMH IPTV Viewer menu

The patient profile view displays basic patient data saved on the EMH system backend. Additionally, it enables users to view recent doctor's notes in the system, as well as their previously sent questionnaire answers.

The measurements view displays measured values from the sensors the patient has been assigned. EMH currently offers a range of sensors with the measurement data being automatically sent to the EMH servers. When using the measurements view, the data is fetched live from the servers and visualized in chart format. For single shot sensors, e.g. blood pressure monitor, each chart displays multiple measurements from the selected time period. For streaming sensors, e.g. oxygenation meter, each chart displays one measurement and all data points from that measurement (Fig. 3).

The calendar view displays the patient's calendar which contains information about doctor appointments, medication dosages, required measurements and other events that should occur at specified times. Patients can browse the calendar and view events, but the system can also remind them about them. The calendar view was implemented to show the Viewer's unique advantages over other EMH screens. For example, EMH IPTV Viewer can easily remind users to take their medicine while they watch a movie. Its operation is substantially closer to the user than using the web interface and the Viewer can use that fact to enable more direct patient-EMH communication.

The questionnaire view enables users to fill out questionnaires defined by their doctor and send their answers back to the system. This feature is implemented in the EMH service and this view is one additional way of filling out the questionnaires.

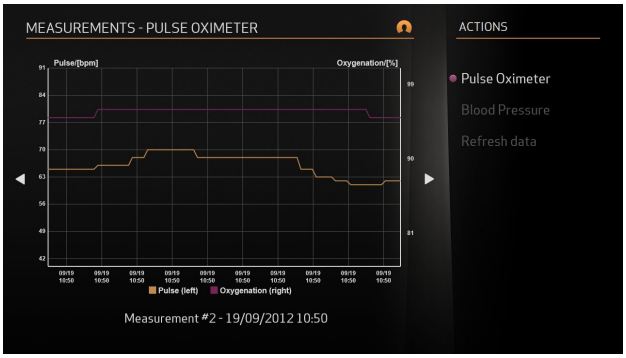


Figure 3. The measurements view showing a Pulse Oximeter measurement

C. Development Challanges

A challenge encountered early in development was a lack of available outside resources (forums, tutorials, code-sharing sites, etc.) due to the proprietary nature of the IPTV portal.

This was a new experience for the developers, as most of their training and practical experience in code development was done on open or semi-open platforms that usually have a dynamic developer community. Transitioning to a platform where, in order to succeed, it is mandatory to follow the manual and reuse existing functionality in a specifically prescribed way was a culture shock to the developers, compared to their training.

Obviously, extensive documentation and developer training is required to prepare developers for the experience of developing for an unconventional proprietary platform like the Ericsson IPTV portal.

During the development process it became apparent that direct communication from the client to the third-party service was not desirable. By using a proxy at the server level, transforms and adaptations can be applied to the data on a system that has significant computational resources, rather than on the STB with limited resources. Such an approach obviously affects the scalability of the solution.

D. Conclusions

The EMH IPTV Viewer application is exactly the type of application envisioned when the extendable STB-based approach was conceived. It brings additional value to customers enrolled in EMH programs, allowing them an easy way to view patient data in a format appropriate and familiar to them. It brings synergy to Ericsson's platforms by bridging two consumer facing platforms that are traditionally thought of in entirely separate contexts.

IV. Scenario: Creating a Natural Interaction Interface

NI is an emergent field of user-machine interaction that focuses on creating user interfaces that are based on natural user actions rather than interaction via a proxy interface device such as a keyboard, remote control etc.

The goal is to make the user interface seamless: there shouldn't be user interface components in the traditional sense but the computer should interpret intent from the user's movement, voice and similar natural actions [7].

One of the premier devices in this space is Microsoft's Kinect add-on for the Xbox 360 gaming console (and more recently, as a standalone peripheral for PCs) [8]. It can track user movements using a combination of infrared (for depth) and traditional RGB cameras. It can additionally pick up ambient sound via an array of microphones [9].

User movement is interpreted to more meaningful data using skeleton tracking of users (the creation of a skeleton model of the user in 3D space based on camera inputs), and gesture recognition based on that skeleton data [10].

More important than the technical characteristics are the user scenarios that the Kinect enables. It allows interaction using motion gestures and removes the need for a remote control when the Xbox 360 is used as a multimedia hub for the living room, which is a use case that is getting more prominence with the growth of online video content.

The Kinect is heavily promoted as such a multimedia hub and it is a direct competitor to IPTV offerings. Therefore, a NI interface is a feature that could be a significant competitive advantage or disadvantage. That being said, NI interface integration into an IPTV platform is a scenario gaining more and more relevance each day and is explored here in the context of such integration into the Ericsson IPTV portal (Fig. 4).

A. State of the Natural Interaction Technology Field

As noted, the NI field is nascent, especially the subfield of motion based interfaces and sensors required for them. The premiere device is Microsoft's Kinect based on technology developed by Primesense [11], with alternative devices slowly becoming available. The Kinect has an officially released software development kit (SDK) [12] that allows development of applications that depend on it for Windows and the Xbox 360. It does not offer support for non-Microsoft platforms or general embedded applications.

OpenNI [13] represents the industry-standard effort in that respect. It is a consortium for building open-source drivers and frameworks in the entire class of NI devices, based on Primesense technology, while offering support for commercial implementations. The OpenNI framework is modular, offering the capability of using different modules for sensor drivers and features such as skeleton detection and gesture recognition. OpenNI is open-sourced, with source and binary distributions available for x86, x86-64 and ARM platforms, in Windows and Linux flavors. Its modularity allows for usage of any device of the same class (e.g. ASUS branded devices are marketed on the OpenNI website), as well as the use of a Kinect through an open-source 3rd party driver [14].

The default provider of skeleton detection and gesture recognition for the OpenNI framework is the related NITE middleware [15]. NITE middleware is a binary-only distribution, free for commercial use with qualifying Primesense based devices (currently all devices available in the NI category). Its gesture detection consists of a limited library of predefined gestures and as such isn't appropriate for complex use cases.

Figure 4. NI interface being used to navigate the IPTV STB menu

It is not currently available on the ARM platform (or any other embedded platform), although there were public promises of support [16] as well as an early leaked binary for the platform indicating actual and continued development.

Alternative skeleton tracking options are the Microsoft Kinect SDK and the open-source community driven SkelTrack [17]. Since Microsoft actually bases Kinect on underlying Primesense technology, it is possible to use the Kinect SDK skeleton tracking binaries in conjunction with OpenNI to enable skeleton tracking in applications based on non-Microsoft development languages and frameworks. However, the Kinect SDK (including the skeleton tracking binaries) is available only for Windows with little prospect of availability on other platforms.

SkelTrack is an open-source community effort at skeleton tracking based on the reverse-engineered open-source OpenKinect drivers [18]. Due to the fully open source nature of the stack, its adaptation to any desired platform should be possible. OpenKinect is also available for the ARM platform. However, there are still significant issues involved. Due to its development history, SkelTrack currently has a large Gnome / glib dependency which could be problematic. It also has fewer features – it can track only arm movement with fewer joints then are available in OpenNI, and its skeleton tracking is poorer.

B. Component Features

The developed NI interface component is based, due to previously mentioned factors, on OpenNI drivers. The Kinect was chosen as the sensor with the greatest availability. However, the software components were engineered in such a way that replacement with another OpenNI-supported device is easily achievable.

The NI interface uses the NITE middleware for skeleton tracking and a custom gesture engine based on the state-machine concept [10]. It enables the user to control the STB with motion gestures from a custom gesture vocabulary (Table II), with support for conflicting gesture resolution. If two gestures conflict, after one is detected, the other cannot be detected for a set period of time. Support is enabled for activating and deactivating the entire gesture recognition system. When the system is deactivated, it is monitoring only for the activation gesture. This protects the user from issuing inadvertent commands.

A custom driver for the Kinect motor was developed and integrated into the component [19]. It allows for control of the vertical Kinect motor and changing of the LED light color and display pattern (Table I). On each update of the skeleton data it moves the Kinect so that the field of view is centered on the point between the shoulders of the user.

C. Component Architecture

The NI Interface is comprised of the NI Server, a custom sensor gateway that communicates with the STB using HTTP JSON messages, and the IPTV STB NI module that receives the messages and acts accordingly (Fig. 5).

TABLE I. SENSOR LED LIGHT STATUS DESCRIPTIONS

LED light	Status
blinking green	the system is currently not tracking an user
constant green	the system is tracking the user, but gesture tracking is inactive
constant orange	the system is tracking the user and gesture tracking is active

The NI Gateway is comprised of the custom NI Server code and the necessary middleware components required for operation of the NI Server. Those include OpenNI drivers and the related NITE middleware used for skeleton tracking. The NI Server is a custom Java application used to interpret the sensor information, manage the current state of the system and dispatch commands to the STB and to the Kinect sensor. It interfaces with OpenNI and NITE to receive raw skeleton data of the detected user which is passed to the integrated custom state-based gesture recognition engine. If a gesture is detected, an appropriate message can be sent to the STB.

The STB IPTV NI module can be deployed, like any other 3rd party platform module, to the IPTV server from which it is loaded to the portal. It receives the NI Server's messages, interprets them and activates the appropriate standard platform actions depending on the event.

D. Development Challanges

Since the goal was the creation of a NI interface for an embedded platform, there had to be a careful consideration of the available technologies in order to attempt integration.

Ideally, the Kinect sensor would be connected directly to the STB. However, the STB that was available had no technical possibilities to run the NI interface, being based on the MIPS processor, and running a custom build of Linux. It had no development tools installed on the system, and we found no way to cross-compile for it. Therefore, the NI Server component had to be created separately, with the intention of running on a different computer or, later, on the STB if an STB with appropriate capabilities became available.

Almost all development effort for embedded systems is targeted at the ARM platform. An ARM based STB would enable use of standard development tools to develop the required functionality. Additionally, Kinect drivers for ARM Linux exist and there are numerous web reports of successful integration on systems based on ARM processors, for example on BeagleBoard or PandaBoard [20]. The main feature required to use the drivers on an embedded platform is support for the libusb library for accessing USB devices [21].

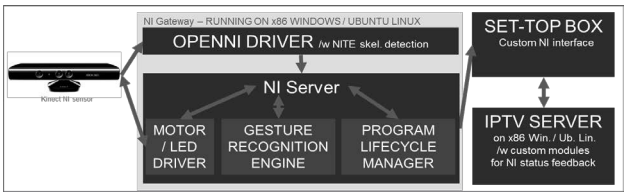


Figure 5. NI interface architecture overview

TABLE II. AVAILABLE GESTURE VOCABULARY

Gestures			
Name	***Description***	***Detected by***	***Action***
Swipe (up / down / left / right)	a swipe of the right hand in the desired direction	monitoring the position of the hand compared to the right elbow of the user; must be performed in front of the user, at approx. chest height	equivalent to pressing the appropriate directional key on the remote control
Wave (left / right)	a swipe of the right hand to the left or right and then back in the opposite direction, while the hand is above the shoulder	monitoring the position of the hand compared to the right elbow of the user; must be performed in front of the user, with the hand above the right shoulder	activation / deactivation gestures
Confirm	a tap by the right hand on the left hand	monitoring the position of the hands (the right one must be above the left one) and their distance	equivalent to pressing OK on the remote control
Back	left hand held up in a STOP sign	monitoring the position of the left hand	equivalent to pressing BACK on the remote control
Menu	right hand is held right below the left hand and then it pulls away straight down emulating a menu opening animation	monitoring the position of the monitoring the position of the hands (the right one must be below the left one) and their distance	equivalent to pressing MENU on the remote control

Unfortunately there were no such systems available to test this possibility. An emulation-based approach was tried, but ultimately failed due to problems with fast USB support in ARM emulators (primarily QEMU [22]).

E. Conclusions

Natural Interface integration is a feature with a bright future in the STB-based TV experiences domain, with the Xbox 360 with Kinect already setting expectations.

Modern STBs almost always run some flavour of Linux or BSD, allowing integration at the operating system level, if the driver support is available for the platform. Unfortunately, currently only the ARM platform is getting attention in that regard, and even there the entire OpenNI ecosystem is not yet available.

That being said, the viability of creating such an interface for the Ericsson IPTV portal was demonstrated, showing the adaptability of the software platform – actual usage using the NI UI is possible and intuitive for the user once the gesture library is learned, and that result was accomplished with basically no modification of the underlying platform.

The field of natural interaction and gesture recognition in particular is currently a field of great interest for research and development, both academic and commercial, and as such rapid advancement can be expected in models that can be applied, and also in commercial middleware for existing frameworks.

V. GENERAL CHALLENGES OF STB DEVELOPMENT

Based on the experiences of the team developing for the Ericsson IPTV portal, some general conclusions on STB development can be drawn. The system has to be adaptive on multiple fronts, from functionality to overall look and feel of the portal. Vendors that buy the system should be able to easily customize the portal to their requirements. The portal look and feel, including localization of the user interface, should be easily changed.

Additionally, the vendors should have the ability to implement custom functionality to the portal, and potentially expand it to include access to the other services they want to offer. EMH IPTV Viewer is an example of such extension.

These desired capabilities are in contrast to the whole development ecosystem for such portals. To enable such functionality, the development technology has to adapt to this specific requirements.

Loose coupling is integral to every part of the system. However, technologies that enable such architecture work outside their typical scope of operation. That complicates development, since the required workflow typically is not supported by conventional development tools. Custom development tools have to be designed and implemented. Also, developers coming to such platforms typically have no experience with the similar systems, so they can be expected to face a steep learning curve.

VI. CONCLUSIONS

This paper presents study results of the development of advanced components for Ericsson IPTV, a modern IPTV platform. The platform proved to be flexible and all functionality was successfully implemented. After the initial problems with the relatively steep learning curve, development was fast and the results were well integrated into the portal. An additional advantage of this platform is its ability to easily deploy and update portal applications of this type.

The first scenario explored is a scenario of connecting to a 3rd party service via the developed IPTV component. The EMH IPTV Viewer application studied in the scenario allows patients to interact with the EMH system using their television sets and an Ericsson IPTV STB, in a form familiar to them, merging two traditionally separate contexts: IPTV, a consumer-facing technology and the EMH system, a traditional enterprise-level professional system, while providing a uniform interface for their access.

The second scenario explored is a scenario of the integration of a NI interface into the IPTV portal. As noted, the field of natural interaction is still in the early stages, but is being rapidly developed. This presents challenges for development of NI interfaces for embedded systems. At the moment, no embedded platform is fully supported by the tools considered in this paper. However, the ARM platform should attain full support rapidly.

Additionally, integration of NI interface with the STB can in general be problematic in the even lower level, due to the need for sensor drivers, which is a feature that an IPTV platform might not support. Despite those issues, a standalone prototype NI interface component was successfully developed with bindings for the Ericsson IPTV platform.

Overall, the experience stresses the need for flexibility in an IPTV platform and the capability of advanced and unorthodox extension and modification, as well as the great potential of modern platforms that do meet those requirements.

ACKNOWLEDGMENT

The authors would like to thank their colleague Merim Džaferagić, who participated in the development of the EMH IPTV Viewer during the Summer Camp. They would also like to thank Goran Kopčak for his mentorship during the Summer Camp project and Saša Dešić for his efforts at making the Summer Camp an enjoyable and structured experience.

The initial version of this paper was honored with a 3rd place Ericsson Nikola Tesla Award for the best student paper in the field of telecommunications in the academic year 2011/2012, jointly awarded by Ericsson Nikola Tesla and the University of Zagreb Faculty of Electrical Engineering and Computing. The authors are grateful to Ericsson Nikola Tesla and the Faculty for the award and recognition.

REFERENCES

[1] J. Blackwell, "Forecast and Outlook 2011: Multiplay Services Forecast." Strategy Analytics, 2011.

[2] J. Hjelm, *Why IPTV? : interactivity, technologies, and sevices.* Chichester, West Sussex, United Kingdom: John Wiley & Sons Ltd, 2008.

[3] Ericsson Nikola Tesla, "Ericsson Nikola Tesla Summer Camp," 2012. [Online]. Available: https://www.facebook.com/entsummercamp. [Accessed: 28-Jan-2013].

[4] Ericsson, "Ericsson IPTV Solution – enabling a new kind of IPTV," 2009. [Online]. Available: http://www.ericsson.com/au/ericsson/press/facts_figures/doc/iptv.pdf. [Accessed: 28-Jan-2013].

[5] Ericsson, "Ericsson Mobile Health." [Online]. Available: http://www.ericsson.com/hr/ict_solutions/e-health/emh/index.shtml. [Accessed: 28-Jan-2013].

[6] Netflix, "Netflix Ready Devices." [Online]. Available: http://www.netflix.com/NetflixReadyDevices. [Accessed: 28-Jan-2013].

[7] A. Valli, "The Design of Natural Interaction." Microsoft Research, 2006.

[8] F. X. Shaw, "Feeling the Kinect Effect," *The Official Microsoft Blog*, 2011. [Online]. Available: http://blogs.technet.com/b/microsoft_blog/archive/2011/10/31/feeling-the-kinect-effect.aspx. [Accessed: 28-Jan-2013].

[9] Microsoft, "Kinect for Windows Sensor Components and Specifications," *MSDN*. [Online]. Available: http://msdn.microsoft.com/en-us/library/jj131033.aspx. [Accessed: 28-Jan-2013].

[10] MCS UK Solution Development, "Writing a gesture service with the Kinect for Windows SDK," *MSDN Blogs*, 2011. [Online]. Available: http://blogs.msdn.com/b/mcsuksoldev/archive/2011/08/08/writing-a-gesture-service-with-the-kinect-for-windows-sdk.aspx. [Accessed: 28-Jan-2013].

[11] M. Schramm, "Kinect: The company behind the tech explains how it works," *Joystiq*, 2010. [Online]. Available: http://www.joystiq.com/2010/06/19/kinect-how-it-works-from-the-company-behind-the-tech/. [Accessed: 28-Jan-2013].

[12] Microsoft, "Kinect Developer Downloads," 2012. [Online]. Available: http://www.microsoft.com/en-us/kinectforwindows/develop/developer-downloads.aspx. [Accessed: 04-Feb-2013].

[13] OpenNI, "Home." [Online]. Available: http://www.openni.org/. [Accessed: 28-Jan-2013].

[14] "SensorKinect." [Online]. Available: https://github.com/avin2/SensorKinect. [Accessed: 28-Jan-2013].

[15] Primesense, "Primesense natural interaction," 2011. [Online]. Available: http://www.primesense.com/nite.

[16] J. C. Hall, "The Future of OpenNI, Primesense (Makers of Kinect)," *Creative Distraction*, 2011. [Online]. Available: http://www.creativedistraction.com/news/the-future-of-openni-primesense-kinect/. [Accessed: 28-Jan-2013].

[17] J. Rocha, "SkelTrack," 2012. [Online]. Available: https://github.com/joaquimrocha/Skeltrack . [Accessed: 28-Jan-2013].

[18] OpenKinect, "Main Page," 2012. [Online]. Available: http://openkinect.org/wiki/Main_Page . [Accessed: 28-Jan-2013].

[19] A. Davison, "Kinect Chapter 6. Tilt Motor, etc," *Java Programming Techniques for Games*, 2011. [Online]. Available: http://fivedots.coe.psu.ac.th/~ad/jg/nui16/motorControl.pdf . [Accessed: 28-Jan-2013].

[20] Pansenti, "Kinect Installation on BeagleBoard-xM and PandaBoard running Ubuntu 12.04," 2012. [Online]. Available: http://www.pansenti.com/wordpress/?page_id=1772. [Accessed: 28-Jan-2013].

[21] "libusb." [Online]. Available: http://www.libusb.org/. [Accessed: 04-Feb-2013].

[22] "QEMU." [Online]. Available: http://wiki.qemu.org. [Accessed: 04-Feb-2013].

Upravljanje rehabilitacijskim okruženjem koristeći M2M tehnologiju

Gabrić Maša
Studentica, Veleučilište u Rijeci, Poslovni odjel, Specijalistički diplomski stručni studij Informatike, Rijeka, Hrvatska
masa.gabric@gmail.com
Jadranko Novak
Mentor, Nomen d.o.o. Rijeka, Hrvatska
jadranko.novak@mynomen.hr

Sažetak - U ovom radu je predstavljena ideja korištenja senzora u sklopu rehabilitacijskih centara.

U svakom rehabilitacijskom centru postoje prostorije u kojima se na dnevnoj bazi moraju provjeravati pravilna temperatura, vlažnost, osvjetljenje i buka. Ovisno o tome koji faktor prevladava, određuje se i uspješnost rehabilitacije. Postavljanjem senzora na različite dijelove tih prostorija, te periodično prikupljanje tih podataka, omogućava brži i bolji uvid u stanje prostorije.

Senzori komuniciraju na temelju M2M tehnologije, a prikupljeni podaci se šalju direktno na računalo ili mobilni uređaj osobe zadužene za prostoriju. Uređaji (poput klima uređaja) mogu biti priključeni na cijelu mrežu, kako bi se nakon učitanih podataka moglo kontrolirati željeno okruženje. Korištenjem ove tehnologije, stvara se opuštenija i kvalitetnija atmosfera u rehabilitacijskim centrima.

I. UVOD

U periodu od 3. do 7. rujna 2012. godine održana je, sedma po redu, ljetna škola „*senZations*“. Neke od glavnih tema su bile „*IoT – Internet of things*“ i „*M2M communications*“. Nakon održanih predavanja i kroz komunikaciju sa ostalim sudionicima, začela se ideja o ovom radu.

Primjena tehnologije u rehabilitaciji je relativno nova pojava. Primjenom modernih rješenja smanjuje se čovjekov udio u procesu rehabilitacije. Tada se posao fizioterapeuta, koji je obično zadužen za rehabilitaciju, svodi na nadzor i kontrolu. Primjer takve terapije bi bila elektrostimulacija atrofirane muskulature. Uz takvu terapiju, postoje još uvijek one terapije koje nemaju primjenu tehnologije. Takve terapije, kao što su grupno vježbanje i vježbe u bazenu, zahtijevaju određene elemente u prostorijama gdje se izvode. Elementi kao što su temperatura, vlažnost, osvijetljenost prostorije, te razina buke, važni su faktori koji mogu utjecati na ishod terapije. Npr. premala temperatura i prevelika buka u prostoriji sa bazenom, stvara neugodu pacijentu. Svaka terapija mora se obavljati u smirenom i pravilnom okruženju, inače je upitan ishod, što upućuje na važnost tih faktora u terapiji.

Stoga, uvođenjem senzora u te prostorije, mogu se prikupljati stalni podaci o željenim faktorima, te ukoliko je potrebno, oni se mogu odmah podešavati.

II. OPĆENITO O M2M TEHNOLOGIJAMA

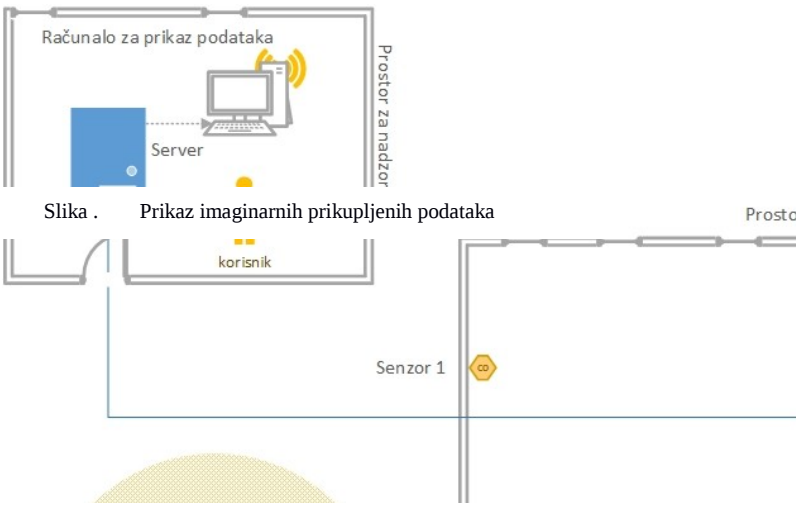
M2M (engl. *Machine- to- machine; Machine-to-mobile; Mobile-to-mobile)* koristi uređaje kao što su senzori, kako bi se pokupili podaci o određenom objektu ili pojavi. Podaci se prenose preko određene mreže (Wi-Fi, žičane ili hibridne) do aplikacije (softversko rješenje), te se pretvara u smislene podatke (npr. prikaz oscilacije temperatura pomoću grafikona).
Arhitektura ovisi o dobavljaču tehnologije.

Uređaji koji mogu biti i u pokretu, moraju se konstantno nadgledati, kako bi se detektirala eventualna nepravilnost u njihovom radu. No, M2M omogućava, uz stalno nadgledanje dobivenih podataka, upravljanje samim uređajima kod nepravilnosti u radu. Na korisnikovom sučelju se može prikazati poruka o nepravilnosti u radu, te se može pokrenuti izoliranje pojedinog uređaja s mreže ili isključenje pojedinih dijelova mreže, bez utjecaja na ostalo prikupljanje podataka.

M2M također, ima relativno malu potrošnju energije, mogućnost povezivanja velikog broja uređaja, dugi životni vijek komponenti, te na kraju veću mrežnu sigurnost. Na kraju, može se reći da M2M nije nova pojava u tehnološkom svijetu, no povećano zanimanje za te tehnologije je nastupilo radi sve većeg broja mogućnosti povezivanja.

III. PRIMJER TEHNOLOGIJE U REHABILITACIJSKIM CENTRIMA

Suprotno mišljenju da je rehabilitacija uspješna onoliko koliko i terapeut, okruženje je upravo ono koje najviše pogoduje uspješnosti terapije. Jedna od osnovnih ideja pri korištenju M2M tehnologije u svrhu kontroliranja okruženja, jest prikupljanje važnih podataka o faktorima koji ga određuju.



Slika . Prikaz imaginarnih prikupljenih podataka

Idealno bi bilo da u svakoj prostoriji bude minimalno četiri senzora (po jedan senzor za svaku stranu svijeta), koji će prikupljati željene podatke.

S obzirom na poziciju prostorije (viša pozicija – bolje osvjetljenje; niža pozicija – veća količina vlage i manje osvjetljenje), podaci mogu varirati i potrebno je prikupit više varijabla, kako bi oni bili točniji. Ti senzori bi bili spojeni na *router*, koji bi slao podatke na glavni server. Sa servera, ti podaci bi se slali na računalo osobe zadužene za prostorije, eventualno na mobilni uređaj te iste osobe. Osoblje bi pomoću aplikacija

Slika . Prikaz moguće strukture uređaja za jednu prostoriju

Uz pregled podataka, aplikacija bi pružala mogućnost pohrane podataka, uključivanje/isključivanje pojedinih uređaja u mreži, te na kraju upravljanje uređajima kao što su klima uređaji, rasvjeta i sl.

Na slici 1. se može vidjeti primjer kako bi izgledala eventualna struktura za jednu prostoriju čiji se podaci prikupljaju.

Podaci bi se prikupljali u određenom vremenskom periodu, te bi prikaz podataka bio graf ili krivulja konstantnog prikupljanja podataka. (Primjer Slika 2.)

Kao primjer je uzeto mjerenje temperature prostorije. Na slici je prikazano imaginarno razdoblje od deset sati, gdje se mjerenje podataka izvodi svakih sat vremena.

Osoba zadužena za temperaturu bi se morala pridržavati dozvoljenog minimuma i maksimuma, te težiti optimalnoj temperaturi od 24 Celzijevih stupnjeva. Može se uočiti da se osoba na početku odlučila za pojačano grijanje prostorije, smatrajući da jutarnja temperatura prostorije nije pogodna za provođenje rehabilitacije, te kad je uvidjela pogrešku, pokušala promijeniti to sa pojačanim hlađenjem koristeći klima uređaje. Nakon postignute željene temperature, istu je potrebno samo održavati. Kroz klima uređaje se može i podešavati željena vlažnost prostorije, koja je važna kao faktor ukoliko se radi o pacijentima sa plućnim bolestima.

Kao dodatak na upravljanje temperaturom u mrežu mogu se nadodati inteligentni prozori i sjenila. Određivanjem pozicije sjenila u trenu se određuje količina sunčevih zraka, a samim time se utječe na temperaturu i osvjetljenje. Uz ta dva faktora, mora se spomenuti i onaj najvažniji – količina UV zračenja.

Iako se većinom smatra štetnim, njegovo djelovanje, kao što su vazodilatacija krvnih žila, koje dovodi do brže cirkulacije krvi i limfe, relaksacije mišića, te konačno i ubrzanje cijelog metabolizma, može predstavljati uvodnu proceduru u neku drugu terapiju.

Sljedeći primjer bi bilo upravljanje rasvjetom. Kvaliteta rasvjete je obično definirana rasporedom svjetala, izborom svjetiljki, te krajnje veličina površine koja radi refleksiju. Iako je osvjetljenje obično proizvoljno, dokazano je da veća količina svjetlosti pogodno utječe na motivaciju pacijenata u izvođenju vježbi.

poželjno je ipak stalno nadgledati i utjecati na njih, radi bolje fleksibilnosti prilagodbe okruženja samim pacijentima. Na ovaj način, taj proces promatranja i održavanja željenih faktora, smanjen je na minimum. Kvaliteta okruženja je poboljšana, a korisnici zadovoljniji.

IV. ZAKLJUČAK

Stečenim znanjem iz ljetne škole „*senZations*“, otvara se novi pogled na određene tehnologije.

Iako je ovaj rad samo prijedlog ideje, iskorištavanje spomenutih tehnologija u različitim područjima kao što su rehabilitacija, otvara nove puteve za njeno razvijanje i širenje.

Stoga, ukoliko bi se ova ideja provela u djelo, ovaj pristup bi uvelike pomogao fizioterapeutima i ostalom osoblju u smanjenju vremena provedenog za stvaranje ugodnog okruženja za provođenje terapije, a javnost bi se lakše upoznala sa novim tehnologijama, kao što je M2M.

LITERATURA

[1] C.A. Haro, J.Matamoros, Parameter and Spatial Field Estimation with Wierless Sensor and M2M Capillary Networks“ , senZations summer school, 2012.

[2] P. Barnaghi, P. Navaratnam, „Middleware and Semantic Technologies for M2M communication“, CCSR University of Surrey, 2012.

[3] Dr. D. Pfisterer, „SPITFIRE: Toward a Semantic Web of Things“, University of Lubeck, Germany, 2012.
[4] B. Lazar, C. O'Reilly, D. Emmanouil, M. Gabrić, M. Vucnik, N. Gouvy, S. Nastić, „Comfice project“, senZations summer school, 2012, neobjavljen
[5] Pr. D. Simplot-Ryl, Dr. N. Mitton, „RFID and Sensor Tutorial“, senZations's summer school, 2010., neobjavljen
[6] S. Nikoletseas, „Sensor placement and energy optimization in green buildings of the future Internet“, senZations's summer school, 2010.

Sustav za podršku razvoju edukacijskih aplikacija zasnovanih na potpomognutoj komunikaciji

Šimo Brtan*, Dijana Perić*, Bruno Hemar*, David Tomić*, Hrvoje Novak*, Želimir Kompes*, Tomislav Vodopijak*, Marin Vuković*, Barbara Zagajšek**, Željka Car*

*Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, Zagreb, Hrvatska
** Hrvatska poštanska banka, d.d., Jurišićeva 4, Zagreb, Hrvatska

***Sažetak* — Potpomognuta komunikacija omogućuje dopunu ili potpunu zamjenu govora i/ili pisanja primjenom simbola, pri čemu se komunikacija ostvaruje kombiniranjem simbola, odnosno grafičkih entiteta koji predstavljaju stvari i pojave iz svakodnevnog života. Edukacijske usluge zasnovane na potpomognutoj komunikaciji mogu djelomično zamijeniti ili nadopuniti postupke podučavanja osoba sa složenim komunikacijskim potrebama. Za prevladavanje problema raznovrsnosti tehnoloških platformi za razvoj i implementaciju usluga te nehomogenosti korisnika, ove usluge trebaju biti interoperabilne i dinamički prilagođene korisnikovim mogućnostima.**

Rad opisuje sustav za podršku razvoju edukacijskih aplikacija zasnovanih na potpomognutoj komunikaciji. Cilj sustava je razvoj višestruko iskoristivih programskih komponenti za kreiranje i upravljanje edukacijskim sadržajem, nagradama i personaliziranom galerijom simbola. Primarni fokus sustava stavljen je na podatke o pojedinom korisniku, odnosno na mogućnost kreiranja korisničkih profila koji su jedinstveni za svakog korisnika edukacijske aplikacije. Kako bi se omogućio razvoj novih edukacijskih aplikacija koje bi koristile podatke o pojedinom korisniku, implementiran je sustav s dva klijenta (Web i Android) te pozadinskom komponentom predstavljenom kroz Web-usluge i bazu podataka.

I. UVOD

Komunikacija je proces razmjene poruka putem dogovorenog sustava znakova i priznata je kao jedno od temeljnih ljudskih prava. Današnjim ubrzanim razvojem informacijskih i komunikacijskih tehnologija (eng. *Information and communications technology*, ICT), komunikacija se podiže na sasvim novu razinu. Uz već standardne komunikacijske mogućnosti, današnja raznovrsna tehnološka rješenja nude brojne alate za komunikaciju, popraćene bogatom ponudom dodatnih sadržaja poput ugrađenih slikovnih, zvučnih i video predložaka. Tehnološki napredna komunikacijska rješenja često su vođena krupnim kapitalom i potrebama većine te ostaju neprilagođena za ionako marginalizirane skupine ljudi sa složenim komunikacijskim potrebama. Osobe sa složenim komunikacijskim potrebama iz različitih razloga nisu u mogućnosti ili im je otežano komunicirati putem osnovnog sredstva komunikacije, a to je govor. Neki od primjera poremećaja koji mogu uzrokovati ovakve teškoće su cerebralna paraliza, Downov sindrom i poremećaji iz autističnog spektra. Metode komunikacije koje potpomažu i/ili nadomještaju njihov govor podrazumijevaju uporabu potpomognute komunikacije.

Potpomognuta komunikacija, (PK) je hrvatska prevedenica za augmentativnu i alternativnu komunikaciju (eng. *Augmentative and Alternative Communication*) te se njome upućuje na činjenicu da je komunikacijski proces omogućen nekim alternativnim sredstvima ili pojačanjem (augmentacijom) postojećih komunikacijskih sredstava tj. sustava znakova koji se koriste pri komunikaciji [1].

Uspješno uspostavljanje komunikacije između osoba koje koriste različite komunikacijske tehnike najlakše je ostvarivo uporabom mogućnosti i resursa iz područja informacijsko-komunikacijske tehnologije (ICT). Zbog nedovoljne zastupljenosti PK u ICT uslugama i proizvodima pokrenute su određene inicijative u ICT domeni kojima se nastoji nadvladati postojeće komunikacijske barijere. Inicijative poput e-Uključenosti (eng. *e-Inclusion*) i e-Pristupačnosti (eng. *e-Accessibility*) imaju za cilj omogućiti socijalnu integraciju osoba s poteškoćama u komunikaciji te im učiniti pristupačnim obavljanje svakodnevnih vlastitih i društvenih uloga [2].

Ovaj rad je rezultat istraživanja u sklopu projekta „ICT sustavi za osobe sa složenim komunikacijskim potrebama“ (http://www.ict-aac.hr) na Sveučilištu u Zagrebu , kojim su analizirani i rješeni neki od problema u razvoju ICT usluga zasnovanih na PK. [1][2].

Rad opisuje implementaciju programskog sustava koji predstavlja osnovu za budući razvoj edukacijskih aplikacija zasnovanih na PK te integraciju unutar sustava aplikacije ABA DTT namijenjene osobama s poremećajima iz autističnog spektra.

U poglavlju II opisana je aplikacija za podučavanje ABA DTT, problem koji se rješava ovim radom opisan je i ilustriran na primjeru navedene aplikacije u poglavlju III. Arhitektura i implementacija predloženog rješenja prikazani su u poglavlju IV te su izneseni proizašli zaključci.

II. ABA DTT APLIKACIJA

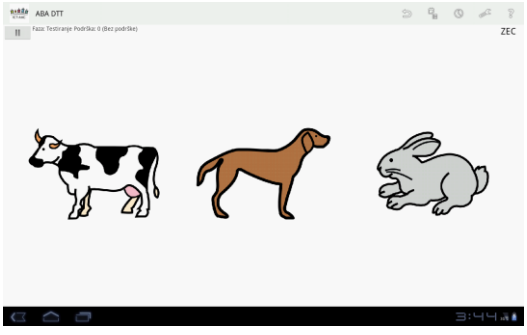
U sklopu projekta „*ICT sustavi za osobe sa složenim komunikacijskim potrebama“* na FER-u, razvijena je edukacijska aplikacija za osobe sa složenim edukacijskim potrebama pod nazivom ABA DTT (Applied Behaviour Analysis – Discrete Trial Training). Aplikacija je razvijena za uređaje na dodir (engl. *touch-screen*) s operacijskim sustavom Android i koristi

lokalnu bazu podataka Sqlite. *Applied Behaviour Analysis* (ABA) [4] jedna je od najviše istraženih i najviše korištenih metoda podučavanja osoba sa složenim komunikacijskim potrebama. ABA uključuje specifičnu i sveobuhvatnu upotrebu principa i teorija učenja po kojima se provodi proces učenja. Temelji se na psihologiji ponašanja i pomaže pri učenju jednostavnih i zahtjevnih vještina kroz sustav nagrada. ABA daje bitan naglasak individualnom pristupu učenja pojedinca, što je i preporuka znanstvenih istraživanja na području ponašanja i socijalnog razvoja osoba s posebnim potrebama. Discret Trial Training (DTT, podučavanje diskriminativnim nalozima) je jedna od tehnika podučavanja koje koristi ABA. Temelji se na učenju u malim koracima umjesto učenja cijele vještine. Svaki pokušaj izvodi se na jednak način, a to omogućava psiholozima da ustanove koje je elemente potrebno modificirati u procesu učenja za određenog pojedinca. Jedna jedinica učenja po ovoj tehnici sastoji se od tri komponente:

1. Podražaj (smjernica ili instrukcija)
2. Odgovor djeteta
3. Posljedica (povratna informacija)

Cilj razvijene aplikacije je učenje novih riječi putem simbola. Simbol u aplikaciji predstavlja edukacijski sadržaj, a u suštini to je prikaz nekog pojma odabranog za učenje.

Aplikacija tijekom sjednice podučavanja prolazi kroz ABA edukacijske faze za pojedini odabrani simbol koji se u toj sjednici podučava. Djetetu u svakom nalogu daje određenu podršku (svojevrsnu pomoć u rješavanju zadatka) te ga na taj način vodi kroz odgovor. S vremenom se ta podrška ukida. Nakon točnog odgovora djeteta slijedi nagrada. Nagrada je kombinacija slike i popratnog zvučnog zapisa, no može biti i kratka animacija ili video, a služi kao motivacija djetetu da i dalje nastavi s edukacijskom sjednicom zadržavajući jednaku razinu entuzijazma. Nagrade su izuzetno bitan faktor u procesu učenja jer djecu ohrabruju i motiviraju za daljnje učenje, stoga je nužno da nagrada bude osobno odabrana za pojedino dijete. Na slici 1 nalazi se primjer jednog naloga iz aplikacije ABA DTT u kojem dijete uči novi pojam: zec.



Slika 1. Primjer naloga (aplikacija ABA DTT)]

III. OPIS PROBLEMA

Osnovni problem razvijene edukacijske aplikacije ABA DTT je njezina vezanost za samo jedan uređaj. Nakon što korisnik instalira aplikaciju na uređaj, svi korisnički podaci, a to uključuje predefiniran skup edukacijskih kategorija, simbola i nagrada, nalaze se u lokalnoj bazi podataka na uređaju. Kod navedenog pristupa možemo uočiti problem ažuriranja, sigurnosti i mobilnosti podataka.

Korisnik je vezan uz uređaj na kojem se nalazi predefiniran skup kategorija i simbola. Simboli su objedinjeni u galerije radi lakše organizacije i pretraživanja. Galerija je entitet u sustavu koji sadrži skup simbola, odnosno edukacijskog sadržaja.

U sustavu su korištene predefinirane galerije čiji su simboli preuzeti iz već postojećih galerija simbola namijenjenih domeni potpomognute komunikacije, a koje su dostupne na svojim web-sjedištima (primjer su ARASAAC, Sclera i Mulberry galerije [1]). Sam postupak preuzimanja simbola iz dostupnih galerija s interneta ovisi o vrsti galerije. Naime postoje besplatne i komercijalne galerije. Ako se radi o besplatnoj javnoj galeriji preuzimanje se izvodi samo uz poštivanje licence i autorskih prava o korištenju sadržaja, a ako se radi o komercijalnoj galeriji potrebno je plaćanje na pravo korištenja zbirke simbola.

Ograničenje vezanosti korisnika na uređaj uzrokuje sljedeće probleme:

- nemogućnost ažuriranja predefinirane galerije simbola na uređaju,
- nemogućnost stvaranja vlastite personalizirane galerije, koja će objedinjavati simbole isključivo po preferencama korisnika te
- nemogućnost migracije korisnika s uređaja na uređaj.

Navedeni problemi za posljedicu imaju veoma ograničene mogućnosti korištenja aplikacije, kao npr. ako se na tržištu pojavi potpuno nova galerija ili samo poboljšana verzije neke već postojeće galerije korisniku aplikacije ABA DTT ti novi simboli ostaju nedostižni.

Ne postoji mogućnost prebacivanja osobnih podataka s jednog uređaja na drugi. Korisnički profil je u potpunosti vezan za uređaj na kojemu je aplikacija instalirana. U slučaju brisanja aplikacije s uređaja, korisnički profil se gubi u potpunosti. Na ovom primjeru možemo uočiti kako dolazi do značajnog gubitka vremena ukoliko se korišteni uređaj želi zamijeniti novim.

Rješenje navedenih problema je razvijanje raspodijeljenog sustava koji bi omogućio ukidanje vezanosti podataka samo za jedan uređaj te dostupnost podataka svim korisnicima neovisno o korištenoj tehnologiji. Ovo rješenje omogućilo bi korisnicima pristup željenim podacima preko Android uređaja te preko web-aplikacije putem web-preglednika. Korisnik će imati mogućnost pristupa svim javnim galerijama koje su integrirane u sustav s bilo kojeg uređaja s instaliranom aplikacijom, a dodavanje novih ili osvježenih galerija u sustav biti će omogućeno administratoru baze podataka. O samoj bazi podataka više u idućem poglavlju.

Raspodijeljeno rješenje, uz javnu galeriju koja je dostupna za pregledavanje i korištenje svim korisnicima aplikacije, uvodi i privatnu galeriju koja je namijenjena isključivo jednom korisniku koji po vlastitim

preferencijama dodaje simbole u nju te ih ažurira ili briše.

Raspodijeljeno rješenje korisnički profil pohranjuje na udaljenom poslužitelju te time eliminira problem vezanosti za uređaj. Navedenim poboljšanjima riješen je razmatrani problem aplikacije ABA DTT.

IV. OPIS RAZVIJENOG SUSTAVA

A. FUNKCIONALNOSTI SUSTAVA

Razvijeni sustav proširuje postojeću aplikaciju ABA DTT, integrira je u okviru platforme za razvoj i implementaciju usluga za PK te nudi rješenje za analizirani problem u kontekstu pohrane korisničkih podataka. Kao što je prethodnim poglavljem navedeno, trenutno stanje ABA DTT aplikacije je u kontekstu pohrane podataka orijentirano isključivo na lokalnu razinu uređaja na kojem se aplikacija izvodi. Ovim se rješenjem uvodi raspodijeljeno okruženje koje podatke pohranjuje na poslužitelj. Novo postavljeno rješenje koje donosi sustav omogućuje korisnicima da aplikaciji pristupaju s raznih uređaja, od osobnih računala do mobitela i tableta, neovisno o njihovoj tehnološkoj izvedbi, te da pritom uvijek imaju vlastito personalizirano okruženje u vidu korisničkih podataka. Rad aplikacije podrazumijeva stalnu internetsku povezivost. Aplikaciji se pristupa pomoću web-preglednika, koji podržava *JavaScript*, ili preko razvijene Android aplikacije.
Glavne funkcionalnosti koje razvijeni sustav pruža su personalizacija galerije simbola i personalizacija nagrada za svakog korisnika aplikacije. Za pristup navedenim funkcionalnostima, korisnik se mora prijaviti na sustav kako bi dobio prava pristupa.

Personalizacija simbola je ključan dio cijelog sustava. Korisniku je omogućeno da iz postojećih javnih galerija simbola (nekomercijalne galerije simbola Mulberry, ARASAAC[1], ScleraPicto's [6]) odabere simbole koji mu odgovaraju i koje će koristiti za komunikaciju i učenje. Sustav omogućuje pretraživanje dostupnih javnih galerija, prethodno uvezenih u AAC bazu podataka, tako da korisnik unosi dio naziva simbola, te odabire galeriju nad kojom će se pretraživanje izvesti. Kao rezultat pretraživanja korisniku se prikazuje lista simbola, te on ima mogućnost dodavanja željenih simbola u vlastitu personaliziranu galeriju. Osim odabira simbola iz javnih galerija, korisnik ima mogućnost stvaranja novih simbola. Pritom svaki novo stvoreni simbol mora imati naziv te sliku i/ili zvučni zapis. Personaliziranu galeriju korisnik može uređivati dodavanjem i brisanjem simbola dok samim simbolima može mijenjati naziv, sliku prikaza te prethodno pridjeljeni zvučni zapis.

Personalizacija nagrada ostvarena je istim konceptom kao i personalizacija simbola. Razlika je u tome što se nagrade moraju stvarati, dok se simboli pored mogućnosti za vlastito stvaranje mogu i dodavati iz javnih galerija. Omogućeno je pridjeljivanje više personaliziranih nagrada istome korisniku. Prilikom stvaranja nagrade korisnik odabire naziv nagrade, sliku i/ili zvučni zapis. Korisnik može pregledavati svoje nagrade te ih po potrebi brisati ili ažurirati. Omogućeno je ažuriranje naziva nagrade i/ili zvučnog zapisa.

B. ARHITEKTURA SUSTAVA

Rad aplikacije zasnovan je na arhitekturi klijent-poslužitelj. Za klijentsku stranu su razvijene:

- aplikacija za pokretni uređaj s operacijskim sustavom Android te
- web-aplikacija.

Klijenti s poslužiteljem komuniciraju koristeći uslugu tipa REST (*Representational State Transfer*), putem koje se pristupa i upravlja resursima na aplikacijskom poslužitelju. Web-aplikacija ne zahtjeva instalaciju te je dostupna kroz web-preglednik koji ima podršku za *JavaScript*. Razvoj web-aplikacije izveden je pomoću tehnologije JSP (*JavaServer pages*).

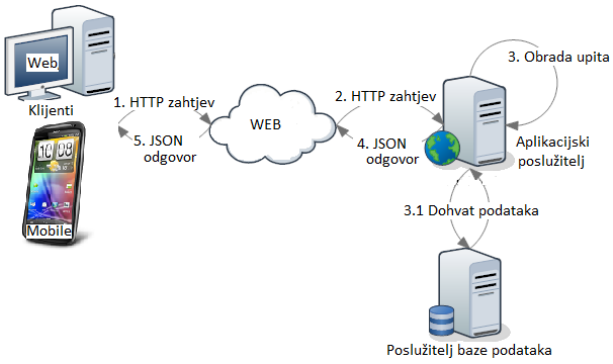
Aplikacija za pokretni uređaj s Android operacijskim sustavom ima ostvarene jednake funkcionalnosti kao i web-aplikacija te s aplikacijskim poslužiteljem komunicira koristeći REST [3].

Poslužiteljska aplikacija je razvijena u programskom jeziku Java, koristeći radne okvire *Spring* [7] i *Hibernate* [8] *te bazu podataka MySql*.

Komuniciranje klijenta i poslužitelja podijeljeno je u faze:

- Korisnik putem korisničkog sučelja internetom šalje zahtjev HTTP do poslužitelja
- Aplikacijski poslužitelj obrađuje zahtjev
- Ako je potrebno, podaci se spremaju ili dohvaćaju iz baze podataka
- Aplikacijski poslužitelj vraća rezultat obrade u formatu JSON [9] koji se prenosi natrag do klijenta

Klijentske aplikacije na temelju JSON odgovora generiraju prikaz (*View*) korisniku. JSON (engl. JavaScript Object Notation) je standardan i jednostavan format za razmjenu podataka, prvenstveno namijenjenih arhitekturi klijent-poslužitelj, jednostavan je za parsiranje i potpuno neovisan o programskoj okolini klijenta ili poslužitelja. Kao alternativno rješenje razmatrana je implementacija web-usluga tehnologijom SOAP[10] no zbog loših performansi u mobilnom okruženju je odabrana tehnologija JSON.



Slika 2. Arhitektura sustava

[1] Autor: Sergio Palao Izvor: ARASAAC (http://catedu.es/arasaac) License: CC (BY-NC-SA)

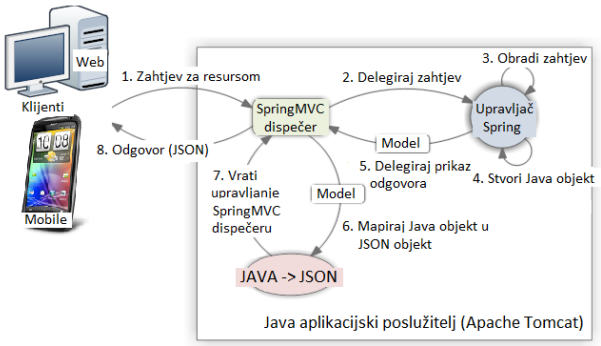
POSLUŽITELJ

Poslužiteljska aplikacija zadužena je za obradu, ažuriranje i pohranu podataka u bazu. Navedene operacije se izvode pomoću web-usluga *RESTful [11]*. Radna okolina poslužitelja sastoji se od tri glavna dijela:

1. Upraviteljski sloj (*CONTROLLER*)
2. Sloj pristupa bazi podataka (*DAO*)
3. Sloj modela (*MODEL*)

Upraviteljski sloj

Sve klijentske zahtjeve prihvaća *Springov* dispečer koji na temelju URL-a poziva odgovarajuću metodu nadležne klase upravljača (Controllera) (slika 3). Klase sadržane u sloju upravljača sadrže svu potrebnu aplikacijsku logiku sustava, obrađuju primljene podatke te ovisno o pozvanoj usluzi komuniciraju s pripadajućim metodama u sloju DAO, radi operacija s bazom podataka.



Slika 3. SpringMVC arhitektura sustava

Sloj pristupa bazi podataka

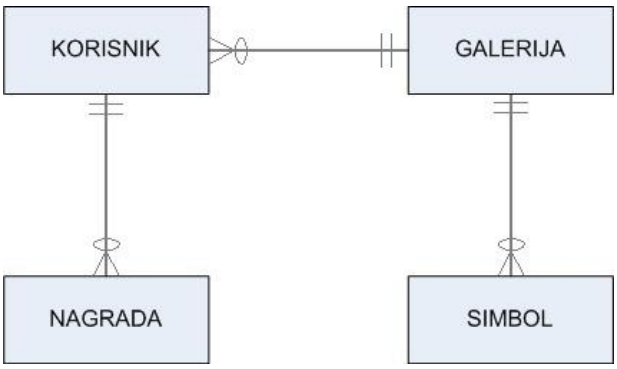
U ovome sloju odvija se sva logika pohrane, ažuriranja i brisanja podataka iz baze podataka. Hijerarhija klasa izvedena je analogno upravljačkom sloju.

Sloj modela

Sloj modela definira sve klase potrebne za stvaranje domenskih objekata u skladu s radnim okvirom *Hibernate*. Radnim okvirom *Hibernate* ostvarena je podrška automatskom prijenosu objekata između dvaju programski različitih koncepata oblikovanja; objektnog *Java* i relacijskog *MySQL*.

BAZA PODATAKA

Za trajnu pohranu podataka odabrana je relacijska baza podataka MySQLCommunity Server [12].
MySQLCommunity Server se razvija i održava unutar globalne zajednice razvijatelja i entuzijasta te je slobodan za korištenje pod uvjetima GPL licence. Baza je relacijskog tipa što znači da se sastoji od entiteta i njihovih pripadnih atributa te veza.
Kreirana baza podataka ima četiri entiteta (tablice) u koje se spremaju svi potrebni podaci (Slika 4). Entiteti su korisnici, galerije, simboli i nagrade.

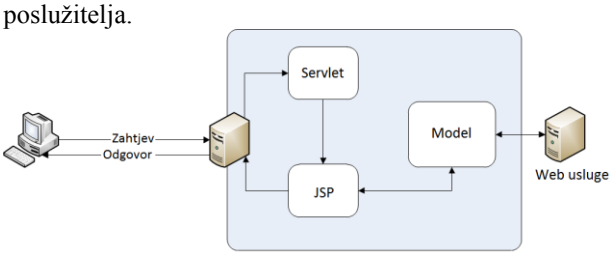


Slika 4. Relacijska shema baze podataka

Iz relacijske sheme se vidi da svaki korisnik može imati jednu personaliziranu galeriju simbola, ali može imati više personaliziranih nagrada.

WEB-KLIJENT

Web-aplikacija je pogonjena poslužiteljem Apache Tomcat. Korišteno programsko sučelje za ostvarenje web-aplikacije je Servlet API, a za prikaz sadržaja služe Java Server Pages (JSP).
Servlet je web-tehnologija bazirana na programskom jeziku Java te omogućuje stvaranje dinamičkog sadržaja koji se prikazuje na JSP stranicama. To se ostvaruje na način da Servlet dohvaća zahtjev korisnika na poslužitelj te generira odgovor koji prosljeđuje JSP stranici koja prikazuje sadržaj korisniku (Slika 5).
Java Server Page (JSP) je tehnologija koja omogućava miješanje običnog statičkog HTML-a i dinamički generiranog sadržaja iz servleta.
Arhitektura web-aplikacije je MVC (Model-View-Controller) gdje Controller-e predstavljaju Servleti, View JSP stranice, a Model obuhvaća reprezentacije objekata iz konteksta projekta te poslovni sloj za komunikaciju između aplikacije i web-usluga glavnog poslužitelja.



Slika 5. Arhitektura web-klijenta

Korisnik nakon prijave u sustav odabire funkcionalnost koju želi koristiti (administracija personalizirane galerije ili nagrada). Klijent za upravljanje personaliziranom galerijom pruža mogućnosti dodavanja simbola iz javnih galerija ili dodavanje vlastitog simbola (Slika 6.). Simboli se sastoje od imena, slike te opcionalno, zvučnog kao i *text2speech* zapisa. *Text2Speech* zapis je zvučni zapis dobiven iz javne usluge *text2speech* [13], kojom se za tekstualni unos kao rezultat vraća zvučni zapis unesenog teksta. Korisnik može pregledavati i pretraživati svoju galeriju te ažurirati i brisati simbole.

Slika 6. Pregled personalizirane galerije [5]

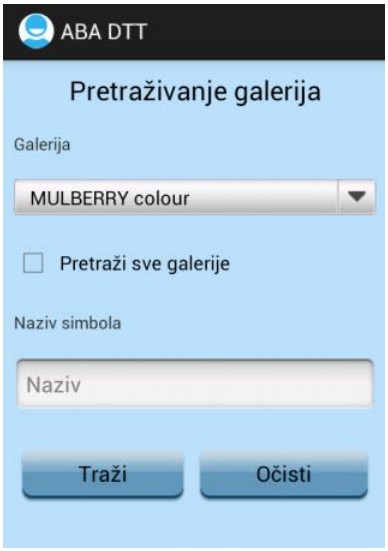
Klijent za kontrolu sustava nagrada radi na sličan način kao i za personaliziranu galeriju, s mogućnošću učitavanja nove nagrade te ažuriranje postavki i brisanje postojećih nagrada. Svaka nagrada se sastoji od imena te slike i/ili zvučnog zapisa.

ANDROID KLIJENT

Android aplikacija omogućava personalizaciju simbola za potpomognutu komunikaciju. Simboli su reprezentirani slikom i zvučnim zapisom koji je opcionalan za pojedini simbol. Simboli se nalaze u javnim i privatnim galerijama. Javne galerije su dostupne svim korisnicima, dok su privatne galerije vidljive samo vlasniku dotične galerije.

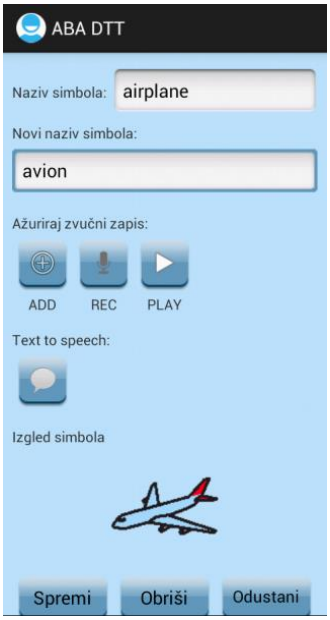
Personalizacija simbola obuhvaća izradu simbola na samom uređaju, kao i dohvat simbola iz postojećih galerija dostupnih korisniku putem klijentske aplikacije, te njihovo uređivanje u skladu s korisnikovim željama. Također je omogućeno kreiranje personaliziranih nagrada. Mogućnosti koje pruža razvijena aplikacija detaljnije su opisane u nastavku.

Kod upravljanja simbolima korisnik ima mogućnost učitavanja simbola iz javnih galerija koje se nalaze na udaljenom poslužitelju i dostupne su svim korisnicima (slika 7).



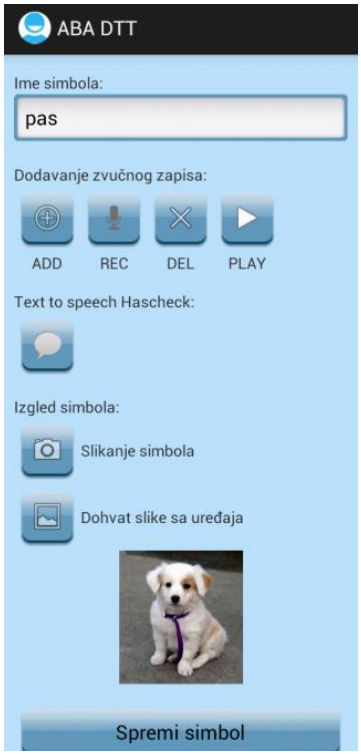
Slika 7: Pretraživanje simbola iz javnih galerija

Nakon odabira željenog simbola iz javne galerije, on se pohranjuje u privatnu galeriju i omogućena je njegova personalizacija. Personalizacija simbola obuhvaća promjenu naziva simbola, dodavanje ili ažuriranje zvučnog zapisa simbolu te brisanje simbola iz privatne galerije (slika 8). Zvučni zapis se može snimiti pomoću mikrofona na uređaju ili odabrati iz memorije uređaja. Za svaki simbol dostupna je i opcija *text2speech* koja iz tekstualnog naziva simbola generira zvučni izgovor naziva simbola. Takav personalizirani simbol pohranjuje se u privatnu galeriju korisnika koja se nalazi na poslužitelju i moguće joj je pristupiti s bilo kojeg uređaja putem izrađene aplikacije.



Slika 8: Personalizacija simbola iz privatne galerije [5]

Drugi način dodavanja simbola u privatnu galeriju je izrada vlastitog simbola na uređaju (slika 9).



Slika 9: Izrada simbola na uređaju

Slika simbola se učitava iz memorije uređaja ili se može izraditi korištenjem ugrađene kamere. Simbolu se dodaje naziv i zvučni zapis, te se pohranjuje u privatnu galeriju koja se nalazi na poslužitelju. Dodavanje zvučnog zapisa je opcionalno. [3]

V. ***ZAKLJUČAK***

Rad opisuje implementaciju programskog sustava koji predstavlja osnovu za budući razvoj edukacijskih aplikacija zasnovanih na PK te integraciju unutar sustava aplikacije ABA DTT namijenjene osobama s poremećajima iz autističnog spektra.
Problemi vezanosti korisnika aplikacije za uređaj otklonjeni su raspodijeljenom arhitekturom sustava te su omogućene nove funkcionalnosti poput stvaranja personaliziranih privatnih galerija edukacijskog sadržaja i nagrada.

Korisnik prijavom na aplikaciju s bilo kojeg uređaja na raspolaganju ima vlastite korisničke podatke kao i pristup javnim galerijama simbola. Implementirane funkcionalnosti, a posebno opcija stvaranja vlastitih simbola pruža dodatne mogućnosti i poboljšava edukacijski proces. Opisani sustav predstavlja infrastrukturni temelj budućim razvijateljima aplikacija pružajući im mogućnost da glavni naglasak u razvoju edukacijskih aplikacija bude implementacija aktivnosti podučavanje, pri čemu se koriste već implementirani komunikacijski i prezentacijski mehanizmi. Kao što je prethodno opisano, sustav nudi gotove funkcionalnosti s raspodijeljenom podrškom koje se lako mogu integrirati u bilo koje razvojno okruženje.

Ovim rješenjem pridonosi se bržem razvoju ICT aplikacija zasnovanih na PK koje omogućavaju poboljšanje vještina osoba sa složenim komunikacijskim i edukacijskim potrebama, čime im se posredno omogućuje efikasnije uključivanje u život društvene zajednice.

LITERATURA

[1] I. Lovrek, Ž. Car, M. Vuković, J. Frey-Škrinjar, M. Ljubešić, Z. Stančić, J. Ivšac Pavliša, V. Vlahović-Štetić, N. Pavlin Bernardić, J. Pibernik, L. Mandić, J. Dolić, "Katalog znanja o Potpomognutoj Komunikaciji", http://www.ict-aac.hr/index.php/katalog-znanja/sadrzaj

[2] I. Vučak, „Razvoj korisničkih sučelja za pristupačne ICT usluge zasnovane na simbolima", kvalifikacijski doktorski ispit, Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, 2012, http://www.fer.unizg.hr/_download/repository/kvalifikacijski_clanak_ivan_vucak.pdf

[3] Š. Brtan, D. Perić, D. Tomić, B. Hemar, H. Novak, T. Vodopijak, Ž. Kompes, „Sustav za podršku razvoju edukacijskih aplikacija zasnovanih na potpomognutoj komunikaciji" Diplomski projekt, Fakultet elektrotehnike i računarstva u Zagrebu, 2013

[4] J.O. Cooper, T.E. Heron, W.L. Heward, „Applied Behavior Analysis", 2nd edition, Pearson, Upper Saddle River, New Jersey, 2007.

[5] Android Developers: http://developer.android.com/reference/packages.html

[6] Sclera Pictos: http://www.sclera.be

[7] What is Object/RelationalMapping?, http://www.hibernate.org/about/orm,

[8] R Johnson, J. Hoeller, K. Donald, C. Sampaleanu, R. Harrop, Spring Framework - Reference documentation, http://static.springsource.org/spring/docs/current/spring-framework-reference/pdf/spring-framework-reference.pdf.

[9] JSON in JAVA, http://json.org/java/

[10] Programming Web Services with SOAP, http://oreilly.com/catalog/progwebsoap/chapter/ch03.html

[11] L. Richardson, S. Ruby, „RESTful Web Services: Web services for the real world", O'Reilly Media Released: May 2007

[12] MySQLCommunity Server, http://downloads.mysql.com/docs/mysql-refman-5.5-en-pdf-1-20130204.noarch.rpm

[13] Hrvatski akademski spelling checker, http://hacheck.tel.fer.hr/

Benchmarking of Parallel Ray Tracing Using Different Load Granularity

Edin Mujagic, Admir Demir, Nedim Omerbegovic, Edin Mujalo, Maja Kozar, Novica Nosovic

Faculty of Electrical Enginnering, University of Sarajevo
Department of Computer Science and Informatics

em14896@etf.unsa.ba
ad14635@etf.unsa.ba
nomerbegov1@etf.unsa.ba
em14711@etf.unsa.ba
mk15000@etf.unsa.ba
nnosovic@etf.unsa.ba

***Abstract* — This paper examines the efficiency of different load granularity in ray tracing. Benchmarking of image rendering using raytracing algorithm with different load balancing scenarios is presented. Open source Monte Carlo raytracer was modified in order to enable measurement of individual thread execution time. It is showed that, with specific load balancing scenarios, specific threads' idle time can be reduced and thus overall program execution time improved. The presented results are analyzed and future work suggested.**

***Keywords* — Ray tracing, load balancing, benchmarking**

I. INTRODUCTION

Ray tracing in computer graphics represents a technique for producing an image by tracing the path of light through pixels in an image plane and simulating the effects of its encounters with virtual objects. The technique is able to produce a very high degree of visual realism, higher than that of typical scan-line rendering methods, but at greater computational costs. This characteristic makes ray tracing best suited for applications where the image can be rendered ahead of time, such as in still images and film, and more poorly suited for real-time applications like video games where speed is crucial. Ray tracing can simulate a wide variety of optical effects, such as reflection and refraction and dispersion phenomena. Ray tracing describes a method for producing visual images constructed in 3D computer graphics environments, with more photorealism than either ray casting or scan line rendering techniques. It works by tracing a path from an imaginary eye through each pixel in a virtual screen, and calculating the colour of the object visible through it. Typically, each ray must be tested for intersection with some subset of all the objects in the scene. Once the nearest object has been identified, the algorithm will estimate the incoming light at the point of intersection, examine the material properties of the object, and combine this information to calculate the final colour of the pixel[1].

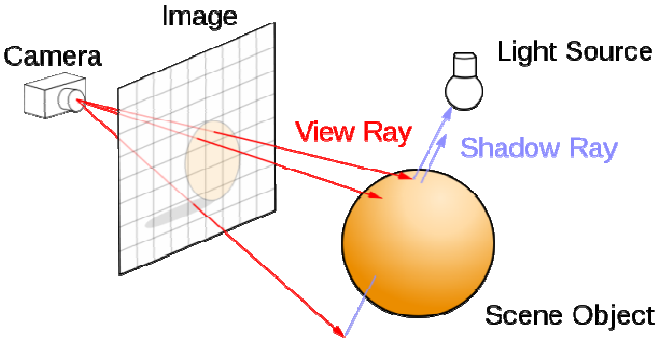


Fig. 1. Building image by extanding rays into a scene

It may at first seem non-intuitive to send rays away from the camera, rather than into it, but doing so is many orders of magnitude more efficient. Since the overwhelming majority of light rays from a given light source do not make it directly into the viewer's eye, a forward simulation could potentially waste a tremendous amount of computation on light paths that are never recorded. Oldest ray tracing algorithms cast rays from the eye into the scene until they hit an object, further ray tracing was not in the plan. Turner Witted continued process.

When a ray hits a surface, it could generate up to three new types of rays: reflection, refraction, and shadow. A reflected ray continues on in the reflection direction from a shiny surface. It is intersected with objects in the scene. The nearest object it intersects is object that will be seen in the reflection. To avoid tracing all rays in a scene, a shadow ray is used to test if a surface is visible to a light. A ray hits a surface at some point. If the surface at this point faces a light, a ray is traced between this intersection point and the light. If any opaque object is found in between the surface and the light, the surface is in shadow and so the light does not contribute to its shade. This new layer of ray calculation added more realism to ray traced images. Basic ray tracing advantage over other similar methods is that reflection and shadows, which are difficult to simulate using other algorithms, are natural products of the ray tracing algorithm. It is relatively easy to implement and it gives impressive visual results, and represents first foray into graphics

programming. Each ray can be computed independently, which makes ray tracing suitable for parallelization[1]. Shooting one ray through pixel gives us poor results so, often more than one ray is shot through each pixel. Number of rays shot through each pixel is called sample number. Combining these multiple samples to color of that one pixel is called sampling. There are multiple sampling techniques.

Finding ray object intersection is very expensive, viewing from computation aspect. Without acceleration, each ray has to be tested with all objects and then smallest value is obtained as the closest intersection. For a scene containing O number of objects and a resulting image composed of I pixels, the complexity is IxO. Acceleration techniques that are in use include methods like: fast intersections, fewer rays and generalised rays.

II. PREVIOUS WORK

Ray tracing algorithm is trivial for parallelization. Research on ray tracing parallelization exists over a decade. Reinhard and Jansen presented [6] ray tracer with data parallel and demand driven (hybrid) scheduling approach. Johansson et. al. presented parallel ray tracing system using message passing implemented using OpenMPI with master-workers model where master node in star shaped layout distributes work to worker nodes[7]. Simple load balancing strategy first-come first-served was used.

In [5] one is introduced with combination of image space subdivision with distributed object database using modified freeware ray tracer POV-Ray. Load balancing algorithm used tries to balance between two extremes (minimal and maximum task granularity). Ways of managing distributed object database are presented, which made possible rendering of large scenes that cannot fit in memory of one worker. Results showed near-linear speedup and high cache hit ratio.

Simple parallel ray tracer is presented by Ashraful et. al. [3] which has implementation for both MPI and OpenMP and, as claimed, achieved near optimal efficiency using simple static load-balancing technique. One is presented with load distribution for both OpenMP and MPI implementations as well memory and time complexity calculations.

III. CURRENT LOAD BALANCING TECHNIQUES

Rendering time of each pixel is influenced by a multitude of factors, such as shading algorithms, texture accesses, acceleration data structure queries. Uneven load balancing approach consists in subdividing the image in equally sized tiles, then using a scheduling load balancing strategy. Current techniques mostly used for rendering are Image space subdivision and Object space subdivision[3][5].

A. Image space subdivision

In this technique every working process gets a chunk of image to process. Processes are designed to work independent of one another.

Disadvantages of the technique is every single working unit must store process. Processes a whole scene in own memory and might cause a problem during a time of rendering a huge and complex scenes.

Advantage of the technique is it's easy to implement. Renderer used for this paper uses this technique.

B. Object space subdivision

It is a process of dividing a 3d scene into distributed regions, and regions are located in a memory of the working units. The techniques allows rendering the larger scene and the limit of memory is a sum of each memory units. Difficulties may occur in the diving of processes and increased communication between units. This techniques is harder to implement.

This paper presents benchmarking of load granularity with image space subdivision technique and parallelization of algorithm using OpenMP.

IV. SCENE SETUP FOR MESURMENTS

Small monte carlo based raytracer is used for measurement purposes[2]. It is modified so execution time of specific thread can be measured. Ray tracer is simple, and supports reflection, refraction and diffusion. The goal was to take advantage of all available processing resourses, so algorithm's code is parallelized using OpenMP for multi-processor CPU-s. Sampling technique of this ray tracer is 2x2 subpixel box sampling.

To achieve all core resources renderer uses load balancing method, which includes that each task was given to one thread. For each measurement threads are given different size of image to process.

Intention is to show how different load granularity scenarios can influence thread's and program's total execution time, so algorithm execution measurements are conducted using single machine (4-core i7 2.1GHz processor) and thus results are relative, meaning, execution times would be lower using more number of more powerful cores.

Primary goal of measurement phase was to show in practise that execution time of programming code, who is adjusted for running of multicores, is different for different load granularity scenarios. Image that we render has 4 spheres in lower third. So thread that renders bottom third has some of its primary rays hitting spheres first.

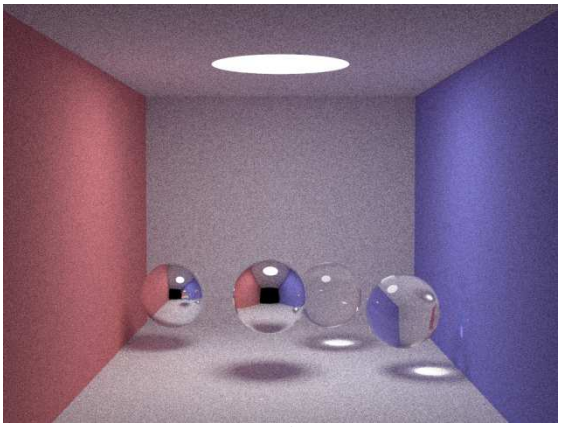


Fig. 2. Scene used for measurements

Figure 2. shows scene used in measurements. It is typical cornell box with four spheres. Two sphere have reflectivity of 1. Two have some transparency. This is aimed to present all features of raytracing (reflectivity, refractivity and shadowing).

Modification of open source ray tracing code included setting up number of threads for execution (that is 4 for 4-core measurement computer CPU) with OpenMP function *omp_set_num_threads()*.

V. MEASUREMENTS

Measurements were conducted using image space subdivision. For first measurement one thread gets whole image region while other threads get none. For first measuremnt that is equal to single-threaded execution. For every other measurement, image region assigned for execution to threads is divided by 2. Image regions are horizontal, meaning each region has full width of an image and specific height.

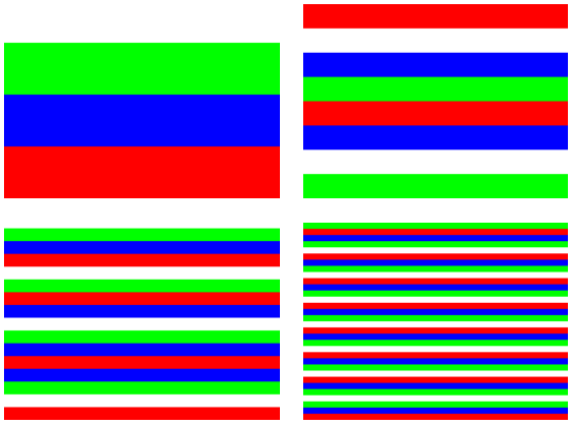


Fig. 3. Division of image regions for each thread for different measurements

Figure 3. shows different measurements' image region divisions. On upper left of figure 3, image is subdivided to 4 regions and each region is assigned to different thread for execution. Figure shows each thread's region with different colour for presentation purposes of these scenarios (one of colours being white). On three other parts of figure it is displayed greater load granularity by subdividing image regions by 2.

Because there are only 4 threads, when each thread finishes its task it will get another part of an image to process so it is necessary to say that whole image is not rendered in parallel but only these 4 regions, each being assigned to different thread so when one thread finishes its task it will get region to render (if there are any regions left).

One measurement is conducted using 32 and another 64 samples. First measurement is equal to single-threaded execution because whole image is assigned to first thread while other threads get none. Second measure is equal to two-threaded execution.

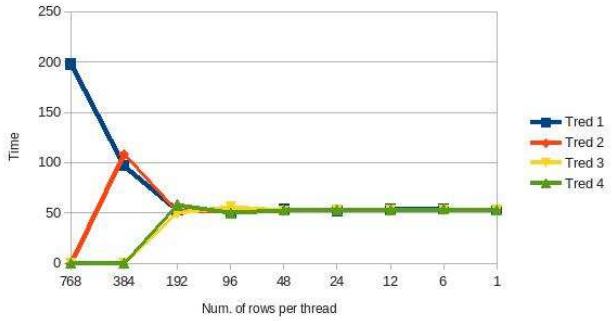


Fig. 4. Graph showing times of execution per thread for differenti image region sizes.

Figure 4. is a diagram presenting first series of measurements using 32 samples on an image shown in figure 2 width height of 768 pixels. Number of pixels of each region's height is shown on x-axis. Y-axis shows time of execution needed for each of 4 threads. Threads themselves are shown as separate lines. Maximum execution time of all threads is time needed for program to complete.

First measurement is equal to single-threaded execution and program takes most time to complete. Second measurement uses two threads because image region used for first measurement is cut in half, and two threads are employed. Total program's execution time is almost half the first case.

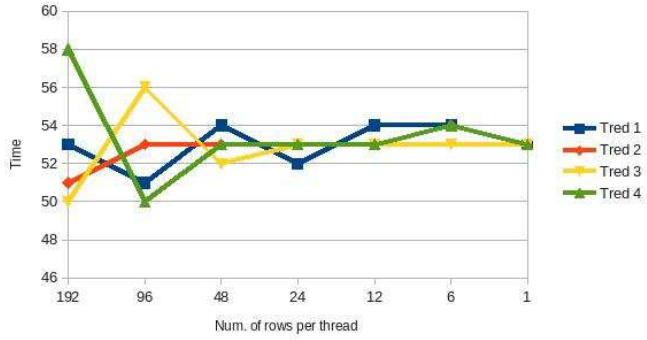


Fig. 5. Figure 4 diagram excluding first two measurements showing more clearly threads' execution times.

For other cases all threads are used. Detailed presentation of measurements' results are shown on figure 5.

As image regions are smaller, and load granularity is higher, it can be observed that execution times of threads begin to converge. Program's execution time is maximum of execution times of all threads. So when image region is 192 pixels in height time needed for program to complete is 58 s, which is time of longest running thread. Other threads took less time to complete.

When load granularity is maximum, meaning image region assigned to thread has height of 1 pixel each thread needed almost as same time to execute and overall program's execution time is lower than in previously discussed case.

Second series of measurements was conducted using 64 samples. Doubling of sample number should result in doubling execution times of previously conducted measurements.

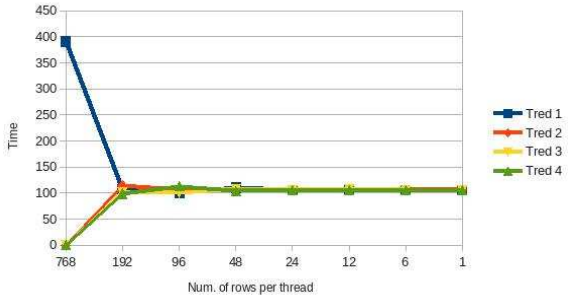


Fig. 6. Same measurements conducted using 64 samples

Measurements showed similiar behaviour as when lower number of samples is used. Figure 6 ommits displaying measurement when image is subdivided by 2. First measurement (equivalent of single-thread execution) is showed for reference.

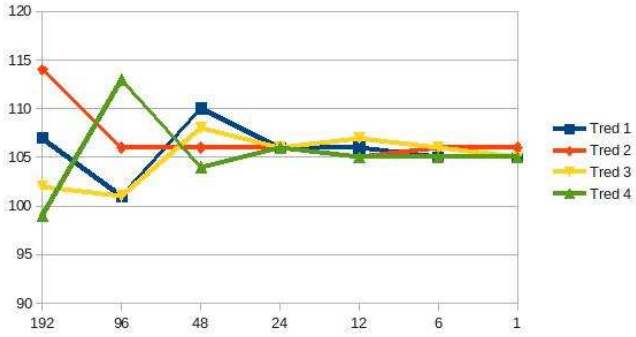


Fig. 7. More detailed figure 6

In figure 7 similar results are obtained. While maximum granularity didn't yield same execution times of all threads those times are less different than ones in 192 pixel region height and total program's execution time is almost 10 seconds lower.

VI. CONCLUSION AND FUTURE WORK

This paper presented measurements for different load granularity scenarios and showed that with strategic employment of different threads one use resources more efficiently.

This differs with scenes and object placement. Ray tracer used for measurements is Monte Carlo based ray tracer that has secondary rays randomly directed after bounce. This means there is no guarantee secondary ray will not bounce off to an area of a scene which is complicated for calculation, and bounce again into same area.

It would be interesting for future work to have this load granularity method measured for ray tracer with some indirect illumination technique implemented and measure different cases for primary and secondary rays. Threads that would get area of an image that contains complicated objects would have more intensive computation for primary rays. Primary rays would be combined with photons from photon mapping to give indirect illumination illusion.

REFERENCES

[1] Peter Shirley, Michael Ashikhmin, Steve Marschner, "Fundamentals of computer graphics", A K Peters, 2009;

[2] Kevin Beason, Small Global illumination raytracer http://www.kevinbeason.com/smallpt/ - last access date: Januar, 2013

[3] S. M. Ashraful Kadir and Tazrian Khan, "Parallel Ray Tracing using MPI and OpenMP", Royal Institute of Technology, Stockholm

[4] Henrik Wann Jensen, "Global Illumination using Photon Maps", Proceedings of the Eurographics Workshop in Porto, Portugal, June 17–19, 1996

[5] Tomas Plachetka, "Perfect Load Balancing for Demand-Driven Parallel Ray Tracing", Euro-Par 2002 Parallel Processing 8th International Euro-Par Conference Paderborn, Germany, August 27–30, 2002 Proceedings pp. 410-419, 2002

[6] Erik Reinhard and Frederik W. Jansen, „Rendering large scenes using parallel ray tracing", Volume 23, Issue 7, Elsevier Science Publishers B. V., 1997

[7] Johansson Gunnar, Nilsson Ola, Sderstrm Andreas and Museth Ken, „Distributed Ray Tracing In An Open Source Environment", Proceedings of the SIGRAD 2006 Conference (2006), Linkping University Electronic Press, 2006

A HTTP-Enabled Cryptographically Secure Hardware Random Number Generator

Mislav Božičević* and Aleksander Radovan **
* Polytechnic of Zagreb, Zagreb, Croatia, mb@mail.mislav-bozicevic.iz.hr
** University of Applied Sciences Velika Gorica, Velika Gorica, Croatia, aleksander.radovan@vvg.hr

Abstract - A cryptographically secure hardware HTTP enabled random number generator based on a nondeterministic entropy source has been implemented using open source hardware development kit. Internally, randomness is distilled from the entropy source using Algorithm M where modified-RC4-based and BLAKE-512-as-counter-mode pseudo-random number stream generators are combined. Users access the device via HTTP on a TCP/IP network and platform specific device drivers are not required. The random number files were tested using stringent statistical tests including DIEHARD, STS 2.1.1, dieharder 3.31.1, TestU01 1.2.3 SmallCrush and ENT. Emphasis was put on low cost of hardware.

I. INTRODUCTION

From the very beginnings of the computer systems, random numbers played an important role. John Von Neumann, who is regarded as the father of computer science, strongly cautioned people against the use of any form of software algorithm to generate random numbers [7]. Von Neumann recognized that only a physical process can produce a truly random, unpredictable number. The output of a mathematical algorithm, by its very nature, can be predicted if the algorithm is returned to the initial condition of previous time.

A good random number generator should have several characteristics in order to satisfy strict mathematical requirements of randomness:

- It should produce numbers that pass major known criteria for randomness, like statistical tests including DIEHARD, STS 2.1.1., dieharder 3.31.1, TestU01 1.2.3 SmallCrush and ENT.
- It should not have repetitive pattern in its output.
- It should have no initial conditions that reset the output to repeat a sequence.
- It should not be spoofable.
- It should not be susceptible to commandeering, in which the generator is forced to produce a predictable stream.

This work was supported by the University of Applied Sciences Velika Gorica and the Polytechnic of Zagreb.

Unlike pseudo-random number generators, which can be placed in a predictable state by introducing a "seed" number into the algorithm, hardware-based random number generator represents a powerful alternative.

Hardware random number generator generates random numbers from a physical process, rather than a computer program. It consists of a device that converts physical phenomena to a signal that is converted to the output digital number. Series of such values can be obtained by repeatedly sampling the input signal.

In order to be maximally utilized in a network environment, hardware random number generator must be able to connect to different devices via USB or RJ45 interface. This example shows an implementation of hardware number generator using nondeterministic entropy source like zener diode or avalanche photodiode, LeafLabs Maple (Rev5) board with microcontroller unit and network module Wiznet WIZ820io. Collected random number data stored in files were tested using stringent statistical tests that proved the truly randomness of the numbers.

II. RANDOM NUMBER GENERATION

A. Block diagram

Microcontroller unit samples a nondeterministic entropy source using an analog-to-digital converter. Samples are stored internally at the MCU in the entropy pool which is cryptographically transformed in an irreversible manner. Cryptographically processed, the numbers are available to the user via HTTP over TCP/IP.

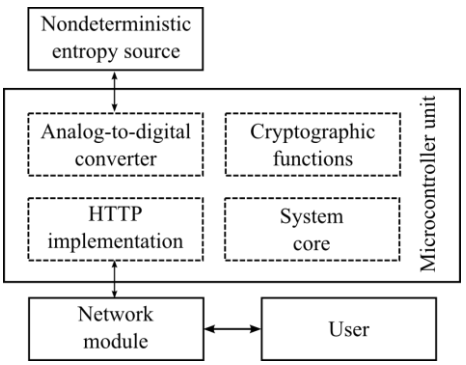


Figure 1. Block diagram that shows connections between the nondeterministic entropy source, the microcontroller unit and the network module

The nondeterministic entropy source is connected to the analog-to-digital converter pin of the MCU. The network module is connected via SPI with the MCU. The user connects to the device via HTTP over a TCP/IP network (as shown in Figure 1).

B. Microcontroller unit

Maple (Rev5), open source microcontroller board, was the basis of the hardware number generator. The board is based on STM32F103RB microcontroller unit ARM Cortex-M3 32-bit RISC core operating at a 72 MHz frequency capable of handling the input, transformation and output tasks. It is equipped with 2 x 12-bit A/D converters, which were used for processing the analog data collected from nondeterministic entropy source.

The board offers custom programming capabilities through uploading new code onto it using USB interface, which is directly supported by the Cortex-M3 MCU. This option was used for customizing the data processing and communication with the Network module WIZ820io.

Our HTTP implementation includes the GET and HEAD methods according to the HTTP/1.1 specification [14]. Individual resources available to the client are fixed length sequences of cryptographically processed numbers. Client may use the GET method to obtain an index of available resources by requesting the root directory from the server, e.g., http://192.168.0.2/. When an individual resource is requested, its URI (for example http://192.168.0.2/1234BEEF) is removed from the index and another URI is put in its place (for example http://192.168.0.2/1234BEF0). Number of URIs in the index, and therefore the number of parallel users (client connections), is limited by MCU memory and network module capability. HEAD method is provided for the client to check if there were changes on the index. HTTP was chosen for its widespread use, but simple modifications in the source code would allow for custom protocol, e.g., UDP-based.

C. Network module

WIZ820io is an internet offload network module that includes W5200, MAG-JACK (RJ45 with transformer) and other glue logic. W5200 is at the core of WIZ820io as a hard-wired TCP/IP chip which includes an Ethernet PHY. The W5200 is TCP/IP embedded Ethernet controller that enables internet connection for embedded systems using SPI. It offers eight sockets to programmer to be used for a web browser, Telnet server or protocols like TCP/IP or UDP.

Connection between Maple board and WIZ820io module was set up by using WizEthernet library and by connecting pins “RESET”, “LOW POWER MODE”, “INTERRUPT” and SPI bus.

D. Nondeterministic entropy source

Hardware random number generators employ a physical process to produce random numbers. The specific physical processes employed to generate random numbers in many cases include thermal noise or quantum processes.

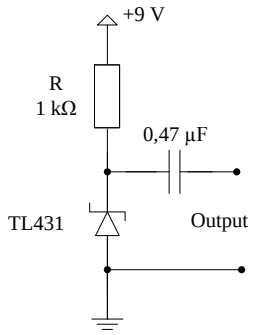


Figure 2. Zener diode based thermal noise source circuit

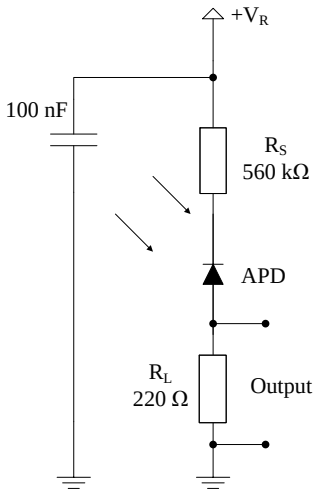


Figure 3. Simple passive quenching circuit for avalanche photodiode SAP500S2

Thermal noise arises from the agitation of current carriers due to environment temperature effects and is independent of external voltage applied to the device. Thermal noise is an example of white noise and does not have its origins in a quantum mechanical process, but rather in stochastic processes governed by the laws of thermodynamics.

One of the implementation of thermal noise is avalanche noise in which the breakdown noise generated by a zener diode operating in reverse bias mode is measured (circuit show in Figure 2). Zener diode was one of the nondeterministic entropy source that was used in the described random number generator.

Other nondeterministic entropy source that can be used in random number generator is avalanche photodiode (APD) that works in the “Geiger” mode to allow detection of single photons. One of these avalanche photodiodes is Laser Components SAP500S2. To allow detection of single photons, the SAP500S2 can be biased above the breakdown voltage using the passive quenching circuit (as shown in Figure 3).

TABLE 1. SELECTED DIEHARDER TEST RESULTS

test_name	tsamples	p-value
diehard_birthdays	100	0.84098386
diehard_operm5	1000000	0.45785809
diehard_rank_32x32	40000	0.12351172
diehard_rank_6x8	100000	0.99121283
diehard_bitstream	2097152	0.31022947
diehard_opso	2097152	0.06986824
diehard_oqso	2097152	0.58840360
diehard_dna	2097152	0.02095990
diehard_count_1s_str	256000	0.75738672
diehard_count_1s_byt	256000	0.42087518
diehard_parking_lot	12000	0.28849924
diehard_2dsphere	8000	0.98750858
diehard_3dsphere	4000	0.04579137
diehard_squeeze	100000	0.90795986
diehard_sums	100	0.16653095
diehard_runs	100000	0.96566529
diehard_craps	200000	0.15539040
marsaglia_tsang_gcd	10000000	0.33710278
sts_monobit	100000	0.93145119
sts_runs	100000	0.94482231
rgb_kstest_test	10000	0.94587621
dab_bytedistrib	51200000	0.85997660
dab_dct	50000	0.30514859
rgb_permutations	100000	0.00407979

III. CRYPTOGRAPHIC DISTILLATION

Our cryptographic building blocks are modified RC4 [6] stream cipher and BLAKE-512 [1] hash function. Only modification on RC4 is that it skips its first 3328 PRGA outputs. That number of skipped outputs is more conservative than the most conservative recommendation of $12 \cdot 256$ skipped outputs [3, p.14].

A PRNG is constructed from BLAKE-512 [1, p.27]. RC4 PRGA is already a PRNG seeded by its KSA.

Algorithm M (Randomizing by shuffling) [4, p.33] is used to combine the two mentioned PRNGs.

Our setup consists of two memory buffers. One is an analog-to-digital converters buffer where signal from the nondeterministic entropy source is digitized and stored. The other is a randomness pool where the hash value (computed using BLAKE-512 algorithm) from the concatenation of the previous randomness pool contents and the entire analog-to-digital converters buffer is stored.

When the analog-to-digital converters buffer is exhausted, it is replenished by new measurements from the nondeterministic entropy source and the randomness pool is updated as previously described.

Random bytes are extracted from the randomness pool by using BLAKE-512 as a counter-mode pseudo-random byte-stream generator similarly to [5, p.427].

Random bytes extracted in that fashion are used as keys for the modified RC4 and to fill the delay buffer for Algorithm M. Algorithm M uses RC4 PRGA to derive an index to the delay buffer. Value from the delay buffer (selected by the index) is served to the user. After that value is used, it is updated by XOR operation with the next sample from the analog-to-digital converters buffer.

IV. STATISTICAL ANALYSIS

Currently, there is no general definition of randomness [9] and no finite number of randomness tests can prove that a sequence of numbers is random; they can only prove that the sequence is not random. Often, a sequence (of bits) is considered random if and only if it is shorter than any computer program that can produce that string.

A. *ENT*

To check the randomness of our cryptographically processed sample, the widely used program ENT was applied to calculate entropy, Chi square distribution, arithmetic mean, Monte Carlo estimation of Pi and serial correlation coefficient. The test results for a typical sequence of 1574397 bytes are as follows:

Entropy = 7.999727 bits per byte.

Optimum compression would reduce the size of this 1574397 byte file by 0 percent.

Chi square distribution for 1574397 samples is 595.11, and randomly would exceed this value less than 0.01 percent of times.

Arithmetic mean value of data bytes is 127.5446 (127.5 = random).

Monte Carlo value for Pi is 3.136277196 (error 0.17 percent).

Serial correlation coefficient is -0.000708 (totally uncorrelated = 0.0).

B. *Dieharder*

The sample passed all tests in the dieharder battery, except one which was marked as a weak pass (rgb_permutations) as summarized in table 1. Dieharder is a random number generator testing suite [12].

C. TestU01 SmallCrush

TestU01 1.2.3 SmallCrush battery of tests was applied and the results are summarized in table 2. TestU01 is a software library implemented in ANSI C, and offering a collection of utilities for empirical statistical testing of uniform random number generators [10].

D. STS

The sample passed 183 of 187 tests in the STS 2.1.1 battery of tests [13]. Among those failed the lowest grade (as proportion) was 87/99 in the ApproximateEntropy test. The other three failed tests (all are NonOverlappingTemplate tests) share the same grade (94/99).

STS 2.1.1 was carried out with the default parameters and the sample was interpreted as 99 sequences 15903 bytes each.

E. DIEHARD

DIEHARD battery of tests was carried out and compared with dieharder results. The results match as expected.

Most of the tests in DIEHARD return a uniform value on [0,1), which is named "p-value", if the input data contains fully independent random numbers. Those values are calculated using assumed distribution [11].

V. Practical Applications and Uses

Hardware random number generators can be used for generating security keys for SSL protocols, generating initialization vectors for a symmetric algorithm for encryption of files, generating cryptographic primitives like *nonce* (number used once) in the Kerberos authentication network protocol, generating asymmetric keys used in PKI, etc.

VI. Conclusion

Results from the empirical type of statistical tests here presented gives strong support to the expectation that our hardware random number generator could be used in a real-world scenario. Cost of hardware was lowered by selecting modular components which are open source hardware and no platform dependency was introduced by using HTTP over the widely available TCP/IP.

The security of the cryptographic transform was built on the thoroughly studied RC4 and BLAKE-512. On a carefully administrated internal network our device can serve several simultaneous users.

There is space for improvement in customization of statistical tests to better suit the cryptographic construct parameters.

TABLE 2. SUMMARY RESULTS OF SMALLCRUSH

test_name	pvalue
Collision	< 1.0e-300
Gap	< 1.0e-300
SimpPoker	< 1.0e-300
CouponCollector	< 1.0e-300
MaxOft	< 1.0e-300
MaxOft AD	1 – (<1.0e-15)
WeightDistrib	< 1.0e-300
HammingIndep	< 1.0e-300
RandomWalk1 H	< 1.0e-300
RandomWalk1 M	< 1.0e-300
RandomWalk1 J	< 1.0e-300
RandomWalk1 R	< 1.0e-300
RandomWalk1 C	< 1.0e-300
MatrixRank	0.86
BirthdaySpacings	0.61

TABLE 3. STS RESULTS

Statistical test	Proportion	p-value
Frequency	99/99	0.246470
BlockFrequency	97/99	0.458035
CumulativeSums	99/99	0.637119
CumulativeSums	98/99	0.341565
Runs	99/99	0.613995
LongestRun	96/99	0.012650
Rank	98/99	0.000004
FFT	98/99	0.834308
Universal	99/99	0.000000
ApproximateEntropy	87/99	0.000000
Serial	98/99	0.834308
Serial	99/99	0.660243
LinearComplexity	96/99	0.522989

List of Abbreviations

APD	…	Avalanche photodiode
HTTP	…	Hypertext Transfer Protocol
IP	…	Internet Protocol
KSA	…	Key-scheduling algorithm
MCU	…	Microcontroller Unit
OSI	…	Open Systems Interconnection
PHY	…	Physical Layer of the OSI

		Model
PKI	…	Public Key Infrastructure
PRGA	…	Pseudo-random generation algorithm
PRNG	…	Pseudorandom number generator
SPI	…	Serial Peripheral Interface Bus
SSL	…	Secure Socket Layer
TCP	…	Transmission Control Protocol
UDP	…	User Datagram Protocol
URI	…	Uniform resource identifier

ACKNOWLEDGEMENTS

The authors are grateful to University of Applied Sciences Velika Gorica for financial support for avalanche photodiodes SAP500S2, to Polytechnic of Zagreb for financial support and to Professor M. Stipčević from the Ruđer Bošković Institute for helpful advices.

REFERENCES

[1] J.-P. Aumasson, L. Henzen, W. Meier, C.-W. R. Phan, "SHA-3 Proposal BLAKE version 1.3", available at https://131002.net/blake/

[2] A. Dujella, M. Maretić, "Kriptografija", Element Zagreb, 2007.

[3] I. Mironov, "(Not So) Random Shuffles of RC4", in Cryptology ePrint Archive, Report 2002/067, http://eprint.iacr.org/, 2002.

[4] E. D. Knuth, "The Art of Computer Programming: Volume 2 / Seminumerical algorithms Third Edition", Addison-Wesley, 1997.

[5] B. Schneier, "Applied Cryptography Second Edition : protocols, algorithms, and source code in C", John Wiley and Sons, Inc, 1996.

[6] Cypherpunks mailing list (1994-09-09), "Thank you Bob Anderson", retrieved from http://web.archive.org/web/20080404222417/http://cypherpunks.venona.com/date/1994/09/msg00304.html

[7] R. Dube, Hardware-Based Computer Security Techniques to Defeat Hackers, From Biometrics to Quantum Cryptography, John Wiley and Sons, Inc., 2008.

[8] M. Stipčević, Quantum random number generators and their use in cryptography, Proceedings of 34nd International Convention MIPRO,2011.

[9] H.-Q. Ma et al., "A Random Number Generator Based on Quantum Entagled Photon Pairs", Chin. Phys. Lett., Vol. 21, No. 10, 2004

[10] P. L'Ecuyer, R. Simard, "TestU01: A C Library for Empirical Testing of Random Number Generators", ACM Transactions on Mathematical Software, Vol. 33, article 22, 2007.

[11] DIEHARD battery of stringent randomness tests, G. Marsaglia, available at http://www.stat.fsu.edu/pub/diehard/

[12] Dieharder: A Random Number Test Suite, R. G. Brown, available at http://www.phy.duke.edu/~rgb/General/dieharder.php

[13] A Statistical Test Suite for Random and Pseudorandom number generators for cryptographic applications, NIST Special Publications 800-22, group of autors, available at http://csrc.nist.gov/groups/ST/toolkit/rng/documents/SP800-22b.pdf

[14] R. Fielding et al., "RFC 2616: Hypertext Transfer Protocol – HTTP/1.1", 1999.

Parallelization of the Ant Colony Optimization for the Shortest Path Problem using OpenMP and CUDA

Maida Arnautović marnautovi1@etf.unsa.ba
Maida Curić mcuric1@etf.unsa.ba
Emina Dolamić edolamic1@etf.unsa.ba
Novica Nosović nnosovic@etf.unsa.ba
Faculty of Electrical Engineering, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

***Abstract* - Finding efficient vehicle routes is an important logistics problem which has been studied for several decades. Metaheuristic algorithms offer some solutions for that problem. This paper deals with GPU implementation of the ant colony optimization algorithm (ACO), which can be used to find the best vehicle route between designated points. The algorithm is applied on finding the shortest path in several oriented graphs. It is embarrassingly parallel, since each ant constructs a possible problem solution independently. Results of sequential and parallelized implementation of the algorithm are presented. A discussion focused on implementing ACO using OpenMP and CUDA provides a basis for analysis of different results achieved on those two platforms.**

***Keywords* – Parallelization; ant colony optimization; shortest path problem; CUDA; OpenMP;**

INTRODUCTION

The shortest path problem is very important and fundamental problem in many applications. The goal of this problem is to find point-to-point shortest path in a directed graph. This paper aims to apply parallel implementation of ant colony optimization algorithm to the well known shortest path problem using CUDA on GPU and OpenMP on shared memory parallel computer architecture.

Ant colony algorithms are computational methods for solving problems that are inspired by the behaviour of real ant colonies. The decision making processes of ants are embedded in the artificial intelligence algorithm of a group of virtual ants which are used to provide solutions to the shortest path problem. The generation and evaluation of ants (solutions), which are the main parts of the algorithm and its principal source of computational cost, offers a low dependency degree, so the structure of the algorithm makes it well suited for concurrent execution. For such reasons, the parallel approach is justified and an implementation was made to study the speedup gains that could be obtained this way.

ANT COLONY OPTIMIZATION ALGORITHM

Ant Colony Optimization (ACO) is a population-based search method inspired by the behaviour of real ants searching for food. As ants find sources of food, they leave pheromone trails that allow other ants to find them and the food. Pheromones also evaporate over time, which is important so ants do not continue to go to the same food source after it has disappeared. Therefore, the ACO algorithm has two phases:

- the construction phase
- the pheromone update phase.

In the construction phase, individual ants construct a sample solution to the problem using a probabilistic function which uses a heuristic function and the amount of pheromone on the edges to decide which city to choose next. In the pheromone update phase, certain solutions (generally the best ones), deposit pheromones on the edges of their solutions. After a certain time period, ants will converge to a nearoptimal path through the graph. More about this algorithm is given in reference 2.

Fig. 1 shows the sequential ant colony optimization algorithm pseudocode.

```
optimalPath ← initializeOptimalPathToZero()
optimalPrice ← optimalPath→calculatePrice()

for every edge e in the graph
  e→initializePheromoneTrail()
end for

repeat
  for each ant in the colony do
    path ← ant→buildSolution()
    price ← path→calculatePrice()

    if price is better than optimalPrice
      optimalPath ← path
    end if
  end for

  for every edge e in the graph
    e→evaporatePheromones()
  end for

  for every visited edge v in the graph
    v→updatePheromones()
  end for
until maximum number of iterations is reached
```

Fig. 1 The sequential version of the ant colony optimization algorithm

SHORTEST PATH PROBLEM

The shortest path problem is the problem of finding a path between two nodes in a graph such that the sum of the weights of its constituent edges is minimized. The goal is to find the shortest existing path between the start node and any of the other nodes in the graph. Each path, therefore, will have the minimum possible sum of its component edges' weights.

This problem involves a weighted, possibly directed graph described by a set of edges and nodes.

Formal definition of shortest path [7]:

G = (V, E) weighted graph, directed or undirected.

Weight of path P =<v0, v1, v2, ... , vk> is

$$w(P) = \sum_{i=1}^{k} w(v_{i-1}, v_i)$$

Shortest path (u,v) from u to v has weight:

$$\delta(u,v) = \begin{cases} min\{w(P): P \text{ is path from } u \text{ to } v\} & \text{if path exist} \\ \infty & \text{Otherwise} \end{cases}$$

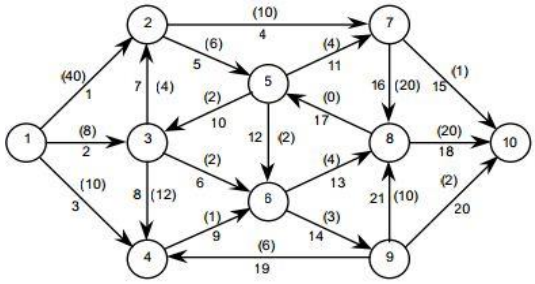


Fig. 2 Example of oriented graph [6]

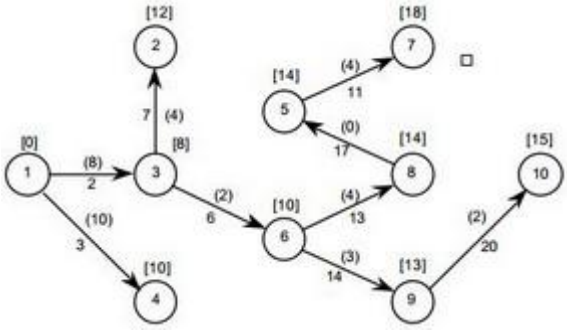


Fig. 3 Shortest path tree with the source at node 1 [6]

Fig. 3 shows the shortest path tree of network on Fig. 2 with the source at node 1.

PARALLEL IMPLEMENTATION USING OPENMP

The structure of sequential ACO algorithms is highly suitable for parallelization. The behaviour of a single ant during the iteration is totally independent of the behaviour of all other ants during that iteration. As Fig. 1 shows, operations in for loop at line 7 can all be calculated independently. Using OpenMP it is possible to experiment on parallel approach and execute parallel program. First step is recognizing parallelizable parts of algorithm and creating parallel ACO model. Fig. 4 shows parallel ACO model. At the beginning of the algorithm, a master process initializes the information needed by each ant, spawns k processes (one for each ant k), and broadcasts the information, in this case pheromone trail matrix. At the beginning of a cycle the τij matrix (the pheromone trail) is sent to each process and the computations (generation and evaluation of solutions) are done in parallel. Then, the solutions and their evaluations are sent back to the master, the τij matrix is updated and a new cycle begins by the broadcasting of the updated τij matrix. This process can be repeated limited number of time, determined prior to algorithm starting, or the finishing cycle can be determined during the runtime.

Parallel implementation of ACO algorithm can be created based on that model. Main difference would be on line 7, where it should include parallel markers and create maximum of n threads, where n is the number of ants. Then each thread executes calculations for that ant only, and master updates the results based on calculations.

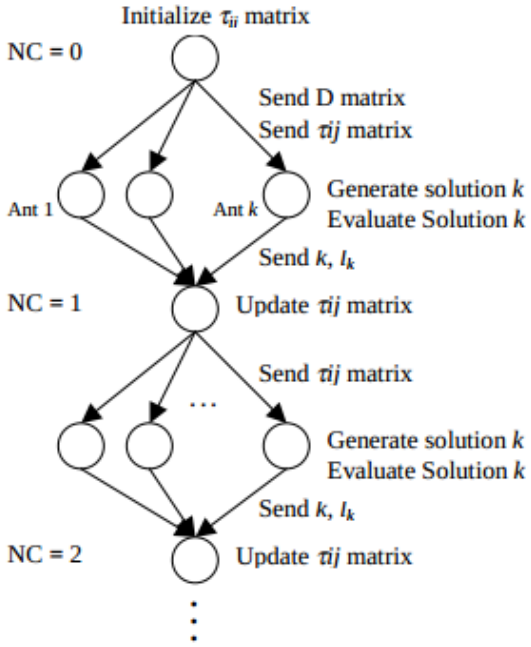


Fig. 4 Parallel ACO model [3]

Using OpenMP parallel *for* loops can be executed using *#pragma omp parallel for* with available number of threads, or number of threads to be used can be specified with *#pragma omp parallel for num_threads(n)* where n is the number of threads to be used. First step could be to use *n = number of ants*, but synchronization problems with pheromone trail matrix would cause a great loss in efficiency. Besides, number of ants can be quite big so with this method number of ants would be limited, or it would require lot of processors. That is why it is recommendable to use *n < number* of ants and *n <= number* of available processors.

```
optimalPath←initializeOptimalPathToZero()
optimalPrice←optimalPath→calculatePrice(
)
for every edge e in the graph
  e→initializePheromoneTrail()
end for

repeat
  parallel for n threads
    for each ant in the colony do
      path ← ant→buildSolution()
      price ← path→calculatePrice()

      if price is better than
         optimalPrice
        optimalPath ← path
      end if

    end for
  merge parallel threads

  for every edge e in the graph
    e→evaporatePheromones()
  end for

  for every visited edge v in the
      graph
    v→updatePheromones()
  end for
until maximum number of iterations
      is reached
```

Fig. 5 Parallel implementation of ACO

Pheromone trail matrix and best solution are kept at the master process. That is why after every iteration master has to wait for all ant threads to finish to update the pheromone trail matrix. That represents the only synchronization barrier for this algorithm. Because the best solution path is dynamically updated, that is it is kept in dynamic array, it may lead to drawback in performance.

PARALLEL IMPLEMENTATION USING CUDA

The GPU has always been a processor with ample computational resources. The most important recent trend, has been exposing that computation to the programmer. Over the past few years, the GPU has evolved from a fixed-function special-purpose processor into a full-fledged parallel programmable processor with additional fixed-function special-purpose functionality. More than ever, the programmable aspects of the processor have taken center stage. One way of harnessing this power was made possible by using the nVIDIA CUDA (Compute Unified Device Architecture) programming model. [8]

The ACO algorithm is embarrassingly parallel, since each ant does an independent task of finding a path in the graph, without any data dependencies during the solution construction phase. Considering the graphics processing unit architecture and the fact that finding the solution for every ant is embarrassingly parallel, algorithm was parallelized using the nVIDIA, CUDA-enabled GPUs, as a relatively cheap and widely available platform for high performance computing.

The graph is represented as an adjacency matrix, where each matrix element is the weight of the corresponding edge. Since CUDA allows execution of multiple threads per block in parallel, and each thread in the same block executes the same instruction, each ant is represented as a single thread. Before the algorithm can be executed, it is first necessary to initialize the graph adjacency matrix, current path matrix, optimal path matrix and pheromone matrix to their respective default values on the host, after which the memory on the GPU device is allocated, and the matrices are then transferred to the allocated memory. Path matrix and optimal path matrix are structured in such way that each row represents a single ant's path.

```
optimalPathMatrix ←
initializeOptimalPathMatrix()

pathMatrix ← initializePathMatrix()
graph ← getAdjacencyMatrix()
pheromoneMatrix ←
getInitialPheromoneMatrix()

allocateAndCopyToDevice(graph, pathMatrix,
optimalPathMatrix, pheromoneMatrix)

repeat
  pathMatrix ←
buildSolution<<<maxAntBlocks,maxAntsPerBlock>
>>(graph)

  parallel for each (path, optimalPath) in
(pathMatrix, optimalPathMatrix)
    if path is better than optimalPath
      replace optimalPath with path
    end if
  end for

  evaporatePheromones<<<numberOfNodes,
numberOfNodes>>>(pheromoneMatrix)

  updatePheromones<<<numberOfNodes,
numberOfNodes>>>(pheromoneMatrix, pathMatrix)

until maximum number of iterations is reached

copyToHost(optimalPathMatrix)
freeAllocatedResourcesOnDevice()

optimalPath ←
extractBestSolution(optimalPathMatrix)

optimalPrice ← optimalPath→calculatePrice()
```

Fig. 6 The parallel version of the ant colony optimization algorithm for the CUDA platform

The path matrix contains paths which are constructed in the current iteration of the algorithm. After each ant has finished with the solution construction, each row of the path matrix is then parallely compared to the optimal path matrix using the path length as criteria, and if the newly

constructed solution is better than the current optimal solution in the corresponding row, the optimal solution is then updated accordingly. This is done by running the kernel for updating optimal path matrix with number of threads equal to the number of ants. In the next phase, the pheromone matrix values are decreased in parallel, and afterwards updated accordingly for each row of the path matrix. The described steps are then repeated until the number of iterations exceed the predefined maximum iterations number. After that, the optimal path matrix is transferred back to the host where the best solution is extracted. Fig. 6 shows the ant colony optimization pseudocode adjusted for parallel execution on the CUDA platform.

RESULT ANALYSIS

In this paper, sequential version and parallel versions of ant colony optimization algorithm are evaluated using different parameters for ant colony size and the input graph size.

The measurements was performed on nVidia GeForce GT 420M graphics card with 96 CUDA cores and 1GB of DDR3 RAM, and AMD Athlon Neo X2 processor running at 1.5GHz. OpenMP version of the algorithm ran with two active threads.

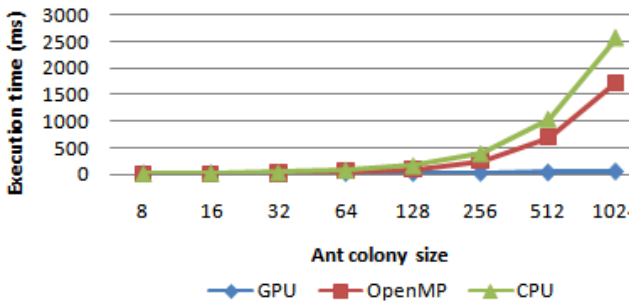


Fig. 7 Execution times in milliseconds with ant colony size increase

One of the factors which was evaluated was the influence of the ant colony size increase on the execution times. Fig. 7 shows the execution times in milliseconds as the ant colony size increases. The difference between execution times for examples becomes more noticeable when the population of ant colony reaches a number of 128 ants. The GPU implementation performs much better as the ant colony population increases, because all of the ants are being processed at once, whereas the sequential version has to process them one by one, hence increasing the execution time. The OpenMP parallelized version of algorithm performs better than CPU proportionally to the number of running threads.

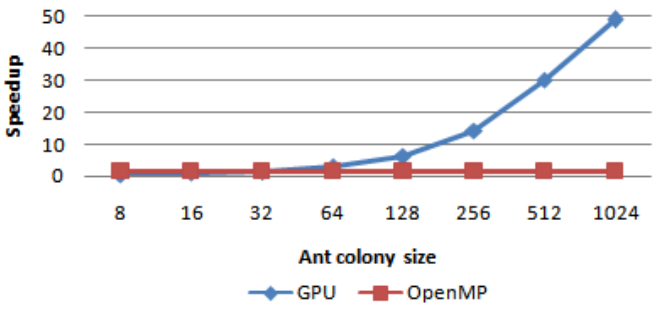


Fig. 8 Speedup based on ant colony size increase

Fig. 8 shows speedup gained by using GPU ant colony optimization algorithm with the increase of the ant colony size. Reason for the very low speedup achieved with GPU for the ant colony size of eight is due to the fixed overhead of kernel launching and the lack of sufficient number of threads required to hide the latency.[8] The CPU execution time is lower than the time necessary to transfer all the data from host memory to device memory, making the CPU version, and with it the OpenMP version of the algorithm an undoubtedly better choice. Fig. 9 shows the comparison of the sequential algorithm, OpenMP parallelized algorithm and the GPU parallelized algorithm execution times for eight, sixteen and thirty-two ants.

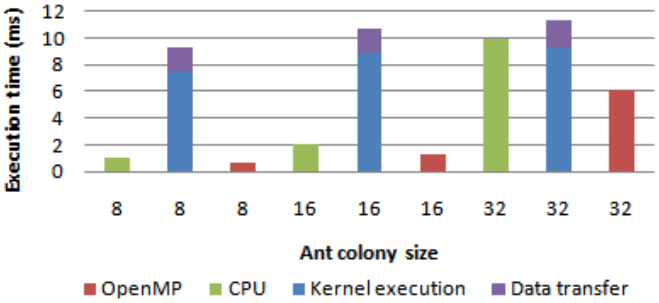


Fig. 9 CPU and GPU execution times comparison: for smaller ant colony sizes GPU parallelized algorithm does not provide speedup which makes the sequential algorithm a better choice.

For smaller ant colony sizes the GPU algorithm performs much worse, as there are not enough threads to fully use the GPU potential. The speedup gained using GPU starts low and increases as the ant colony size increases.

Fig. 10 shows the influence of the graph size increase on the execution times. Larger graph implicates bigger individual ant's path, which in turn increases the total execution time. Since all ants on CPU execute sequentially, this extra time accumulates much faster than on GPU. It is also worth noting that the memory requirements increase also limits the size of problems on which ant colony algorithm can be applied using GPU, considering the fact that the amount of memory on GPU is limited.

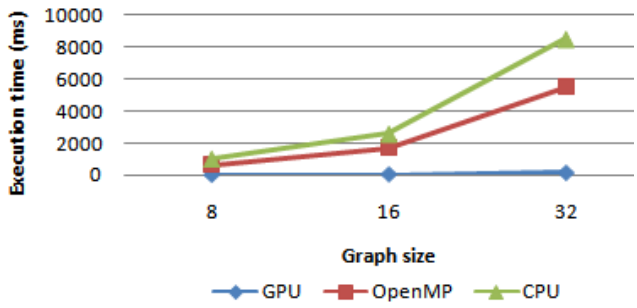


Fig. 10 Execution times in milliseconds with graph size increase

Fig. 11 shows the influence of graph size increase on the speedup factor. GPU provides larger speedup gain as the problem size increases, and OpenMP provides speedup proportional to the number of running threads.

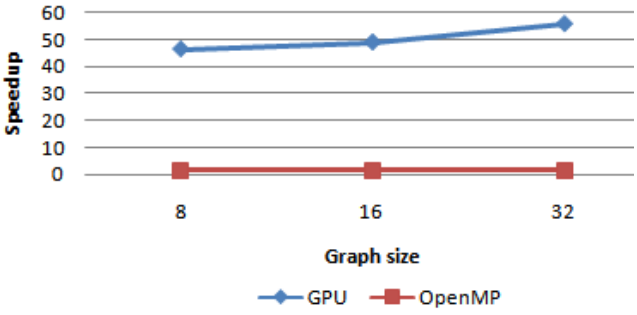


Fig. 11 Speedup based on graph size increase

CONCLUSION

In this paper two ways of parallelizing ant colony algorithm are presented. Results show significant speedup compared to the sequential version of the ant colony algorithm. The first presented parallelization method was OpenMP parallelization technique which is a way of parallelizing sequential code using multiple threads. Obviosly, the speedup gained by this method of parallelization is limited by the number of CPU cores. Other presented parallelization method was based on CUDA, which allows usage of GPU computing power for general problems. Results show up to fifty times faster execution when using the GPU, making the GPU parallelization method better choice over OpenMP parallelization. The application of GPU computing power is limited by the GPU's memory capacity which constrains the problem size. To bypass this constraint, further focus should be given on increasing data parallelism so the problem could be computed in smaller chunks, and possibly using more GPUs in parallel.

REFERENCES

Daniel Angus, *Solving a unique shortest path problem using ant colony optimisation*, 1998.

Stefan Janson, Daniel Merkle, and Martin Middendorf, *Parallel Ant Colony Algorithms*, Parallel Computing and Complex Systems Group, University of Leipzig, D-04109 Leipzig, Germany

Pierre Delisle, Michaël Krajecki, Marc Gravel, Caroline Gagné, *Parallel implementation of an Ant Colony Optimization metaheuristic with OpenMP*

NVIDIA CUDA C Programming Guide, v4.2, April 2012, Available at: http://developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/CUDA_C_Programming_Guide.pdf

Enrique Alba, *Parallel Metaheuristics: A New Class of Algorithms*, Wiley-Interscience, 1st ed, September 8, 2005

Paul Jensen (1999). *Solution Methods for the Shortest Path Tree Problem, 5.2 Shortest Path Problem*. Available at: http://www.me.utexas.edu/~jensen/methods/net.pdf/netshp.pdf

Laura I. Toma (2009*). Introduction to the Design and Analysis of Algorithms* [Lectures].Retrieved from: http://www.bowdoin.edu/~ltoma/teaching/cs231/fall07/Lectures/shpaths.pdf

[1] John D. Owens, Mike Houston, David Luebke, Simon Green, John E. Stone, and James C. Phillips (2007*). GPU Computing*. Available at: http://cs.utsa.edu/~qitian/seminar/Spring11/03_04_11/GPU.pdf

Protein database search optimization based on CUDA and MPI

Dario Pavlović*, Robert Vaser* , Matija Korpar* , i Mile Šikić*
* Faculty of electrical engineering and computing, Unska 3, 10000 Zagreb, Croatia
contact e-mail: dario.pavlovic@fer.hr

Protein database search is an important method in the field of computational biology. There are a large number of sequences in an average database which makes such searches rather time and resource consuming. With the rapid growth in size of these databases in the past years, there came a need to speed up the search and consequently, any alignments performed on such databases.

This paper presents an acceleration of the database search tool sw#DB which is based on a CUDA implementation of Smith-Waterman algorithm. We achieved speed up by reducing database size. The whole database was divided into seeds of a fixed length. The positions of these seeds and the corresponding sequence indexes from the database are then stored in a hash container. This allows for a constant time lookup of all the positions of a seed in every sequence of a database. Potential alignment candidate sequences for a query are filtered using this method, forwarding only those which contain at least one seed from the query to the sw#DB. This reduces the number of alignments performed. Overall, it brings a speedup of around three times compared to the basic sw#DB tool, based solely on Smith Waterman algorithm, with almost no loss of accuracy.

The implementation is written in CUDA and C programming languages. For large queries, MPI implementation with multiple CUDA cards is used.

I. INTRODUCTION

The ever more increasing importance and usage of protein database search in various scientific fields, especially computational biology, makes finding a fast and accurate way to obtain search results as important as ever. The most widely used database search method is BLAST [1][2]. Over time, there have been numerous proposals of how to speed up BLAST and its modules, often at a large cost to its accuracy. Some of these improvements were based on deterministic finite automatons, while others considered speeding up BLAST through modifications made in seeding phase [6][7].

In this paper, we consider another database search tool named sw#DB which is still unrealased. sw#DB (pronounced swSharpDB) is a tool that makes full use of modern 'graphics processing unit' (GPU) computing architecture and its advantages. It is based on Nvidia's 'Compute unified device architecture' (CUDA) technology and implements a parallel Smith-Waterman algorithm. It can run on multiple GPUs at the same time. This enables for huge performance gains on larger databases, such as Uniprot[3], with no loss of accuracy. sw#DB software uses the Smith-Waterman algorithm primarily because of its advantages over BLAST. BLAST is a heuristic algorithm and does not guarantee optimal local alignments. On the other hand, Smith-Waterman is an algorithm which guarantees optimality in alignments. Although GPU architecture enables fast parallel calculations and speed up of Smith-Waterman algorithm, there are further possibilities for improvement of this program. Herein, we propose a couple of optimizations of the sw#DB early phases, aimed at reducing the number of alignments performed in its central stage and ultimately decreasing its runtime.

II. METHODS

To accomplish the improvements to sw#DB, we added a new first stage to it, similar to BLAST's seeding stage. This stage has two substages.

In the first substage, the whole database is stored in a lookup table in the form of pairs which contain seeds of lengths four or five, depending on the implementation, and their respective locations in the database. This means that every possible protein seed of length four or five is located in every sequence of the database, if in any, and its located positions are stored in the lookup table. Lookup table is indexed using ASCII values of the characters in a seed. We accomplished this using 5-bit integers for each character. For example, if a seed is of length four, it is split in four characters, each of which is used as a 5-bit number. Upper three letters of the seed are shifted to the left to reflect their respective positions in the seed and summed with the last letter which is not shifted at all. In this way, they form a 20-bit number which becomes index into the lookup table. All of the seed's positions in the database are then stored in the hash container using that index. If a file with the hashed database seeds already exists, this stage is skipped, saving more time.

The second substage is the core stage of the optimization and is similar to the first substage with a difference in that it actually reads in the positions of all possible seeds from the hashed database file and uses this information for further query processing. In this substage, query is split into seeds of the same length as used in the database lookup table creation. These query seeds are then used as 25-bit number indexes utilized to find all their positions in the database in constant time. This produces a number of database sequences, all of which contain at least one seed the same as in the query. All these

sequences are then forwarded, along with the query, to the Smith-Waterman stage in which the actual alignment is performed. This alignment phase is different than what BLAST does in the sense that the alignment is done on the whole sequences, and not just in the high scoring regions around seeds. We also considered a hybrid implementation in which the sequence selection criteria was that each sequence must contain two exact seeds as in the query. This brings more performance gains compared to only one seed criteria with a good level of accuracy.

The second significant improvement we introduced to the sw#DB is the use of another parallel programming model, the Message passing interface, commonly known as MPI. MPI enables programs to run on clusters comprised of many computers working in parallel. The MPI implementation makes use of the fact that different queries can be processed in parallel, not only on multiple GPUs, but also on multiple computers with many GPUs, all at the same time. We changed and expanded the sw#DB implementation to be able to use this concept. In our implementation, there is one process per cluster node. Each process, depending on its rank, is associated with a different part of the query input file and runs that part of queries through the sw#DB.

III. RESULTS AND DISCUSSION

For the evaluation of accuracy and speed, we used SCOP/ASTRAL release 1.75[4] and Uniprot databases. The former was used for identifying true and false positives and plotting 'receiver operator characteristic' (ROC) curves and the latter was used for the speed testing and comparison with the original sw#DB and blastp modules.

ROC_{10000} curves[5] were used to plot and analyze accuracy for each implementation. All results obtained from the input queries were sorted by E-value in ascending order and all of them with up to 10000 false positive were taken into account. A result is a true positive if the query and a resulting sequence are in the same superfamily. Otherwise, it is a false positive. While querying the Astral database, we only considered the sequences which have at least one other member of the same superfamily as queries. This equaled to about 9000 queries on an 11000 sequence Astral database.

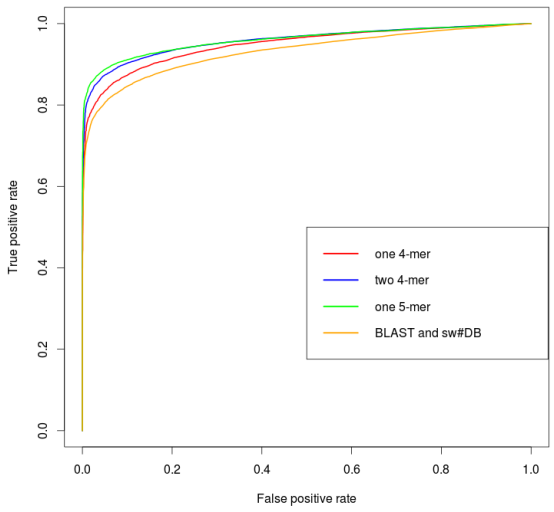


Figure I. ROC_{10000} curves for various seed sizes compared to default sw#db and BLAST

The resulting ROC curves for default sw#DB, blastp and our optimizations are shown in the Figure 1. It is clearly seen that the accuracy of the default implementation was not compromised. In fact, the graph shows slightly higher AUC values than with the original sw#DB and BLAST. Also, the curves for BLAST and sw#DB are the same, so there is only one curve representing both.

We performed the timing test using 500 sequences selected from the Astral database. These sequences were queries into the Uniprot database. We compared the running time to blastp and default sw#DB implementation. All software was tested on an Intel Core 2 Quad Q6600 CPU with Nvidia Geforce GTX 570 GPU. Table 1 shows the resulting running times. It is clear that the default blastp implementation is faster than the pure sw#DB, but that is primarily because of the time needed to allocate GPU memory and transfer data to the GPU. This time difference would go down if the number of queries was larger. On the other hand, our improvements considerably reduce run time of sw#DB, and as Table 1 shows, it is faster than BLAST in both variants. Seeds of size four don't affect overall accuracy, based on ROC_{10000}, and still manage to overcome blastp.

Additional improvements to the execution speed of sw#DB can be made using MPI. Our implementation of sw#DB using MPI greatly reduced run time with the seeds of length four and almost halves the run time of seeds of length five. We tested the MPI implementation using GPU Hybrid cluster with six nodes activated at a time. Each node has an eight core Intel Xeon CPU, 8GB of RAM and 2x Nvidia GT200 graphics. Testing data was the same as in the first speed testing - 500 sequences from the Astral database used as queries to the Uniprot database. Table 1 shows the results of this test for the seeds of length four and five, respectively.[1] The reduction in run time is quite large and we assume that a cluster with more modern hardware, primarily modern GPUs, and more nodes would improve these times even more. It is quite obvious that the greatest speedup with MPI is accomplished with the version of sw#DB which has no modifications. This means that it is necessary to additionally analyze and improve running time of the hashing and seeding parts in the improved versions.

There is further room for improvement of sw#DB, mainly in the MPI part of the optimizations, which are obviously still non-optimal. Additionally, we will consider adding support for mutations in the query seeds, which should give more results in the first phase, thus increasing accuracy. Furthermore, we are a bit sceptical about the validity of ROC_{10000} method and its precision, because of the fact that, with reducing the search space, ROC_{10000}

[1] BLAST currently does not support MPI

curves show better accuracy, which is highly unexpected. We do believe that the actual accuracy does not significantly deviate from the shown results.

TABLE I. RUN-TIMES OF SW#DB AND BLAST WITH VARIOUS SEED SIZES

Seeds/Parallelization	No MPI (seconds)	MPI (seconds)	Speedup
BLAST	460.0	-	-
sW#DB	811.6	233.7	3.47
One 4-mer	421.3	169.6	2.48
Two 4-mer	283.5	156.0	1.82
One 5-mer	130	74.9	1.74

IV. CONCLUSION

In this work, we described a new optimization incorporated into sw#DB sofware which greatly reduces its runtime. We base our improvements on adding a new start stage to the software. This stage is similar to the seeding stage of BLAST and is used to reduce the number of alignments performed in the later phases. We have measured performance of the program with improvements incorporated and have shown that it can be faster than BLAST with larger databases, but at the same time mantaining accuracy level.

REFERENCES

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–10, Oct. 1990.

[2] M. A. Liebert and C. E. T. Al, "Hit Detection in BLAST," vol. 13, no. 4, pp. 965–978, 2006.

[3] T. U. Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt).," *Nucleic acids research*, vol. 40, no. Database issue, pp. D71–5, Jan. 2012.

[4] J.-M. Chandonia, G. Hon, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner, "The ASTRAL Compendium in 2004.," *Nucleic acids research*, vol. 32, no. Database issue, pp. D189–92, Jan. 2004.

[5] M. Gribskov and N. L. Robinson, "USE OF RECEIVER OPERATING (ROC) ANALYSIS TO EVALUATE SEQUENCE MATCHING," vol. 20, no. I, pp. 25–33, 1996.

[6] M. Cameron, H. E. Williams and A. Cannane, "A deterministic finite automaton for faster protein hit detection in BLAST.," *Journal of computational biology*, vol.13, no. 4, pp. 965-978, 2006.

[7] J. Tan, D. Kuchibhatla, F. L. Sirota, W. A. Sherman, T. Gattermayer, C. Y. Kwoh et al., "Tachyon search speeds up retrieval of similar sequences by several orders of magnitude.," *Bioinformatics*, vol. 28, no. 12, pp. 1645–1646, April, 2012.

Parallel Implementation of the Ant Colony System Algorithm for Automated Combinational Logic Circuits Design Using CUDA Architecture

Džanan Bajgorić, Adnan Porović, Tarik Handanović, Novica Nosović

University of Sarajevo, Faculty of Electrical Engineering
Department for Computer Science and Informatics

{dbajgoric1, aporovic1, thandanovi1, nnosovic}@etf.unsa.ba

***Abstract* – This paper propose parallel implementation of Ant Colony System (ACS) algorithm for automated combinational circuit design. Ant Colony System is one of the most popular and widely used Ant Colony Optimization (ACO) algorithm and heuristic algorithm in general. As digital logic circuits become more complex, efficient circuit design is priority and use of heuristic methods are unavoidable. Unfortunately, the optimization problems became so complex in sense of their size, even the most powerful heuristic algorithms can't solve them on single CPU. In order to be able to tackle the problem, parallel version of ACS is needed and this paper presents CUDA (Compute Unified Device Architecture) C language implementation.**

Keywords: Ant Colony System, logic circuits design, CUDA, parallel implementation, sequential implementation

I. INTRODUCTION

Several methods have been developed for solving the combinational logic circuits design (for example Quine – McCluskey method and Karnaugh Maps), and some of them can handle extremely complex circuits with hundreds of inputs. Heuristic algorithms, including Ant Colony System, are restricted to relatively small truth tables and therefore can't be used for solving complex problems in circuit optimization. This paper tries to overcome this difficulty by using GPU computing power. The main goal of this paper was to implement ACS algorithm for digital circuit design using CUDA architecture, in order to improve not just performance but also quality of solutions i.e. generated circuits. Sequential (CPU) version of the algorithm is implemented using standard C language, while for the parallel version CUDA C language was used.

GPGPU (general programming graphics processing unit) – using a GPU processor for arbitrary computations. At the very beginning, GPUs were exclusively used for graphics programming (i.e. 3D computer games). As their computational power increased, GPUs started to attract the experts from other fields of science, who wanted to program the graphics processor to do arbitrary computation.

It was NVIDIA who first offered complete solution to this problem with their CUDA architecture. CUDA allows access and programming of every available ALU on the card, and also has modified instruction set that simplifies general purpose programming. Onwards, every ALU can arbitrarily read and write memory – before there was no choice but to let graphics API and driver manage memory access. Also the new software driven cache called **shared memory** is introduced. CUDA architecture and programming are explained in great detail in [1] and [2].

Ant colony optimization is a metaheuristic in which a colony of artificial ants cooperate in finding good solutions to hard discrete optimization problems [4]. The oldest ACO algorithm is Ant System (AS), developed in 1992 [4]. AS was used in other papers that try to solve the problem of digital circuit design by application of ACO heuristic. However, this paper doesn't use AS because it has the worst performance of all ACO algorithms. Instead it was decided to use Ant Colony System, as one of the several modifications of AS algorithm. There are three main aspects of Ant colony system algorithm: **solution construction**, **global pheromone update** and **local pheromone update**. AS, ACS and many other ACO algorithms are described in great detail in [4].

II. RELATED WORK

The authors couldn't find any registered papers describing the parallel implementation of any ACO algorithms for digital circuit optimization problem. However, there are papers that exploit the possibility of using heuristic algorithms for mentioned problem.

The first work in evolutionary design of digital circuits was proposed in [7]. Genetic algorithms were also used for digital circuit design in [6]. This paper is in fact the

basis for the [3], where AS algorithm is used for solving the problem. In [6] it is shown that Breeder Genetic Algorithm (BGA) can be efficiently used to design relatively small digital circuits, if convenient problem representation is adopted. Paper [3] extends the work and idea of paper [6] to Ant System algorithm. It is said that AS in general produces better results than corresponding BGA implementation. Papers [8,9] claim that using the evolutionary algorithms it is possible to build certain arithmetic circuits, that can't be built using conventional human designer methods.

As seen, there were certain number of attempts to solve digital circuit design optimization problem using heuristic algorithms, but none of them used parallel implementation and specifically none of those approaches used GPU (namely CUDA architecture) for algorithm implementation. It is important to acknowledge that this paper heavily relies on the work in [3], with one important exception which is the use of Ant Colony System while in [3] AS was used. This paper tries to exploit the CUDA architecture computing power in solving of the problem of digital circuits design, as this could improve the performance but also the quality of the constructed solutions.

III. ACS FOR LOGIC CIRCUIT DESIGN – SEQUENTIAL ALGORITHM

In this section the used approach is described, including problem representation, solution construction, global pheromone update and local pheromone update. Most of the information presented in this chapter corresponds to both versions of algorithm (sequential and parallel), but there are some differences in the parallel algorithm that will be addressed in the chapter four.

A. Problem representation

In order to use ACS efficiently, the problem must be represented as the problem in which it is needed to find an optimal path in some graph. However, the task is to develop an algorithm that is able to build digital logic circuits using artificial ants.

One idea, that is also used in [3], is to represent gates (gates are building elements of logic circuits) using the matrix notation. In other words, a MxN matrix is used, where each entry represents a gate. Ants are supposed to traversal this problem matrix, choosing one gate at each step and that way gradually build a whole circuit. In this paper, five types of gates are used: AND, OR, NOT, XOR and WIRE. WIRE is used to represent an absence of the gate – WIRE has just one input that is also its output. In this paper, only circuits with one output were used.

In the beginning of the algorithm, a problem matrix of a given dimension (dimensions are one of the parameters) is randomly generated. For every matrix entry a completely random gate is chosen. One important property of used approach is that these gates are static, that is they don't change during the algorithm execution, while connections between them do change.

The second step is to again randomly generate connections between the gates. Basically, every gate consists of three attributes: two inputs and its type. So for every gate in the matrix, two inputs from the previous column are chosen.

The third step in the preprocessing stage is to calculate the value of the objective function for each gate. The value of the objective function is calculated for each gate, as the number of matches between the output of the partial circuit, whose terminal gate is the gate in question, and the output defined with the truth table.

The fourth and the final step of the preprocessing stage is the computation of the so called the distance matrix. An entry (i,j) in the distance matrix is calculated as the difference between the objective function value in gate "j" and the same value in gate "i". The higher the distance the better, because the higher distance means that the circuit with the terminal gate "j" is better than the circuit with the terminal gate "i" i.e. the number of matches of its output is higher.

B. Building a circuit

Once the problem instance is established, ants can start building their circuits. An ant has to start the process of building a circuit in some gate, and we decided to randomly place an ant to the gate in the first column. So the first step of the building process is to randomly place a colony of ants to the gates in the first column.

Then, ants start to gradually build their circuits by choosing gates in the neighborhood of their current gate. The ACS algorithm's pseudo – random proportional rule [4] was used for this purpose. There are three possibilities when moving from one gate to the other: **1)** randomly generated number q is less than the parameter q_0, when the best unvisited neighbor is chosen; **2)** number q is greater than the parameter q_0, when the roulette wheel selection is used to determine the next gate from the neighborhood. And the third **3)** possibility is actually a special case. It happens when an ant has already chosen all of the gates in certain column (theoretically possible) and is now stuck in the gate that has no neighbors in the next column. In this situation "next – column – choose – best" technique is used, where the best gate in the next column is found, and then the ant is placed to that gate.

The process of building a solution terminates when an ant chooses a gate in the last column, and that gate is considered to be the terminal gate of the built circle. The next part is the solution evaluation. The first step is to check if the built solution is at least feasible i.e. the number of matches between the output of the final gate is identical to the output defined by the truth table. If that's not the case, the solution is discarded as only feasible solutions are considered for the optimal one. If the solution is feasible, the next step is to check if it is shorter than the current best solution. The circuit "i" is shorter than the circuit "j", if it has less gates, whereby gates of type WIRE are not counted when calculating the circuit's length.

C. Local pheromone update

Every time an ant moves from the gate "i" to the gate "j", it uses the local pheromone update procedure to update the pheromone on the arc (i,j).

D. Global pheromone update

The purpose of the global pheromone update is to increase the amount of the pheromone on the arcs that belong to the best so far tour. In our problem, the procedure of global pheromone update has the task to update the pheromone value on the connections that belong to the best so far circuit.

E. Reconnecting gates

After all of the ants have constructed their solutions and global pheromone update is done, it is necessary to reconnect the gates in the problem matrix that don't belong to the best so far circuit. The reconnect procedure is needed because it shows that the initial search space isn't big enough (i.e. the ants can't find good solutions). It is important to underline that the reconnect gates procedure affects only the gates not included in the best so far circuit. That way the best solution is kept while also continuing to look for better solutions by modifying the search space.

IV. PARALLEL IMPLEMENTATION

All of the algorithms from the ACO family including ACS, are suited for parallel processing because every ant in the colony can construct its solution independently, without direct communication with other ants. Of course ants do communicate through the pheromone deposit but that's not direct communication so it doesn't require ants synchronization. Not every part of the algorithm can be parallelized though, because global pheromone update must be performed only once in every iteration. First of all, the authors would like to acknowledge that the best available graphics card was NVIDIA 8600 GTS, whose computing capability is 1.1.

The first question to answer when implementing the ACS using CUDA is: how to model an ant? It was decided to model an ant as a **block** [1,2], that has a number of **threads** [1,2] that do its work in parallel. This approach is convenient, because the ants are already independent during the solution construction phase, and CUDA blocks are also independent. Furthermore, if each ant has at his disposal a number of treads, it'll be possible to parallelize procedures like choosing the next gate (using the best neighbor next and roulette wheel selection technique). Almost every nested for loop in the sequential version of the algorithm, has been converted to make use of enormous number of threads that are on our disposal. The shared memory is also heavily used in these inner procedures (used locally on the level of ants), to reduce the read and write time, and to reduce the memory traffic that is the bottleneck of the today GPUs.

Not only the local procedures that are used by ants individually are parallelized. Global procedures that need to modify the data shared by the entire ant colony, are also changed to allow utilization of the massive number of threads that can be used. However, while mentioned local procedures are parallelized on the level of a block, i.e. using only the threads from the block that executes at the moment, parallel versions of the global procedures make use of the entire set of the threads that the main kernel has been initiated with.

V. MEASUREMENTS AND RESULTS ANALYSIS

In all of three examples that will be shown, the problem matrix has 30 rows and 5 columns, and the number of iterations is 30. Also, every circuit has 3 inputs i.e. the output has 8 possible values.

The outputs of the circuits are defined as follows:

Outpu1 = {1,0,1,1,1,0,0,1}
Outpu2 = {1,1,1,1,1,0,0,1}
Outpu3 = {0,0,0,1,1,1,1,0}

The measurements results for every circuit for **sequential** C implementation, are given in the following table:

Circuit	Boolean function	Length	Time
1.	$\overline{\overline{\overline{A} \oplus \overline{B} \oplus (\overline{A \oplus C})\left[\overline{A \oplus (B + C)}\right]}}$	12	60ms
2.	$\overline{[(A + A) \oplus AB] \oplus [(A + A) \oplus AC]}$	8	57ms
3.	$\overline{\overline{A} \oplus BC}$	4	60ms

and the results for **parallel CUDA** implementation are given in the next table:

Circuit	Boolean function	Length	Time
1.	$[A \oplus C \oplus (A+C) \oplus A \oplus B] + \overline{A+B}\,\overline{A+C}$	11	20ms
2.	$\overline{A[(B \oplus C)+(C \oplus C)]}$	7	17ms
3.	$\overline{A \oplus \overline{BC}}$	6	21ms

The first conclusion that can be derived from the results is that the parallel implementation is more efficient in the terms of performance. For all of the 3 circuits, the parallel algorithm found the solution at least 3 times faster that the sequential one. The speed up wasn't remarkable and the reason is that not every part of the algorithm could be parallelized. Some procedures like local pheromone update, circuit length computation are executed by only one thread per block. Further, some procedures like global pheromone update are executed by only one of all available threads (including all blocks). Another reason that prevented the linear speed up is memory bandwidth and latency. As indicated in [1,2], memory traffic is the bottleneck of the modern GPUs and it is priority to reduce the memory traffic as much as possible. The implemented algorithm ignores this practice in a way that every data structure shared by every ant is placed and accessed in global GPU memory. This is extremely slow, and should be avoided to get the best possible performance.

The second conclusion is regarding the circuit length i.e. the number of gates in the circuit. It shows that in this aspect the performance of the CUDA implementation is comparable to the performance of the sequential implementation. This is because the sequential and parallel implementation essentially do the same amount of computations, the only difference is that the parallel version does the work in much less time. So it is understandable that very similar circuits will be produced.

VI. CONCLUSION AND FUTURE WORK

This paper presents the parallel implementation of the Ant Colony System algorithm for combinational logic circuits design, using CUDA architecture. Two versions of the same algorithms were developed: the sequential and parallel version. As it was expected, the parallel CUDA algorithm performs better than the sequential one, but the speed up wasn't extraordinary. The main reasons for this are the deficiencies of the algorithm and the parts of the algorithm that couldn't be completely parallelized. Current sequential algorithm is able to solve problems that have at most **three** inputs. On the other side, the parallel implementation can handle the circuits that have a maximum of **five** inputs – even though the circuits of this size are considered to be quite small, the fact that CUDA C implementation can handle almost two times bigger circuits shows that CUDA architecture can be successfully used to tackle the problems that could never be solved using the sequential algorithm. In order to be able to handle even more complex circuits, the algorithm itself must be improved, as well as the parallel implementation that currently has several deficiencies.

Therefore, out plan is to improve the sequential implementation of the algorithm to allow its application in solving of more complex problems. One modification that will be made is the regeneration of the GATEs that don't belong to the best so far circuit. This modification will expand the search space, and will therefore increase the probability of finding the feasible (and possibly optimal) solution.

After the main idea of the algorithm is improved, the parallel implementation will also be modified and the other approaches that will allow more aggressive use of the shared memory will be tried. Our goal is to make the adjacency list, pheromone matrix and the total matrix shared, and that would certainly have an enormous positive impact on the performance and, through the performance, on the quality of constructed solutions.

REFERENCES

[1] Jason Sanders, Edward Kandrot: "Cuda by Example: An Introduction to General – Purspose GPU Programming", Addison – Wesley Professional, July 29, 2010.

[2] Rob Farber: "Cuda Application Design and Development", Morgan Kaufmann, November 14, 2011.

[3] Coello Coello, C.A., Gutierrez, R.L.Z., Garcia, B.M. and Aguirre, A.H. (2000). Ant colony system for the design of combinational logic circuits. *Evolvable Systems: From Biology to Hardware*, Edinburgh, Scotland, (Miller, J., Thompson, A., Thomson, P. and Fogarty, T. C., Eds.) Springer – Verlag, 21-30.

[4] Marco Dorigo, Thomas Stutzle: "Ant Colony Optimization", Bradford Books, June 4, 2004.

[5] Mostafa Abd-El-Barr, Sadiq M. Sait, Bambang Sarif, Ant Colony algorithm for evolutionary design of arithmetic circuits, Microelectronics, 2003. ICM 2003. Proceedings of the 15th International Conference on, Cairo, Egypt, 9-11 Dec. 2003 Page(s): 198 – 201, December 2003.

[6] Carlos A. Coello, Alan D. Christiansen, and Arturo Hernandez Aguirre. Automated Design of Combinational Logic Circuits using Genetic Algorithms. In D. G. Smith, N. C. Steele, and R. F. Albrecht, editors, *Proceedings of the International Conference on Artificial Neural Nets and Genetic Algorithms*, pages 335 – 338. Springer – Verlag, University of East Anglia, England, April 1997.

[7] Sushil J. Louis, Genetic Algorithms as a Computational Tool for Design, Ph.D. thesis, Department of Computer Science, Indiana University, Aug 1993.

[8] J. F. Miller, T. Fogarty, and P. Thomson, "Designing Electronic Circuits Using Evolutionary Algorithms. Arithmetic Circuits: A Case Study", *Genetic Algorithms and Evolution Strategy in Engineering and Computer Science, John Wiley and Sons, Chichester,* pp. 105 – 131, 1998.

Performance Analysis of the Parallelized Matrix Multiplication Using C# Threading

Merima Delić, Edina Omerović, Azemina Sokolović, Novica Nosović

Faculty of Electrical Engineering, University of Sarajevo Sarajevo, Bosnia and Herzegovina
md15384@etf.unsa.ba, eo15363@etf.unsa.ba, as15181@etf.unsa.ba, nnosovic@etf.unsa.ba

Abstract-**This paper presents parallel matrix multiplication in C#. Threaded computation was distributed to all available cores in the shared memory system. Number of threads created to parallelize for-loops is equal to the number of cores in the system. The main aim of this work was to accomplish parallelization similar to what Microsoft .NET V4.o provided in their parallel extensions library and make it easy to use.**

Keywords—Matrix multiplication, parallel algorithm, .NET Framework.

I. Introduction

This paper presents parallel **matrix multiplication**. Threaded computation was distributed to all available cores in the shared memory system. Number of threads created to parallelize for-loops is equal to the number of cores in the system. The main aim of this work was to accomplish parallelization in C# similar to what Microsoft .NET V4.o provided in their parallel extensions library and make it easy to use. **Matrix multiplication** is a operation that takes a pair of matrices, and produces another matrix. This term may refer to a number of different ways to multiply matrices, but most commonly refers to the matrix product. There are two types of multiplication for matrices: scalar multiplication and matrix product (two matrices).

A. Scalar multiplication [1]

The simplest form of multiplication associated with matrices is scalar multiplication. There are two types of scalar multiplication: left and right. The left multiplication of a matrix **A** with a scalar λ gives another matrix $\lambda\mathbf{A}$ of the same size as **A**. The entries of $\lambda\mathbf{A}$ are given by

$$\lambda\,(\mathbf{A})_{ij} = (\lambda\mathbf{A})_{ij} = \lambda\mathbf{A}_{ij}$$

explicitly:

$$\lambda\mathbf{A} = \lambda\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix} = \begin{pmatrix} \lambda A_{11} & \lambda A_{12} & \cdots & \lambda A_{1m} \\ \lambda A_{21} & \lambda A_{22} & \cdots & \lambda A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda A_{n1} & \lambda A_{n2} & \cdots & \lambda A_{nm} \end{pmatrix}.$$

Similarly, the right multiplication of a matrix **A** with a scalar λ is defined to be

$$(\mathbf{A}\lambda)_{ij} = \mathbf{A}_{ij}\lambda.$$

B. Matrix product (two matrices)

If **A** is an $n \times m$ matrix and **B** is an $m \times p$ matrix, the result **AB** of their multiplication is an $n \times p$ matrix defined only if the number of columns m in **A** is equal to the number of rows m in **B**. When multiplying matrices, the elements of the rows in the first matrix are multiplied with corresponding columns in the second matrix. One may compute each entry in the third matrix one at a time.

For two matrices:

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mp} \end{pmatrix}$$

(where necessarily the number of columns in **A** equals the number of rows in **B** equals m) the matrix product **AB** is defined by

$$\mathbf{AB} = \begin{pmatrix} (AB)_{11} & (AB)_{12} & \cdots & (AB)_{1p} \\ (AB)_{21} & (AB)_{22} & \cdots & (AB)_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ (AB)_{n1} & (AB)_{n2} & \cdots & (AB)_{np} \end{pmatrix}$$

with no multiplication signs or dots) where **AB** has entries defined by

$$(AB)_{ij} = \sum_{k=1}^{m} A_{ik}B_{kj}.$$

Matrix-Multiply(A; B)

- ***n*** *= A:rows*
- ***let*** *C be a new n * n matrix*
- ***for*** *i = 1 to n*
- ***for*** *j = 1 to n*
- *cij = 0*
- ***for*** *k = 1 to n*
- *cij = cij + aik *bkj*
- ***return*** *C*

Fig. 1. Pseudo code implementation for ***matrix multiplication***

C. Properties of ***matrix multiplication***

Analogous to numbers , matrices satisfy the following general properties. Although there is one subtlety, due to the nature of matrix multiplication.

1. Not commutative
   In general $AB \neq BA$ because AB and BA may not be simultaneously defined, and even if they are they may still not be equal.

2. Associative

$$A(BC)= (AB)C$$

3. Distributive

$$A(B+C)=AB+AC,$$
$$(A+B)C=AC+BC$$

4. Scalar multiplication

$$\lambda (AB) = (\lambda A)B \text{ and } (AB) \lambda = A(B\lambda)$$

5. Transpose

$$(AB)^T=B^TA^T$$

where T denotes the transpose, the interchange of row *i* with column *i* in a matrix. This identity holds for any matrices over a commutative ring, but not for all rings in general.

6. Hermitian conjugate

If **A** and **B** have complex entries, then

$$(AB)^\dagger = B^\dagger A^\dagger$$

where † denotes the Hermitian conjugate of a matrix (complex conjugate and transposed).

7. Traces

The trace of a product **AB** is independent of the order of **A** and **B**:

$$tr(AB)=tr(ba)$$

II. **MATRIX MULTIPLICATION** PARALLELIZATION

Matrix multiplication has very specific characteristics as regards the design and implementation of a **parallel algorithm** in the parallel algorithms context in general:

1. Computation independency: each element computed from the result matrix C, cij, is, in principle, independent of all the other elements. This independence is utterly useful because it allows a wide flexibility degree in terms of parallelization.

2. Data independence: the number and type of operations to be carried out are independent of the data. In this case, the exception is the algorithms of the so-called sparse **matrix multiplication**, where there exists an attempt to take advantage of the fact that most of the matrices elements to be multiplied are equal to zero.

3. Regularity of data organization and of the operations carried out on data: data are organized in two-dimensional structures (the same matrices), and the operations basically consist of multiplication and addition.

The first characteristic makes **matrix multiplication** specially appropriate for parallel machines called multiprocessors, where a set of processors, or processing elements, share the same memory.

Parallel algorithms for multiprocessors often follow the basic lines of decomposition or division.

In general, all of them have a previous static or dynamic period for partitioning or dividing data quantity (or parts of the multiplication result matrix) to be computed in each processor to compute the final result.

The last two characteristics make the proposed algorithms for parallel **matrix multiplication** follow SPMD (Single Program - Multiple Data - multiple autonomous processors simultaneously execute the same program at independent points) parallel computing model in general. In this way, a same program is executed asynchronously in each processor of the parallel machine and it is eventually synchronized and/or communicated to the other processors. It is worth to

mention that SPMD is independent of whether the implementation is carried out on a multiprocessor or multicomputer parallel machine, or on a parallel computer with processing architecture distributed as a workstation network.

In general, it is really difficult to find in appropriate parallel algorithms for a certain type of parallel computing architecture. Even though algorithms can be adapted in a more or less complex way to each of the available parallel processing architectures, it is also true that there exists an adaptation and implementation cost at an algorithmic level. And even more important is the fact that, in the context of applications with great computing requirements, the cost in terms of throughput obtained may be too high. Then, in most of the cases, a close relation between each algorithm and a particular parallel computing architecture can be found.

*D. **Parallel algorithm** implementation*

Main aim of this paper is to make something similar to what Microsoft provides in their parallel extensions library and make it easy to use. There will be implemented a variant of *Parallel.For()* which is provided by Microsoft. The only differences is just a single definition of the method, which accepts start and stop indexes of the for-loop and the loop's body as a delegate.

The only difference between this code and Microsoft code is just the namespace name where the Parallel class is defined :

Microsoft's solution : *System.Threading.Parallel.For(0, s, delegate(int i)*

this implementation : *AForge.Parallel.For(0, s, delegate(int i).*

E. Implementation details

The main idea of this implementation is concentrated in three main routines – initialization, job scheduling and execution.

The implementation of parallelism that is presented in this paper is based on regular classes from *System.Threading* namespace which are available in any version of .NET- *Thread*, *AutoResetEvent* and *ManualResetEvent.* Name of **parallel algorithm** presented in this paper is *Aforge*. The amount of threads that were created to parallel for-loops is equal to the amount of cores in the system. The system that was used for this measurement is Intel core i5-2400 and it has four cores so there were four threads used. First step of this implementation is to initialize Parallel class's instance creating required number of threads and synchronization objects. All implementation details are available on this link *Implementation details*.[5]

III. PERFORMANCE ANALYSIS

There was used a simple technique to measure performances, running routines many times in a loop, checking how much time is spent for it.

Processor Number	i5-2400
#of Cores	4
#if Threads	4
Clock Speed	3.1 GHz
Max Turbo Frequency	3.4 GHz
Intel Smart Cache	6 MB
Bus/Core Ratio	31
DMI	5 GT/s
Instruction Set	64-bit
Embedded Option Available	Yes

Fig. 2. *Computer performance which is used for measuring speed of execution **matrix multiplication**.*

Measurements below were done on this processor and they gave these results:

Matrix size	10
1. C#(ms)	20.000
2. C#(ms)	10.000
3. C#(ms)	20.000
4. C#(ms)	20.000
5. C#(ms)	20.000
AVG time in C# (ms)	18.000
1. Aforge (ms)	20.000
2.Aforge(ms)	10.000

3. Aforge (ms)	10.000
4. Aforge (ms)	10.000
5. Aforge (ms)	10.000
AVG time in Aforge(ms)	12.000
Runs	1000
Number of threads	4

Fig. 3 *Performance measurement for matrix size 10*

Matrix size	1000
1. C#(ms)	206712.290
2. C#(ms)	204761.159
3. C#(ms)	205358.761
4. C#(ms)	209111.509
5. C#(ms)	211316.298
AVG time in C# (ms)	207452.003
1. Aforge (ms)	59540.083
2. Aforge(ms)	58226.898
3. Aforge (ms)	58590.490
4. Aforge (ms)	60481.085
5. Aforge (ms)	58923.326
AVG time in Aforge(ms)	59152.376
Runs	10
Number of threads	4

Fig. 4 *Performance measurement for matrix size 1000*

Results for matrix size 10 and 1000 presents how much time is spent for **matrix multiplication** when there is a 10x10 and a 1000x1000 matrix. As it presents in tables above, there is less time needed for small **matrix multiplication** without any parallelism and there is less time spent for big **matrix multiplication** when the parallelism is used

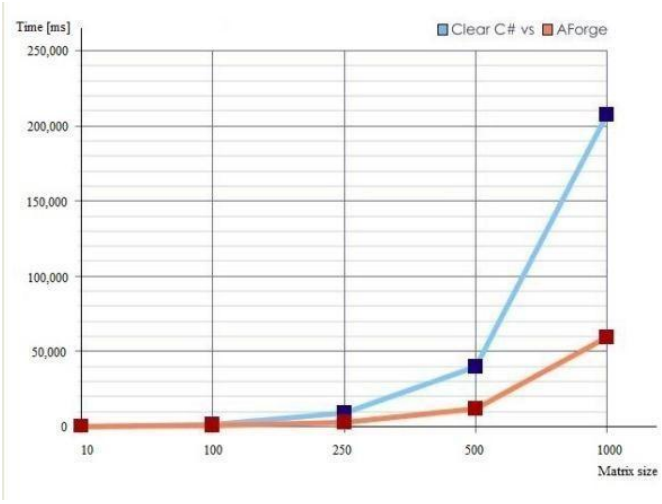


Fig. 5 *Comparison of Speed execution between sequential algorithm Clear c# and **parallel algorithm** Aforge when there are matrix of different sizes used*

As it can be seen from the above result, there is less time spent to multiply small matrices without any parallelism at all. Paralleling such computations just leads to the fact that all additional routines for threads synchronization take more time than the actual useful work.

IV. Conclusion

In this paper a method for solving parallel **matrix multiplication** problem was presented. The current implementation, by default, creates the amount of threads which is equal to the number of cores in the system. As it was shown, it is faster to multiply small matrices without any parallelism at all. But in case of bigger matrices using parallelism takes less time. Some tuning of the existing implementation may be done by investigating the optimal amount of background threads to use for parallelism.

References

[1] http://en.wikipedia.org/wiki/Matrix_multiplication

[2] D.C. Chen and R. H. Price. "A Real-Time TMS320C40-Based Parallel System for High Rate Digital Signal Processing." ICASSP91 Proceedings, May 1991.J.

[3] Ziad A.A. Alqadi, Musbah Aqel, and Ibrahiem M. M. El Emary, Performance Analysis and Evaluation of Parallel Matrix Multiplication Algorith ms, Faculty of Engineering, Al Ahliyya Amman University Amman, Jordan

[4] Anshul Gupta and Vipin Kumar, Scalability of Parallel Algorithms for Matrix Multiplication, TR 91-54, November 1991 (Revised April 1994)

[5] http://www.codeproject.com/Articles/29813/Parallel-Computations-in-C Source code available on this link

.

Comparison of a Sequential and a MapReduce Approach to Joining Large Datasets

Marko Lalić, Emina Memić, Faruk Kešan, Edita Gondžić, Nermin Smajić, Novica Nosović

Faculty of Electrical Engineering, University of Sarajevo/Department for Computing and Informatics,
Sarajevo, Bosnia and Herzegovina
mlalic1@etf.unsa.ba

***Abstract*—MapReduce as a programming model is considered one of the biggest improvements in massive data processing which utilizes parallelization. The increasing amount of data being processed and stored has caused a need to investigate more efficient solutions to common problems, one of which is performing a join operation on two interconnected datasets. In this paper, a classic sequential solution to this problem is compared with a MapReduce approach, with the intent of discovering the relative advantages of the two. The sequential application runtime for datasets of negligible sizes in today's terms is proven prohibitively slow. Furthermore, a MapReduce cluster of five Amazon EC2 nodes is shown to process, in the same time period, ten times larger data than the sequential application.**

Keywords—Hadoop, MapReduce, join, distributed join, cluster

I. Introduction

MapReduce [1] is applicable in a wide spectrum of scenarios which pertain to processing very large data. Using MapReduce for such operations consists of implementing two functions: map and reduce. The map function transforms the input into a set of (key, value) pairs. After that, the reduce function receives the set of all values with the same key and emits the final result based on aggregating these values.

Recently, information explosion has become a well known phenomenon which causes datasets upon which applications operate to grow to an extremely large scale. Examples of such datasets may be found in social networking applications where millions of users produce terabytes of data each day, e.g. Facebook stores at least 60 TB of new data daily [2]. This kind of system requires a frequent join operation to match the users with their content: comments, status updates, etc.

In this paper, for the purpose of comparing a sequential solution to the join problem with a distributed MapReduce based approach, a large set of data was generated. The data is made of two sets - users and comments - linked by the user's ID. The data was fed to both a sequential implementation of a join algorithm running on a single machine and a parallel one, using MapReduce, running on a cluster of varying size. The results will clearly show the power of the new programming model.

Every MapReduce implementation is highly scalable, but it is always hard to know when an increased number of nodes is essential for improved performance, as opposed to causing an unnecessary cost increase. The results of running the application on clusters of different sizes, particularly on clusters which consists of one, three, five and eight nodes, are presented.

II. Infrastructure

MapReduce is a programming model for concurrent processing of large datasets, based on the concepts of map and reduce functions, stemming from functional programming [1]. There are a number of frameworks implementing the MapReduce model, including Google's proprietary implementation and Apache's open source Hadoop. For the purpose of running a MapReduce task for this paper, the Hadoop framework was used. [3]

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides a cloud based compute platform [7]. In this paper, it was used to run the Hadoop framework on a cluster of nodes. Furthermore, a single micro instance was used for running the sequential implementation. This way, the measurement results are comparable since they were obtained on homogenous hardware and the same execution environment.

III. Used algorithms

A short description of the algorithms used to implement the sequential join operation and the MapReduce version follows.

A. Sequential Join

The sequential join operation implements the well known hash join algorithm [4]. In its simplest form, it first reads the smaller of the two tables upon which it operates to an in-memory hash table, where the hash value is calculated based on the join attribute. After this, it iterates through the larger table calculating the hash value for each row's join attribute and, in turn, using it to retrieve rows from the hash table, thereby achieving the join.

In the case which this paper discusses, the users table is always small enough to be stored to the in-memory hash table, so there is no need to use a more sophisticated scheme for partitioning the input set.

B. *MapReduce Join*

The MapReduce implementation of the join operation uses an approach called repartition join [5]. The map function can receive and process a row from either of the two tables. It emits a (key, value) pair where the key is the value of the join attribute and the value is the record itself, additionally tagged with an identifier indicating which table it belongs to.

The reduce function will receive all records from both tables which have the same value for the join attribute, thanks to the MapReduce framework. It then splits the records into two sets based on the tag value and performs a cartesian product on the two sets, thus resulting in a new set of joined records.

IV. Measurement scenarios

In order to compare the MapReduce approach with the sequential solution to the join problem, two groups of measurements were used.

A. *First class of scenarios*

The first group contains four scenarios which measured the performance of MapReduce on a varying number of nodes in a cluster. Each scenario used the same dataset, previously generated to contain a hundred million comments and two million users, resulting in a dataset input size of 14,000 MB. In the first scenario the job was run on a cluster containing one node. In the second scenario the cluster consisted of three nodes, whereas in the third and fourth scenario the cluster was made of five and eight nodes, respectively.

B. *Second class of scenarios*

The second group of scenarios measured the performance of the sequential algorithm running on a single machine. The size of the input data was gradually increased by 70 MB for each scenario in this group, starting from 350 MB up to 980 MB, resulting in nine different cases.

V. Measurements and results presentation

Figure 1 shows how much time is needed to execute a job, depending on the number of nodes. The same job was executed four times, on 1, 3, 5 and 8 nodes respectively, while data size remained the same (14,000 MB).

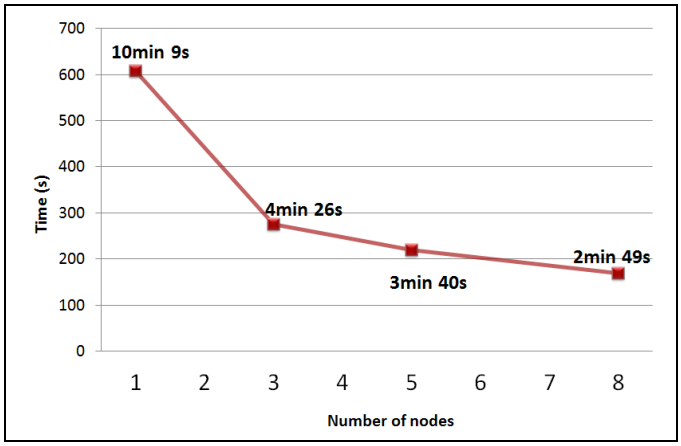


Figure 1. Job execution on four different clusters for 14,000 MB of input data

Figure 2 consists of two curves. The first one shows how execution time depends on data size, while the second represents linear trendline. The job was executed sequentially for different data sets from 350MB to 980MB.

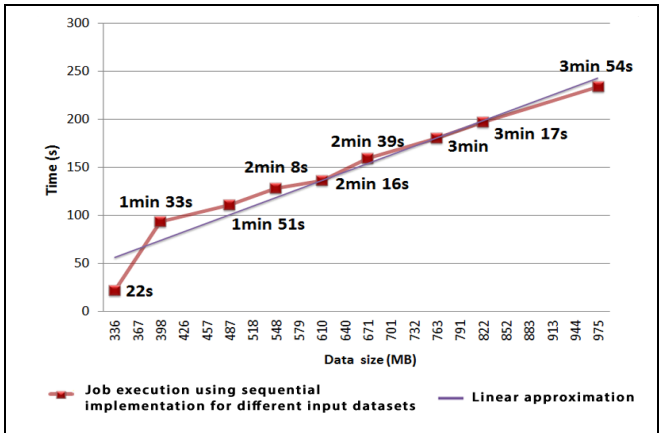


Figure 2. Job execution using sequential implementation for different input datasets

I

Figure 3 combines the information shown in Figures 1 and 2 and compares the execution times after the problem was solved both sequentially as well as using clusters with different numbers of nodes running Hadoop MapReduce.

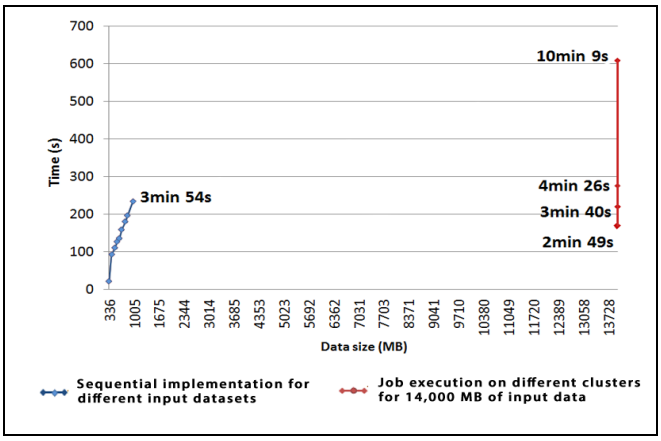


Figure 3. Comparison between sequential implementation and implementation on clusters of a different number of nodes

VI. Analysis

As noted previously in the paper, the problem was first solved using a sequential algorithm running on a single machine and then on clusters of a different number of nodes in the cloud.

The number of nodes was one, three, five and eight. Each execution resulted in a different execution time. From the previous diagrams it can be seen that by increasing the number of nodes, the execution time decreases. The reduction of the execution time between the case with eight nodes and the one with five nodes is smaller than that between the cases with five and three nodes. Furthermore, the largest decrease is noticed when going from a single node to three nodes in the cluster. (Figure 1)

In other words, one can conclude that, at a larger node count, the performance is only slightly changed by moving to a cluster with even more nodes. This is due to the fact that the execution time is dependent on, not only the number of nodes processing the data, but also on the speed of access to the data and the overhead of partitioning the input set, which is constant, regardless of the number of processing nodes, due to the fact that it is performed by the single master node [6]. Thus, increasing the number of nodes becomes an unnecessary waste of resource allocation.

Using Amdahl's law it is calculated that the parallel proportion of the program was 84.4%, 79.8% and 82.5% for the scenarios with 3, 5 and 8 nodes when compared to the single node execution. This confirms that MapReduce splits the input data in a convenient manner for parallel processing, having only 15% of non-parallel work in the form of job preparation overhead.

When it comes to the sequential algorithm, its execution time increases linearly as the size of the input data is increased. (Figure 2). In addition, already at an input size of 980 MB, the execution time has exceeded that of the MapReduce version on a cluster of five nodes where the input size is 14,000 MB. (Figure 3).

The reason the performance of the MapReduce job running on a single node is better than that of the sequential algorithm is that, even though there is only one node, multiple map and reduce processes are run in parallel on it.

From this analysis, it can be established that, as the amount of data to be processed is increased, a sequential approach becomes unusable and impractical. Using a cluster of nodes for execution of the map and reduce functions, instead of using sequential execution, becomes a much better solution, especially when the speed of retrieving the results is crucial.

VII. Conclusion

This paper shows the advantages of using parallelism, realised with Amazon Elastic MapReduce, over a sequential implementation for large datasets.

After the job was executed sequentially, it was also executed on four clusters of one, three, five and eight Amazon EC2 virtual nodes. As the number of nodes was increased, it took less time to finish the job. However, going from one node to three nodes in the cluster gave a larger increase in performance than going from five to eight nodes, since the execution time also depends on speed of access to the data and the overhead of partitioning the input set. This leads to a conclusion that a further increase of nodes doesn't improve performance greatly.

Also, the execution of the job on a single node showed better performance than the execution of the sequential algorithm, because multiple processes can be run in parallel on the single node in the cluster.

Furthermore, for input datasets whose size exceeds 1 GB, the execution times of the sequential algorithm implementation become prohibitively slow, especially compared to the performance of a Hadoop MapReduce cluster.

Acknowledgment

This work is supported in part by the software development company AtlantBH based in Sarajevo, Bosnia and Herzegovina.

References

[1] Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. Google, Inc.

[2] Ashish Thuso, Zheng Shao, Suresh Anthony, Dhruba Borthakur, Namit Jain, Joydeep Sen Sarma, Raghotham Murthy, Hao Liu. Data Warehousing and Analytics Infrastructure at Facebook. Facebook, Inc.

[3] Puneet Singh. Hadoop - Map Reduce. Indiana University.

[4] Hansjiirg Zeller, Jim Gray. An Adaptive Hash Join Algorithm for Multiuser Environments. Tandem.

[5] Spyros Blanas, Jignesh M. Patel, Vuk Ercegovac, Jun Rao, Eugene J. Shekita, Yuanyuan Tian. A Comparison of Join Algorithms for Log Processing in MapReduce. Computer Sciences Department University of Wisconsin-Madison, IBM Almaden Research Center.

[6] http://aws.amazon.com/ec2/

[7] http://hadoop.apache.org/docs/

Acceleration of DCT transformation in JPEG image conversion

Nikolina Frid*, Hrvoje Mlinaric** and Josip Knezovic***
Faculty of Electrical Engineering and Computing / Dept. of Control and Computer Engineering, Zagreb, Croatia
*nikolina.frid@fer.hr, **hrvoje.mlinaric@fer.hr, **josip.knezovic@fer.hr

Abstract - This paper explores the alternatives to a standard software implementation of DCT transformation algorithm, an IEEE standard 1180 definition, and compares the obtained results on FPGA development boards such as Spartan-3E and Virtex-5 with 32-bit MicroBlaze™ soft-core processor. Three cases of implementation are presented: a software implementation of AAN algorithm, a special FPGA ip-core that accelerates the standard DCT algorithm and a special FPGA ip-core that implements the AAN algorithm. Both of the ip-cores work as coprocessor to a standard MicroBlaze processor.

I. INTRODUCTION

JPEG image format is one of the most commonly used image formats because of the high level of compression it provides and the fact that almost every web browser has a support for it. The algorithm that is based on CCITT Rec. T.81 (1992 E) [1,2,3] poses no computational challenge when it is to be implemented on middle to large-scaled computer system but when considering small built-in computers, the traditional way of implementation has a very negative impact on performance. The major setback in this algorithm is the DCT frequency transformation of image data, the calculation of which is extremely CPU demanding and in case of low frequency cores, which are found in built-in systems, can take too much time to be carried out. Hence, new and faster methods for transformation are needed. Acceleration can be achieved by introducing new software algorithms [5,6] for DCT transformation but also by a combination of hardware implementation of computationally demanding DCT transformation and software implementation of other parts which result in a faster but still a very flexible system.

II. JPEG COMPRESSION STANDARD

A. About the Standard

JPEG is a still-image compression standard, it supports a broad range of image quality level and is applicable to almost any type of digital image. The standard has two compression methods. First one is called *Baseline method and* is based on DCT. It produces an image of lesser quality than the original but the limitations of human visual system make this invisible to naked eye.

The compression algorithm has 7 consequential steps. Firstly, conversion of the pixel color from RGB model to YUV model takes place. Secondly, every pixel is shifted by -128 so that the whole range of DCT is used. Then the image is divided into 8x8 pixel blocks which are DCT transformed. The transformation broadens the pixel data range to 12 bits, and to scale back to 8 bits per pixel quantization is performed. Next data is arranged in a zig-zag manner so that the coefficients with higher value come first which is important for the efficiency of the following step – the Huffman entropy coding. Finally, coded data is formatted and written to a .jpeg file according to Standard. This is shown on Figure1.

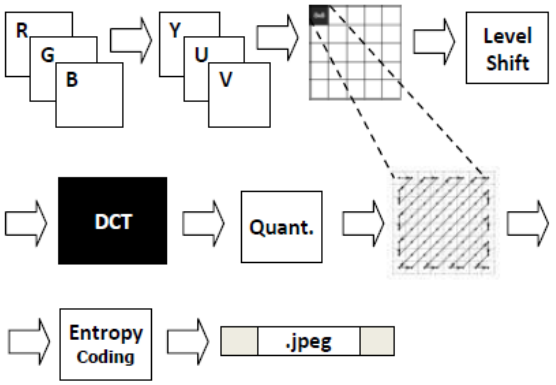


Figure 1. JPEG compression flow diagram

B. DCT transformation

DCT transformation performs conversion of incoming data to frequency domain the goal of which is to acquire the frequency characteristics of every image component (Y, U and V image components). As mentioned earlier, it is performed on 8x8 pixel block. The standard formula for calculation is (1) with extensions (2) to (4)

$$F(u,v) = \frac{1}{4}C(u)C(v)G(u,v) \tag{1}$$

$$G(u,v) = \sum_{i=0}^{7}\sum_{j=0}^{7} f(i,j)\cos\left(\frac{(2i+1)u\pi}{16}\right)\cos\left(\frac{(2j+1)v\pi}{16}\right) \tag{2}$$

$$C(u) = C(v) = (1/2)^{1/2}, \quad \text{for } u, v = 0 \tag{3}$$

$$C(u) = C(v) = 1, \quad \text{for } u = 1,2,...M-1, v = 1,2,...N-1, \tag{4}$$

where *F(u,v)* denotes the transformation coefficient, *f (i,j)* denotes the amplitude of image element in a block, *u* and *v* are coordinates in area of spatial frequencies, *i* and *j* are coordinates in area of image elements.

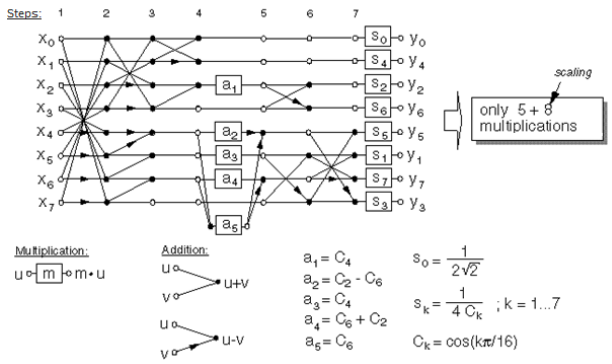


Figure 2. AAN algorithm diagram

One can easily notice that (1) requires 64x64 passages through the pixel block and its core (2) includes 3 multiplications per each step. The cosine function can be simplified by storing pre-calculated values in a lookup table but still, this way of performing DCT takes up more time than all other parts of the JPEG compression together.

Through exploration of the specific properties of DCT, special algorithms were devised that help achieve significant speedup. In this paper the AAN [4] algorithm for scaled 1D DCT (8 points) was chosen for implementation and testing. Single pass of this algorithm requires 5 multiplication, 29 additions and 16 2nd complements. Two passes are needed for transformation of one block of pixels.

Figure 2 shows the flow of the AAN algorithm for a single pass (7 steps marked on the diagram). Details about the algorithm itself can be found in [2].

III. IMPLEMENTATION OF DCT TRANSFORMATION

A. *Software implementation of DCT tranformation*

The standard way of implementing JPEG compression method is to write a piece of software [5,6,7,8,9] that implements the entire procedure. This implies the direct implementation of (1) in the source code. As discussed earlier, this is possible only on middle to large – scaled computer systems.

On small computer systems with modest CPU performance, such as MicroBlaze on Xilinx FPGA boards considered in this paper, the standard way is not an option because of the time the DCT would take to perform. The easiest way to overcome this issue is to replace the standard way of calculating DCT by a much faster algorithm such as the AAN algorithm, although several other fast algorithms exist and it would be interesting to make a comparison of all of them along with the AAN algorithm if the future work.

By following the flow diagram shown in Figure 2, source code of this new way of DCT calculation can be easily written. One should have in mind that this algorithm is 1D, which means that it needs to be performed on each 8x8 block of pixels: first passing the rows (one row per pass) and then the columns (one column per pass), which is 16 passes in total.

Significant speedup can be achieved this way, which will be explained in details in the next chapter, but it does not exploit all the benefits of this algorithm as hardware implementation of the same does. The main benefit of this approach is that it does not require any additional circuitry to be added to the existing computer system or any other modifications to be carried out. If this is the case, this is the only solution for speedup possible.

B. *Hardware implementation of DCT tranformation*

In case when it is possible to add special components to the system beside the standard ones like CPU, memory and peripheral controllers, a good idea is to create a dedicated component that acts like a coprocessor and its only task is to sequentially receive blocks of pixels, perform DCT on them and send them back to main processor. The other parts of the JPEG compression algorithm should remain implemented in software as an application which runs on system preserving the flexibility.

In research done for this paper, two Xilinx FPGA development boards were taken into consideration: Spartan – 3E and Virtex5 – University Board. Since any computer system based on FPGA technology consists of soft-core design which is downloaded onto the board and programmed into the reprogrammable FPGA field, circuitry – wise, it is built from scratch every time and it is no problem to create any additional cores needed.

By following the standard framework provided by Xilinx for creating soft-core systems, a new ip core which calculates DCT was created and connected to MicroBlaze, the 32-bit CPU provided by the framework along with its standard peripherals, through FSL interface. Every FSL interface is bidirectional and consists of two separate connections, one for each direction, and two roles on each connection: master – who always sends data, and slave who receives. The component block diagram of the developed test system is shown in Figure 3.

The coprocessor works as a simple synchronous (clock triggered) state machine with 5 states altogether as shown in Figure 4. After system-reset it is first initialized and then goes to a state where it waits for the first word of

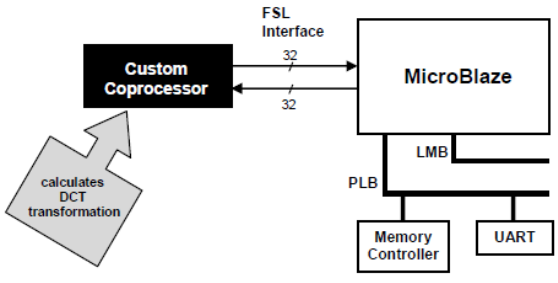


Figure 3. Component block diagram

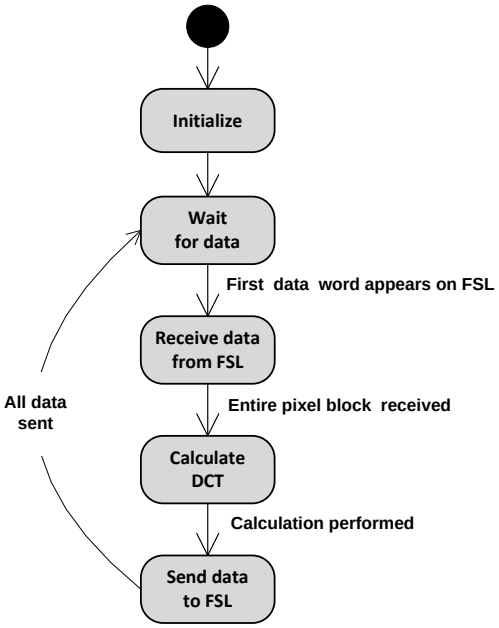


Figure 4. Coprocessor state diagram

data to appear on FSL bus port. Since the bus is 32-bit by default at first it appears to be a good idea to send 4 pixels through bus at once in order to get additional speedup, but almost as much time or even a little more takes for MicroBlaze to paste 4 pixels together (done in the source code of application). Then the DCT calculation is performed on the block of pixel, the implementation of which can vary and is described later in this chapter. Having the transformation finished the core sends the data back to MicroBlaze goes back to the wait state once again and stays there until new data is ready.

1) Calculation of AAN algorithm

To get all the benefits from the structure of the AAN algorithm one should make a hardware implementation of this algorithm. It is by far the fastest method for calculation of DCT. Of course, the algorithm needs to be adapted to achieve a balance of FPGA resource used and calculation duration.

As discussed earlier, this algorithm should have two passes (firstly over rows and then over columns), each pass consisting of 8 points i.e. 8 steps (one row/column is on step). Furthermore, it is impossible to perform the whole step in one clock tick so the step is divided into 7 sub steps (indicated on Figure 2), each step performed on a single clock tick. This way, a good balance between FPGA resource usage and timing performance is achieved. Nevertheless, this method of calculating DCT is still very FPGA resource intensive because it requires a 64x12bit field to hold the input/output data, 8x32bit registers to hold in-between-sub-step data, 13 constants and several other auxiliary registers depending on the exact implementation in VHDL. Also, one must notice that auxiliary registers are accessed all at once which is also very demanding for mapping and routing in FPGA chip. Thus, this method cannot be implemented on smaller chips such as Spartan-3E but only on those with much more resource at hand like Virtex-5.

More details on resource usage and performance results is provided in the next chapter.

2) Calculation of standard formula

The alternative to hardware implementation of AAN algorithm is adjusting (1) so that it can be implemented in hardware. This is a slower method of calculation but is much less hardware intensive and can be implemented even on resource-limited chips such as Spartan-3E.

First thing that needs to be done is to pre-calculate values of cosine function and store them in a lookup table directly in VHDL code. Secondly, (1) needs to be divided in two parts the second one being (2). Finally, the algorithm of calculation is conducted as follows: on each clock tick one multiplication is performed; if the summation part is done then the rest of multiplication with predefined constants is performed. This process is performed 64 times: once for each new pixel value in the pixel field.

This method requires only a 64x12bit field to hold the input/output data and predefined constants (lookup table included). On each clock tick only one data register is accessed so all of the described features make this method much less FPGA resource demanding and possible to implement on a smaller board like Spartan-3E. Detailed discussion on resource usage and performance results is conducted in the next chapter.

IV. PERFORMANCE COMPARISON

All three methods for calculation DCT described earlier in the paper were implemented using Xilinx XPS IDE. They underwent the process of bit stream generation and were downloaded to the evaluation boards.

TABLE I. PERFORMANCE COMPARISON

Method of implementation		**Target FPGA Board**	**Performance**		
			Duration per one pixel block	***Resource demand***	
				Flip-Flops used	***LUTs used***
Software	Standard formula	Spartan-3E/Virtex-5	Not feasible		
	AAN algorithm	Spartan-3E/Virtex-5*	95 µs / 40 µs	-	-
Hardware	Standard formula	Spartan-3E	82 µs	2811	5717
	AAN algorithm	Virtex-5	1.12 µs	8209	19053

* MicroBlaze frequency on Spartan-3E is 50MHz and on Virtex5 100MHz

Results obtained from these measurements along with data about resource usage obtained from XPS synthesis report are shown in Table 1. The target board is indicated because not every method of implementation is available on both target boards.

V. CONLUSION

There are several methods for implementation of DCT transformation over a block of pixels. On middle to large-scaled computer systems, with sufficient CPU capabilities, any alternative to the standard method is not worthwhile. As for small built-in computers, an alternative is a must. The best option is to create a dedicated core, whose only task is to calculate the DCT transformation, keeping the other parts of the entire JPEG compression algorithm as they are. In this paper, two implementations were presented: based on the standard algorithm and based on AAN algorithm. They both have upsides and downsides: The AAN algorithm takes more advantage of hardware implementation and is undoubtedly the fastest, but it requires a great amount of FPGA resource and its implementation on smaller FPGAs like Spartan - 3E is not feasible. In those cases the hardware implementation of the standard method, which needs less FPGA resource, is a good option. Finally, if having a dedicated core is not an option, it is possible to implement the AAN algorithm as software but this is the slowest method of the three.

REFERENCES

[1] Gregory K. Wallace, “The JPEG Still Picture Compression Standard”Communication of the ACM, vol 34., Issue 4, ACM Press, New York, NY, USA, pp.30-44 , 1991.

[2] Leonardo Chiarligione, “Impact of MPEG Standards on Multimedia Industry”, Proc of IEEE, June 1998, Bol. 86, No. 6, pp 1222-1227, 1998

[3] Joan L. Mitchell, William B. Penebaker, Chad E. Fogg, Didier J. LeGall, “MPEG Video Compresion Standard”, Champ & Hall, New York, USA 1996

[4] Arai Y., Agui T., Nakajima M., “A Fas DCT-sQ Scheme for Images”, Trans. IEICE, Vol. E71, No.11 pp. 1095-1097, 1988

[5] M. Kovac, M. Zagar and N. Ranganathan, ”Design and Optimization of a VLSI Architecture for Discrete Cosine Transform Used in Image Compression”, Journal of Computing and Information Technology, 3(4), Sept. 1996, Pp. 159-170.

[6] M. Kovac and N. Ranganathan, ”JAGUAR: A VLSI Architecture for JPEG Image Compression Standard,” Proceedings of IEEE, Special Issue on Image and Video Compression, 83(2), Feb 1995;

[7] Mlinarić, Hrvoje; Kovač, Mario; Knezović, Josip. “Application of dynamically reconfigurable processors in digital signal processing”, International Conference on Signal Processing and Multimedia Applications - SIGMAP 2006 Portugal, INSTICC Press, 2006. 343-346

[8] Duracic, Kresimir; Mlinaric, Hrvoje; Kovac, Mario. “Optimization methods for MPEG-4 Algoritgm in multimedia application running PDA”, PROCEEDINGS EC-VIP-MC 2003 Zagreb : FER, 2003. 305-310

[9] Mlinaric, Hrvoje; Kovač, Mario; Đuračić, Krešimir. Implementation of MPEG audio decoder on TMS320C3x DSP processor // Proceedings of the 2nd International Workshop on Design, Test and Applications - WDTA 99 Dubrovnik : FER, 1999. 83

A preliminary study of Croatian Language Syllable Networks

K. Ban, I. Ivakić and A. Meštrović*
* Department of Informatics University of Rijeka, Rijeka, Croatia
kristina.ban89@gmail.com, ivan.ivakic@net.hr, amestrovic@inf.uniri.hr

Abstract - This paper presents preliminary results of Croatian syllable networks analysis. We analyzed networks of syllables generated from texts collected from the Croatian Wikipedia and Blogs. Different syllable networks are constructed in a way that each node in this network is a syllable, and links are established between two syllables if they appear together in the same word (co-occurrence network) or if they appear as neighbours in a word (neighbour network). As a main tool we use network analysis methods which provide mechanisms that can reveal new patterns in a complex language structure. We aim to show that syllable networks differ from Erdös-Renyi random networks, which may indicate that language has its own rules and self-organization structure. Furthermore, our results have been compared with other studies on syllable network of Portuguese and Chinese. The results indicate that Croatian Syllables networks exhibit certain properties of a small world networks.

I. INTRODUCTION

Network analysis have become significant method in different research areas such as biology, computer science, economics, sociology, medicine and linguistics. Complex networks are a class of networks that exhibit specific topological features, such as high clustering coefficients, small diameters, power-law degree distribution, community structures, one or several giant components, hierarchical structures, etc. Two important classes of complex networks that can be further differentiated are small-world networks [1, 2, 3] with high clustering as a main property and scale-free networks [4, 5] which can be characterized by power-law degree distribution.

Language can be viewed as a complex network if it is presented as a system of interacting units. Network analysis provides mechanisms that can reveal new patterns in a complex structure and can thus be applied to the study of the patterns in language structures. This, in turn may contribute to a better understanding of the organization, structure and evolution of a specific language.

Network properties of written human languages have already been analyzed in different research studies [6]. Networks based on co-occurrence of words in sentences are analyzed in [7, 8, 9]. The topology of human written language, through a network representation of Orwell's 1984 is presented in [10], while the co-occurrence properties of words in different languages are studied in [11]. All these studies have shown that language networks exhibit certain properties indicative of small-world networks, e. g. Pemble and Bingol [12] have constructed two complex networks out of Wikipedia English and German corpora and analyze conceptual networks in different languages.

So far, syllable networks have been constructed exclusively for Portuguese [13] and Chinese [14]. Both networks have a large clustering coefficient and power-law degree distribution, as opposed to the Erdös-Renyi (ER) random networks [15], which have low clustering coefficient and Poisson-like degree distribution. In [13] the syllable network is used to demonstrate that language in itself resembles a living organism, evolving in time and space. The growth of the syllable network was analysed and it was concluded that such growth can be compared to the growth of a child's vocabulary: the most frequently occurring sounds have a higher probability to be accepted.

Syllables are important phonetic elements in language evolution. Therefore, a study of the syllables that make up a language is crucial in determining the phonetic structure of a language, as well as providing necessary grounds for further linguistic research [16]. As far as the structure of the Croatian language is concerned, a parallelism between the sound structure of the language and its syllabic hyphenation exists. The issues presented here are further accounted for and expanded on in the domain of natural language processing. The most common type of syllable-based text processing is evident in speech synthesis approaches [17], while a statistical analysis of syllable networks may be important for corpora building.

In this paper we describe our experiments with syllable networks designed for the Croatian language. We propose a number of possible models in syllable network construction and detail a network construction process in the second section. In the third section we denote main network properties such as diameter, average path length and clustering coefficient, followed by a comparison of the Croatian syllable networks with the Chinese and Portuguese syllable networks. In conclusion, we elaborate on the obtained data and provide concluding remarks in the fifth section.

II. NETWORKS CONSTRUCTION

A. Syllable networks construction strategies

Different strategies can be applied in building syllable networks from text. The idea of a syllable network is to represent syllables as nodes and establish links between them according to their connections within words. Generally speaking, a syllable network can be either undirected or directed and unweighted or weighted. In a directed syllable network, a directed link indicates the direction of the connection; displaying which syllable

(node) is the initial and which syllable (node) is the target. By using a directed network, the successor or the predecessor of an intended syllable can be seen, possibly providing the grounds for further statistical analysis of language structure on the phonetic level. Weighted syllable networks contain information about the number of established links between two syllables, which is again significant in phonetic structure analysis.

A question of how to establish the links between the nodes (syllables) must be discussed. One way is to connect the syllables that belong to the same word (syllable co-occurrence network) and another way is to connect only the neighbour syllables (first-neighbour network). This results in eight different syllable network models (Table I).

TABLE I. SYLLABLE NETWORK MODELS

	co-occurrence network		first neighbour network	
	unweighted	weighted	unweighted	weighted
undirected	undirected unweighted co-occurrence network	undirected weighted co-occurrence network	undirected unweighted neighbour network	undirected weighted neighbour network
directed	directed unweighted co-occurrence network	directed weighted co-occurrence network	directed unweighted neighbour network	directed weighted neighbour network

In [13, 14], the network is constructed in a way that two nodes (syllables) are connected if they belong to the same word, making the network undirected and unweighted. This simplified model of a syllable network is constructed in order to study the evolution of the language using phonetic elements [13]. We constructed three networks according to this model.

In our opinion, for some proposals that include natural language processing and linguistic studies, it also makes sense to construct a directed and weighted syllable network of neighbour syllables. Therefore we additionally constructed one directed and weighted neighbour network.

B. Data

We analysed different networks of syllables from different text corpora. The texts used for building the networks are two large corpora. The first corpus is the Croatian Wikipedia. The second corpus contains 3.218 articles collected from different Croatian blogs (including 4 religious and 5 political portals, 6 blog spaces, 3 web-pages with comments and 4 columns from the daily newspapers).

The reason why we have chosen these corpora is because our future work is focused on the text collected from the Web. Another possible approach is to choose a dictionary of Croatian language as a network source. But in [13] it is shown that there is no big difference between syllable networks constructed from the book and syllable network constructed from the dictionary. A problem we encountered was the Wikipedia corpus containing a certain number of foreign words. This is the reason why the initial network had certain syllables unusual for the Croatian language. Therefore, we examined a filtered network from which all nodes with small degree (meaning that they contain some rare and unusual syllables) were excluded. There is a linguistic difference between the two corpora. The Wikipedia corpus is more formal, so there are more standard words with a pattern in writing. On the other hand, the blog corpus is mostly written in an informal manner, with the use of dialect, slang or abbreviations. However, all of the mentioned texts specifics collected from the web are essential for our future work.

C. Syllable networks

We constructed four different networks. Three of them were designed as word co-occurrence syllable networks: the first from the Wikipedia text - C_W, the second from the blog text – C_B, and the third was devised from both corpora - C_A. The fourth network was constructed as a directed and weighted first-neighbour syllable network from a single part of the Wikipedia text. The number of nodes and edges for all four networks are displayed in table II.

TABLE II. SYLLABLE NETWORKS CONSTRUCTED

	C_w	C_B	C_A	C_w *-Dir*
Nodes (N)	4284	2000	4067	4438
Links (K)	170248	36202	173660	33341

Network construction is implemented in the Python programming language which contains the NetworkX software package developed for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [18]. For network visualization we used Gephi software [19]. The algorithm described in [16] is implemented and used for the hyphenation process. Both corpora where in *txt* file format which made the reading and processing easy and the only problem was the encoding because of our diacritical signs such as *č,ć,š* etc. The NetworkX module provided us with all the necessary commands to construct a graph and then export it in the desired format. We chose the *gml* format readable in Gephi that supports UTF8 encoding.

The co-occurrence syllable network constructed from texts from Wikipedia (C_w) visualized using Gephi is shown in Figure 1. The most frequent syllables are pointed out.

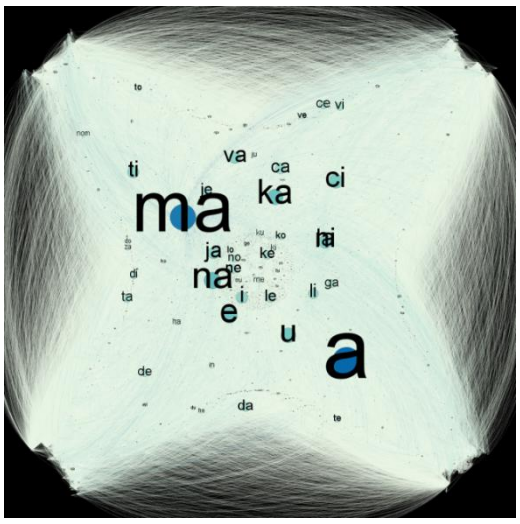


Figure 1. Syllable network from Wikipedia

Another co-occurrence syllable network constructed from blog corpus (C_B) is shown in Figure 2. This is a smaller network with smaller number of nodes, but the most frequent nodes (syllables) are similar to the first network, which is discussed in the fourth section.

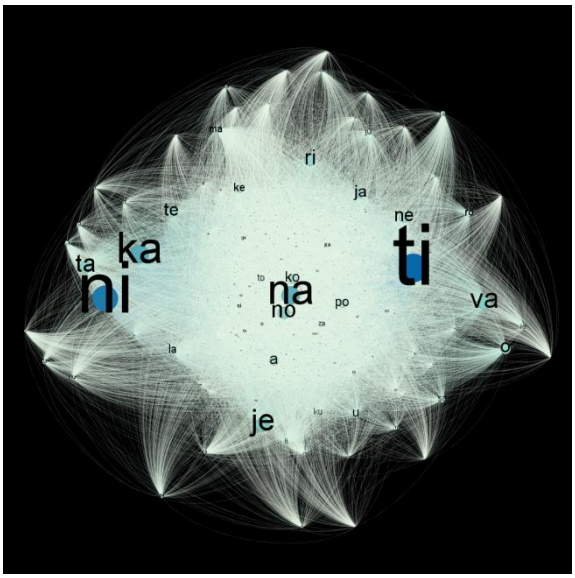


Figure 2. Syllable network from blog corpus

The third network is constructed from both Wikipedia and blog corpora. Syllables with the most connections with other syllables are pointed out and are almost the same as in the first network.

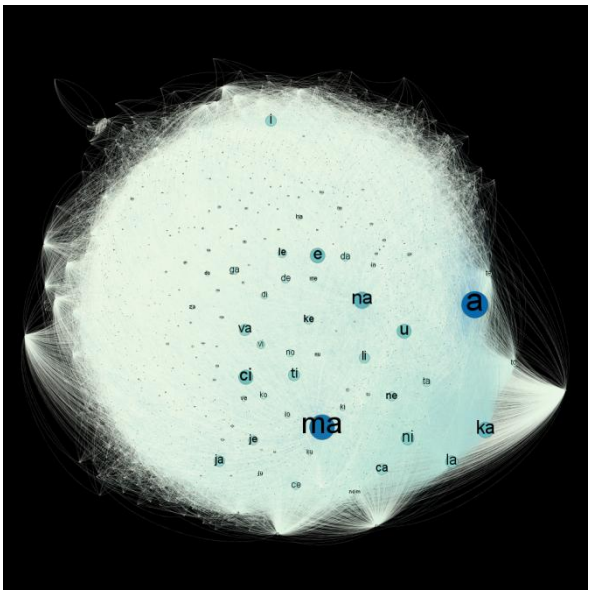


Figure 3. Syllable network from

The fourth network is constructed as a directed and weighted network of first neighbor syllables from words that appear in the texts of the Croatian Wikipedia. The idea was to compare this network to the other three networks and to see if it had potential in phonetic structure analysis.

III. THE NETWORK STRUCTURE ANALYSIS

In this section we explain the most important measures for network analysis. Every network has an N number of nodes and K links.

The degree of a node i is the number of connections of the node and is denoted by k_i. Thus, the average degree of the network is:

$$< k >= \frac{2K}{N}$$

For every two connected nodes i and j the number of links lying on the path between them is denoted as d_{ij}, therefore the average distance of a node i from all other nodes is:

$$d_i = \frac{\sum_j d_{ij}}{N}$$

From where we easily obtain the average path distance L as the average value of di of all nodes:

$$L = \frac{\sum_i d_i}{N}$$

and the maximum distance results in the network diameter, D:

$$D = \max_i d_i.$$

The clustering coefficient is described as a presence of connections between the nearest neighbours of a node. The clustering coefficient C_i of a node i is defined as a ratio between the number of edges E_i that actually exist among these k_i and the total possible number:

$$C_i = \frac{2E_i}{k(k-1)}$$

The average clustering of a network C is the average value of the clustering coefficient of all the nodes:

$$C = \frac{\sum_i C_i}{N}$$

The main property of small-world networks is that the distance between two random nodes grows proportionally to the logarithm of the number of nodes. Therefore, small-world networks tend to have small diameter and short average distance which is the property of random ER networks. Another important property is the high clustering coefficient in comparison to random ER networks. Furthermore, for complex networks it's typically a power-law degree distribution.

IV. RESULTS

In this section the characterization of the constructed syllable networks with the measures commonly used in the study of complex networks such as average degree $<k>$, average path distance L, diameter D, and clustering coefficient C described in the previous section, will be

provided. One of our objectives in this experiment is to see if constructed syllable networks of the Croatian language have properties of small–world networks. Small-world properties have already been proven for syllable networks of the Portuguese and the Chinese language; therefore we expected to find similar results for the Croatian language. For the purpose of comparing constructed networks with random networks, ER networks with the same number of nodes and edges have been constructed and all the important properties have been analysed.

Our original aim was to analyse the network using Pajek software but with a high number of nodes Gephi became a better option. Using Gephi we filtered the networks and determined the average degree <k>, diameter D, average distance L, average clustering coefficient C and some other network values. The correspondent values of these coefficients are shown in table III.

TABLE III. NUMERICAL RESULTS FOR CO-OCCUERNCE SYLLABLE NETWORKS

	C_w	ER_w	C_B	ER_B	C_A	ER_A
N	4284	4284	2000	2000	4067	4067
<k>	39.74	39.74	18.1	18.1	42.7	42.7
D	4	3	4	3	3	3
L	2.151	2.209	2.310	2.489	2.113	2,143
C	0.691	0.017	0.687	0.016	0.690	0.021

The results show that all three co-occurrence syllable networks have a small diameter and average path distance. Furthermore, for all three networks it holds $< k > \ll N$ which shows that syllable networks are sparse as it is expected for complex networks. In comparison to the ER networks with the same number of nodes and edges these networks show a high clustering coefficient: $C(C_w) \approx 40C(ER_w)$, $C(C_B) \approx 42C(ER_B)$, $C(C_{wA}) \approx 33C(ER_{wA})$.

All these results lead to a conclusion that co-occurrence syllable networks of Croatian language exhibit small world network properties.

We compared our results with the results obtained for the Portuguese language and concluded that there is a similarity between these two families of syllable networks. Syllable networks of the Portuguese and of the Croatian language are similar in size, are both sparse and both contain a small diameter and small size of average path length. Both network families have a high clustering coefficient (syllable networks of the Portuguese language have a roughly 50 times larger clustering coefficient than random networks of the same size).

We additionally analysed the fourth, first-neighbour syllable network that is weighted and directed. For the purpose of comparison to the random network, we transformed it into an undirected network. The results are displayed in table IV. Although C for the first-neighbour syllable network was smaller than in the co-occurrence syllable networks, in comparison with the random network, it was still about 30 times larger than random network C.

TABLE IV. NUMERICAL RESULTS FOR THE FIRST-NEIGHBOUR SYLLABLE NETWORK

	C_w -Dir	C_w -Undir	ER
N	4438	4438	4438
K	33341	33341	33341
D	9	8	5
C	0.153	0.208	0.007

These values indicate that the first-neighbour syllable network may be a small-world network as well, however, more experiments with larger corpora need to be conducted.

In these preliminary experiments we didn't estimate the degree distributions for syllable networks. However, we did use NetworkX functions to plot degree distribution on the log-log scale and we got the result shown in Figure 4.

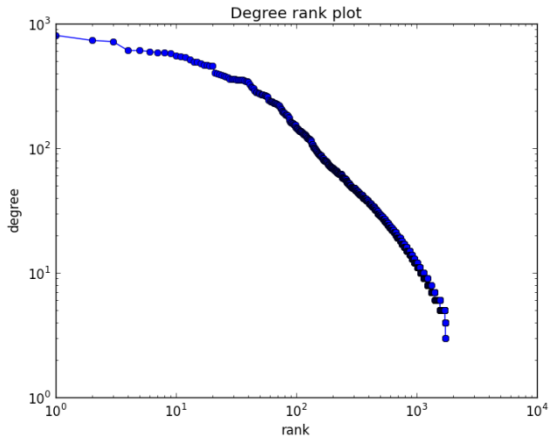


Figure 4. Degree distribution for co-occurrence syllable network

Furthermore, small subnetworks were filtered with ten nodes with the highest degree from all three networks. The results are shown in table V. It is shown that all three networks have almost the same nodes with the highest degree. This indicates that different corpora do not create significant differences between the networks.

TABLE V. THE MOST FREQUENT SYLLABLES

C_A		C_w		C_B	
Syll.	Degree	Syll.	Degree	Syll.	Degree
ma	2299	ma	2296	ma	836
na	2166	na	2156	ti	824
ni	1937	ni	1927	na	753
ti	1918	ra	1897	ni	741
ra	1894	a	1890	ka	627
a	1860	ti	1889	ci	626
ne	1808	ne	1792	ra	623
ka	1801	ka	1773	je	604
o	1692	o	1672	no	595
ta	1682	ta	1670	ne	593

V. CONCLUSION

As the only species capable of verbal communication, language plays an important role in human society, and therefore, it is an important research field. We communicate by using words composed of syllables. Syllable network analysis is an important study in the domain of natural language processing and linguistics.

In this paper we presented different approaches in syllable networks construction. Undirected and unweighted word co-occurrence syllables networks have been already constructed and analysed for two languages: Portuguese and Chinese. The same syllable networks constructed for the Croatian language (from different corpora) exhibited similar results. The networks contain a high cluster coefficient compared to random networks of the same size and small diameter and average path length. In conclusion, the Croatian language syllable networks have properties of small-world networks.

Another approach was to construct a directed and weighted first-neighbour syllable network for the Croatian language. As far as we know, this is the first time this syllable network construction type has been utilized. The main idea of this approach is to capture more information about the properties of each syllable (the successor, the predecessor and strength of connections with other syllables). These properties cannot be analysed in the unweighted and undirected network.

These are just preliminary results and there is still a lot of future research to be conducted in this direction. More detailed statistical analysis should be performed. The experiment should be repeated with different corpora such as Croatian literature and dictionaries. However, it is necessary to determine an exact degree distribution for all networks. Our plan is to analyse the network growth.

These results will be further examined for the purposes of text to speech synthesis improvements. Another plan is to study a different kind of linguistic network, starting with the word co-occurrence network.

REFERENCES

[1] D.J. Watts and S.H. Strogatz, Nature 393, pp. 440, 1998.

[2] S.H. Strogatz, Nature 410, pp. 268, 2001.

[3] S. Milgram, "The Small World", Psychology Today 2, pp. 60–67, 1967.

[4] R. Albert and A.-L. Barabasi, Rev. Mod. Physics 74, pp. 47, 2002.

[5] A. L., Barabasi and R. Albert, "Emergence of scaling in random networks", Science, 286, pp. 509–512, 1999.

[6] J. Borge-Hothoefer and A. Arenas, "Semantic Networks: Structure and Dynamics", Entropy, *12*, pp. 1264-1302, 2010.

[7] R. V. Sole, B. Corominas-Murtra, S. Valverde and L. Steels, "Language Networks: Their Structure, Function, and Evolution", *Complexity*, *15*(6), pp. 20-26, 2010.

[8] R. Ferrer i Cancho, R. V.Sole, "The small world of human language", The Royal Society vol. 268 no. 1482, pp. 2261-2265, 2001.

[9] S. N. Dorogovtsev, and J. F. F. Mendes, "Language as an evolving word web", in Proceedings of The Royal Society of London. Series B, Biological Sciences, 268(1485), pp. 2603–2606, 2001.

[10] P. Masucci, G. J. Rodgers, Network Properties of written human language, Physical Review E, 74(2), 2006.

[11] M. Choudhury, D. Chatterjee, and A. Mukherjee, „Global topology of word co-occurrence networks: Beyond the two-regime power-law", In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 162-170). Association for Computational Linguistics, 2010.

[12] F. C. Pemble, H. Bingol, "Complex Networks in Different Languages: A Study of an Emergent Multilingual Encyclopedia", Unifying Themes in Complex Systems, 1, 612, 2008.

[13] M. Medeiros Soares, G.Corso, L. S. Lucena, "The Network of syllables in Portuguese", Physica A 355, pp. 678-684, 2005.

[14] G. Peng, J. W. Minett, W. S-Y. Wang, "The network of Syllables and Characters in Chinese", Journal of Quantitative Linguistics, 2008.

[15] B. Bollobas, Random Graphs. Academic Press, London, 1985.

[16] M. Turk, Fonologija hrvatskoga jezika, Izdavački centar Rijeka, Tiskara Varždin, Varaždin, 1992.

[17] E. Lewis and M. Tatham,"Word and syllable concatenation in text-to-speech synthesis", In:Sixth European Conference on Speech Communications and Technology, pp. 615–618, 1999.

[18] A. A. Hagberg, D. A. Schult and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, 2008.

[19] M. Bastian, S. Heymann, and M. Jacomy, Gephi: An open source software for exploring and manipulating networks, 2009.

Performance Comparison of Sequential and Parallel Execution of the Ant Colony Optimization Algorithm for Solving the Traveling Salesman Problem

Elmedina Fejzagić, Adna Oputić
Faculty of Electrical Engineering
University of Sarajevo
Sarajevo, Bosnia and Herzegovina
efejzagic1@etf.unsa.ba, aoputic1@etf.unsa.ba

***Abstract* - Ant Colony Optimization (ACO) is a metaheuristic alghorithm which uses ideas from nature to find solutions to instances of the Travelling Salesman Problem (TSP) and other combinatorial optimisation problems. ACO is taken as one of the high performance computing methods for TSP. In this paper, the impact of parallelizing an ant colony optimization (ACO) algorithm for the traveling salesman problem in increasing performances is studied, using the task parallel library. One of the main reasons for parallelizing this alghoritm is to reduce the time needed to find a solution while the quality of solution is the same as in the algorithm which is not parallelized.**

***Keywords* - ant colony optimization, travelling salesman problem, parallelization**

I. INTRODUCTION

A system of n parallel processors is generally less efficient than a single n-times-faster processor, but the parallel system is often cheaper to build. The availability of parallel architectures at low cost has widened the interest for the parallelization of algorithms and metaheuristics.

ACO is a relatively novel meta-heuristic technique and it has been successfully used in many applications, especially problems in combinatorial optimization. ACO algorithm models the behavior of real ant colonies in establishing the shortest path between food sources and nests. Ants can communicate with one another through chemicals called pheromones in their immediate environment. The ants release pheromone on the ground while walking from their nest to food and then go back to the nest. The ants move according to the amount of pheromones, the richer the pheromone trail on a path is, the more likely it would be followed by other ants. So a shorter path has a higher amount of pheromone in probability, ants will tend to choose a shorter path. Through this mechanism, ants will eventually find the shortest path. Artificial ants imitate the behavior of real ants, but can solve much more complicated problem than real ants can. Ant Colony Optimization has been widely applied to solving various combinatorial optimization problems such as Job-shop Scheduling Problem, Vehicle Routing Problem, Quadratic Assignment Problem etc. It is also used for solving Traveling Salesman Problem which solution and it's performances are being studied in this paper.

Even if the ACO algorithm in its original sequential form has proven to be a good solution, it is still very demanding in computational time and resources and this is particularly true as the problem size increases. The generation and evaluation of ants , which are the main parts of the algorithm and its principal source of computational cost, offers a low dependency degree, so the structure of the algorithm makes it well suited for concurrent execution. For those reasons, a parallel approach was justified and an implementation was made to study the performance in execution time that could be obtained that way.

The Task Parallel Library (TPL) is used for the parallelization of ACO algorithm. This library introduces the concept of "task". Task parallelism is the process of running these tasks in parallel. A task is an independent unit of work, which runs within a program. TPL provides increasing the performance of the algorithm, efficiently using all the processors available.

II. ANT COLONY OPTIMIZATION ALGORITHM

Ant colony optimization (ACO) takes inspiration from the behavior of ant species. In natural world, while walking towards food from the nest and back, ants leave pheromone trails. Other ants perceive the presence of pheromone and tend to follow paths where pheromone concentration is higher. Through this mechanism, ants are able to transport food to their nest in a

remarkably effective way. Over time pheromone trail starts to evaporate, and thus the pheromone density on short paths becomes higher than on long paths - the short paths get marched more frequently.

In computer science and operations research, the ant colony optimization algorithm is a technique for solving problems which can be reduces to finding good paths trough graph [2]. The original idea for ant colony optimization algorithm comes from observing ant colonies behavior previously described. Pheromone evaporation, that was described, has the advantage of avoiding convergence to a locally optimal solution.

Ant colony optimization has been formalized into a metaheuristic for combinatorial optimization problems [3]. A metaheuristic is a set of algorithmic concepts that can be used to define heuristic methods applicable to a wide set of different problems. The use of metaheuristics has significantly increased the ability of finding very high-quality solutions to hard, practically relevant combinatorial optimization problems in a reasonable time[4]. Ant Colony Optimization metaheuristic is presented in pseudo code:

Set parameters, initialize pheromone trails
while termination condition not meet **do**
ConstructAntSolutions
ApplyLocalSearch (optional)
UpdatePheromones
endwhile

Algorithm 1. The ACO metaheuristic in pseudo-code[1]

Ant Colony optimization algorithm can be presented as described above.

After initialization, the metaheuristic iterates over three phases:

ConstructAntSolutions: This phase manages a colony of ants that concurrently and asynchronously visit adjacent states of the considered problem by moving through neighbor nodes of the problem's construction graph. They move by applying a stochastic local decision policy that makes use of pheromone trails and heuristic information. In this way, ants incrementally build solutions to the optimization problem.

ApplyLocalSearch: In this phase solution obtained by the ants are improved trough a local search. This phase is highly problem-specific.

UpdatePheromones: In this phase values associated with good or promising solutions are increased, and all the pheromone values are decreased through *pheromone evaporation.*

III. TRAVELING SALESMAN PROBLEM

The traveling salesman problem (TSP) is to find a routing of a salesman who starts from home location, visits a prescribed set of cities and returns to the original location in such a way that the total distance travelled is minimum and each city is visited exactly once. TSP in its generality represents a typical 'hard' optimization problem[5]. Thus, it is likely that the worst-case running time for the TSP increases exponentially with the number of cities.

The TSP can be formulated as an integer linear program (the program in which some or all of the variables are restricted to be integers):

Minimize

$$\sum_{i=0}^{n}\sum_{j=1}^{n} cijxij$$

Subject to

$$\sum_{i=1}^{n} xij = 1, \quad j = 1,\ldots,n \qquad \sum_{j=1}^{n} xij = 1, \quad i = 1,\ldots,n$$

$$xij = 0 \text{ or } 1$$

Figure 1. Mathematical formulation of TSP [5]

In the above formulation, c_{ij} corresponds to the distance between cities *i* and *j* and x_{ij} indicates whether the path from city *i* to city *j* is included in the tour.

TSP can be represented by complete weighted graph G(N,A) with N being a set of nodes representing the cities, and A being the set of arcs. Each arc (i,j) ∈ A is assigned a value(length) cij distance between cities *i* and *j*) [4].

Computing a solution can be done by [6]:

- Devising algorithm for finding exact solutions (they work reasonably fast only for small problem sizes)
- Devising heuristic algorithms that deliver either seemingly or probably good solutions, but which could not be proved to be optimal.
- Finding special cases for the problem ("subproblems") for which either better or exact heuristics are possible.

IV. ACO FOR THE TRAVELING SALESMAN PROBLEM

In ACO for the TSP, each node of the graph represents a city, and each arc represents a connection between two cities. A variable called pheromone is associated with each node and it can be read and modified by the ants. At each iteration of algorithm, artificial ants build a solution by walking from node to node on the graph with the constraint of not visiting any node that she has already visited in her walk, and eventually going back to the initial node .

In each step of solution construction, an ant arriving in node (city) i chooses the next city to move to as a function of the pheromone values and of the heuristic values on the arcs connecting city i to the cities j the ant has not visited yet until all cities have been visited (Figure 2).

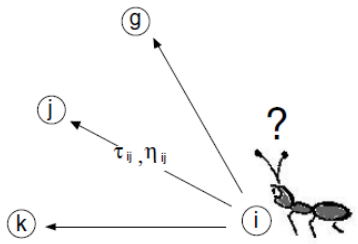


Figure 2. Artificial ant moving to the next city [4]

At the end of an iteration, on the basis of the quality of the solutions constructed by the ants, the pheromone values are modified in order to bias ants in future iterations to construct solutions similar to the best ones previously constructed [1].

V. DATA PARALLELISM USING .NET FRAMEWORK TASK PARALLEL LIBRARY

Data parallelism (also known as loop-level parallelism) is a form of parallelization of computing across multiple processors in parallel computing environments. Data parallelism focuses on distributing the data across different parallel computing nodes [8].

The Task Parallel Library (TPL) library simplifies the process of adding parallelism and concurrency to application and provides writing multithreaded and parallel code. By using TPL it is possible to maximize the performance of the code, efficiently using all the processors available.

TPL supports data parallelism trough System.Threading.Tasks.Parallel class. This class provides method-based parallel implementation of for and foreach loops [9].

The example of parallelizing for loop using TPL is shown in Figure 3 and Figure 4.

```
for (int i = 0; i < matARows; i++)
{
    for (int j = 0; j < matBCols; j++)
    {
        for (int k = 0; k < matACols; k++)
        {
            result[i, j] += matA[i, k] * matB[k, j];
        }
    }
}
```

Figure 3. Sequential for loop

```
Parallel.For(0, matARows, i =>
{
    for (int j = 0; j < matBCols; j++)
    {
        // Use a temporary to improve parallel performance.
        double temp = 0;
        for (int k = 0; k < matACols; k++)
        {
            temp += matA[i, k] * matB[k, j];
        }
        result[i, j] = temp;
    }
}); // Parallel.For
```

Figure 4. Parallelizing for loop using TPL: by writing Paralell.For instead of for, the TPL handles all the low-level work

VI. IMPLEMENTATION

The sequential algorithm used in this paper was written by James McCaffery in C# programming language [7]. Ants are initialized to random trails trough cities. The goal is to find the shortest path that visits each city exactly once.

This algorithm was adapted by using data parallelism as described previously.

VII. PERFORMANCE COMPARISON

Execution time of the program was estimated using Intel Core i3-2350M processor that has 2 cores and 4 threads.

In first performance comparison, number of ants in program was set to a value 10, and the number of cities was changed on each execution. Measurement times for the overall performance test are shown in Figure 5.

Number of cities	Sequential execution time (ms)	Parallel execution time (ms)	Sequential -paralell execution time (ms)
5	102	84	18
10	342	329	13
15	951	915	36
20	1965	1841	124
30	5407	5293	114
50	21551	21368	183
75	67314	67074	240
100	153990	153684	306

Figure 5. Measurement times for the overall performance test 1

As the number of cities was larger, increasing the complexity of problem, the sequential execution time- parallel execution time was increasing also. Figure 6. shows the dependence of sequential execution time- parallel execution time of increasing the number of cities.

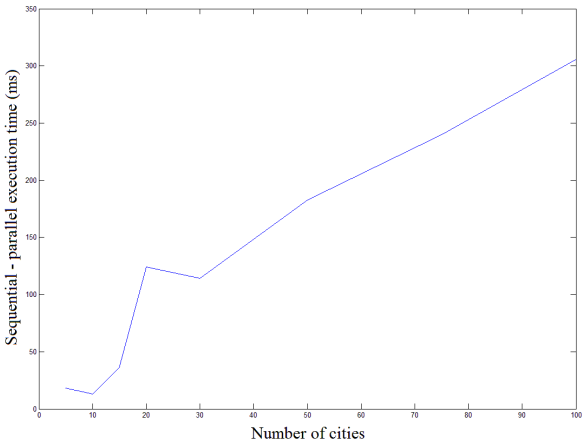


Figure 6. Dependence of sequential execution time- parallel execution time of increasing the number of cities

In second performance comparison, number of cities in program was set to a value 7, and the number of ants was changed on each execution. Measurement times for the overall performance test are shown in Figure 7.

Number of ants	Sequential execution time (ms)	Paralell execution time (ms)	Sequential - parallel execution time (ms)
5	84	81	3
10	163	162	1
15	291	288	3
100	1456	1386	70
250	3505	3394	111
500	6717	6584	133
2500	33164	32943	221
5000	66379	65889	490

Figure 7. Measurement times for the overall performance test 2

As the number of ants was larger, increasing the complexity of problem, the sequential execution time- parallel execution time was increasing also. Figure 8. shows the dependence of sequential execution time- parallel execution time of increasing the number of ants.

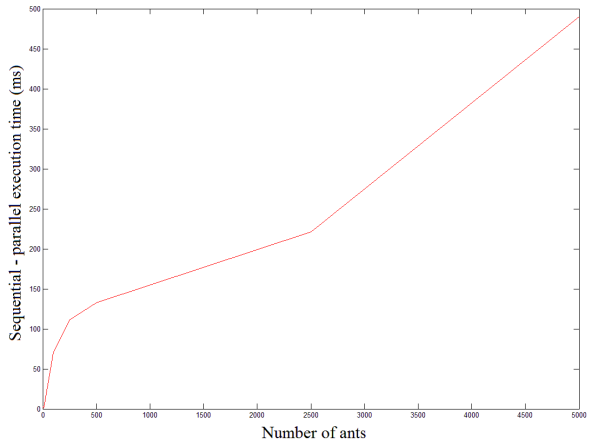


Figure 8. Dependence of execution time- parallel execution time of increasing the number of ants

VIII. Conclusion and Future Work

In this paper a parallel implementation of an Ant Colony Optimization metaheuristic that is applied to Traveling Salesman Problem is presented. The nature of the ACO and the functionallity offered by Task Parallel Library made the

transition from sequential to parallel programming straightforward, while some changes had to be made to the algorithm and to the program to obtain the level of efficiency which is achieved.

The main aim was to reduce the execution time of the algorithm. The resulting implementation has shown that is was possible to design an efficient parallel Ant Colony Optimization metaheuristic using TPL. Graphs shown in Figure 6. and Figure 8. show that acceptable efficiency can be achieved for larger problems, respectivley as the number of key parametres increases.

The future plan is to apply the parallel ACO algorithm to other combinatorial optimisation problems to see whether the same results obtained for the Travelling Salesman Problem can be reproduced.

REFERENCES

[1]Marco Dorigo, Mauro Birattari, Thomas Stützle, "Ant Colony Optimization, Artificial Ants as a Computational Intelligence Technique", IEEE Computational Intelligence Magazine, 2006.

[2]http://en.wikipedia.org/wiki/Ant_colony_optimization_algorithms

[3]M. Dorigo,G. Di Caro, "The Ant Colony Optimization meta-heuristic," in *New Ideas in Optimization*, D. Corne et al., Eds., McGraw Hill, London, UK, 1999.

[4]Marco Dorigo and Thomas Stützle, "Ant Colony Optimization", Massachusetts Institute of Technology, 2004.

[5] G. Gutin,A.P. Punnen, "The Traveling Salesman Problem and Its Variations", Kluwer Academic Publishers, 2002.

[6]http://en.wikipedia.org/wiki/Travelling_salesman_problem

[7] http://msdn.microsoft.com/hr-hr/magazine/hh781027(en-us).aspx

[8] http://en.wikipedia.org/wiki/Data_parallelism

[9]http://msdn.microsoft.com/en-us/library/dd460717.aspx

Parallelization and Performance Analysis of the Simulated Annealing Algorithm for Graph Coloring Problem

Lejla Bećirspahić, Adisa Đulović, Novica Nosović
lbecirspah1@etf.unsa.ba, ad15352@etf.unsa.ba , novica.nosovic@etf.unsa.ba

Faculty of Electrical Engineering, University of Sarajevo

***Abstract* – Vertex coloring is a subset of the graph coloring problem. It is of great importance in many applications. Vertex coloring implies a coloring of the vertices of the graph with minimal number of colors (k), so that adjacent vertices have different color. The paper presents a hybrid implementation of Simulated Annealing algorithm for k-coloring of the vertices of the graph. The programming has been done with the use of CUDA toolkit. In order to find out how the speedup is achieved by parallelization, a sequential implementation for the problem has been used as a starting point. The results of CUDA based program and sequential implementation are analyzed. This hybrid implementation shows significant improvement and can be used as base for future work.**

Key words: parallelization, CUDA, graph coloring, simulated annealing

I. INTRODUCTION

Examples of the use of graph coloring are found in the analysis of computer networks, social networks, the results of simulation etc. Graph coloring can be divided into two main categories: vertex coloring and edge coloring. Vertex coloring is an assignment of colors to a set of vertices of a graph, so that the neighboring vertices are associated with different colors. Edge coloring is a coloring of edges with a minimal number of colors, considering that adjacent edges (both associated with a vertex) must be differently colored. In this paper is shown graph's vertex coloring algorithm.[1]

II. USED TOOLS

A. *Description of the used environment*

In this paper was used for demonstration NVidia's GeForce GT 640M LE graphic card. More about this card can be found on NVidia official site.

SIMD (single instruction, multiple data) is execution model used for general-purpose computing. Similarly, NVidia model is SIMT (Single Instruction Multiple Threads). This implies that all threads run the same instruction at time but with different data. In this model, host program launches a kernel function that generates a large number of parallel threads. Threads are grouped into blocks. Independent blocks are executed concurrently on each multiprocessor. Threads in a same block are assigned to a single multiprocessor and are executed within its own execution environment. [2]

To meet demands for programming for GPU, NVidia has developed the CUDA programming model. CUDA offers a very simple abstraction of parallelism and extends C semantics. Thus, the parallel systems became available to a wider range of programmers. CUDA is a software environment that allows programmers to isolate components of the parallelism at the level of data and their performance on specialized FP unit [3].

CUDA program consists of serial code, executed on CPU, and parallel code, executed on GPU. Before any computation, data must be pulled from CPU to GPU device memory. After calculations have been done, obtained results must be returned from the GPU. Described data transfer between host (CPU) and device (GPU) can harm performances, and can be considered as one of limitations of CUDA programming model.

B. *Used algorithm*

The process of simulated annealing (hereafter SA), was independently researched by S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi in 1983 and V. Cerny in 1985. Annealing is the process of slowly cooling a hot mass (e.g. glass or metal) in order to correctly form crystal structure and obtain a firmer material. Simulated annealing is a heuristic algorithm that mimics finding

of the minimum energy state in a process of annealing metal. [4]

Optimization principle is as follows:

-In each iteration, for state S from the previous iteration is selected random neighboring S' state

-If the energy of state S' is lower than the energy of state S, a particle goes into state S', and if the state S' is the state of higher energy, S becomes S', with a probability that depends on the temperature and the difference between energies of S and S'.

The algorithm starts from some initial high temperature and reduces it through iterations to very small value. The function of probability of transition from a lower state to the higher energy state can be any function, with the restrictions that the function is positive and that its value will decrease as the temperature decreases and with the increase of energy differences between the states. [4]

```
s = initial state
temperature = initial temperature
while temperature > 0 works:
s' = arbitrary neighbor (s)
dE = energy (s') - energy (s)
if dE < 0 or f (temperature, dE)> random () then:
s = s'
temperature = new temperature (temperature)
solution = s
```

Figure 1: Pseudo code of simulated annealing

III. SEQUENTIAL EXECUTION

In order to carry out an implementation of graph coloring using parallel simulated annealing algorithm, first step was to develop sequential code for described problem. In next section an implementation of sequential version of this algorithm will be described. A aim is to compare performance of CUDA version against performance of sequential version of code, to find out how the speedup is achieved after parallelization.

The general schema of graph coloring using sequential simulated annealing:

//initialization

1) color given graph using greedy algorithm in order to get minimum number of colors needed to color nodes so that adjacent nodes have different color
2) initialize adjacency matrix for given graph
3) color given graph randomly, and calculate number of conflicting nodes

//simulated annealing

4) start with a given temperature
5) choose random conflicting node from graph and find random color that is different from current node's color
6) color selected node with chosen color, update adjacency matrix and get number of conflicting nodes
7) calculate profit of the move as a difference between current number of conflicting nodes and new number of conflicting nodes
8) if profit is greater than 0, it means that new solution is better, so it will be accepted, otherwise, possibility of accepting this solution is calculated.
9) possibility of accepting new solution is compared with some random number from range (0, 1). If random number is greater than possibility, solution isn't accepted. Changes, that are made on adjacency matrix in step 6. Otherwise, solution is accepted
10) repeat steps 5-9 until given iteration's improvement, which represents equilibrium state for given temperature, has been reached
11) after an equilibrium has been achieved , decrease current temperature with a given decreasing factor
12) repeat steps 5-9 with new temperature
13) repeat steps 5-12 until achieving minimal temperature

IV. PARALLEL EXECUTION

There are several strategies [5] on how to parallelize SA:

- Application dependent parallelization: execution of operations is divided between processors
- Domain decomposition: searching the space is divided into subdomains, thus each subdomain is assigned to a single processor
- Access of multiple Markov chains: the most natural way for parallelization strategy of SA is tracking multiple Markov chains, where multiple Markov chains execute asynchronously and exchange final states after a certain period or at the end of the process. The clearest approach is that when the Markov chains interact only at the end of the process. This is called asynchronous access. On the other hand, in the synchronous approach Markov chains communicate in inter-temperature levels. Only the function values are exchanged between workers. Communication can take place at any temperature level or after a certain number of temperature levels. Kernel is launched from the host. With given number of threads the CUDA kernel is simultaneously executed in parallel, thus allowing calculating a given number of Markov chains. This kernel launches a grid of given number of blocks. Each block contains given number of threads. Constant parameters like number of colors, number of nodes and temperature factor are stored in constant memory. These parameters are available to all threads. Constant memory is read-only and provides short

latency and high bandwidth when multiple threads access the same memory position.[6] This is important because these consecutive accesses are repeatedly required by the SA algorithm. Each step of the Markov chain three requires a random numbers. NVidia's CURAND library facilitates generation of random numbers to use them immediately by the kernels. This has reduced the time of reading and writing from the global memory. More about CURAND library can be found on [7].

One of the functions that have been parallelized is function for computing number of conflicts for each node. When all threads have finished computing the number of conflicts, the sum is calculate by reduction operation. This operation is carried out in parallel inside the GPU, by means of the specific optimized NVidia's Thrust library, that takes advantage of the consecutive memory access and the involved partial reductions are performed in shared memory. More about Thrust library can be found on [8].

V. MEASUREMENTS RESULTS AND ANALYSIS

In this section are presented results of performances tests of given sequential and parallel implementation of graph coloring using simulated annealing. These tests have been performed on Intel Core TM i5-32 @ 2.50GHz processor and GeForce GT 640M LE. Algorithm was executed with four different dimensions of graphs, three times each. The same SA parameters were used for each execution.

Previously, the best practices for parallelization of SA algorithm have been proposed. In consequent text will be analyzed performances of hybrid implementation of this algorithm. Meaning of the word hybrid is that main part of SA algorithm is executing sequentially. Parallel functions, like function for procurement of conflicting nodes and function for conflict resolution, are called from sequential code. Prerequisite for SA algorithm is that nodes are colored with some initial colors, so the method for random coloring is parallelized.

Number of blocks is 256. Number of threads per block is 128. In each kernel thread id is calculated as BlockIdx.x * BlockDim.x + ThreadIdx.x, where thread id is not greater than number of nodes in graph. In that way, each thread corresponds to one node.

The Visual Profiler® has been used to analyze performance issues. Results of measurements and graphs that are used, are given as follows:

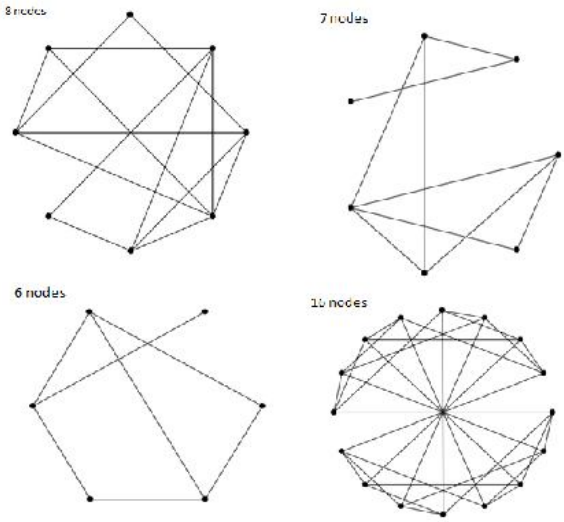


Figure2:Graphs that are used for measurements of performance

	Number of conflicts	Number of colors	Time (ms)
6 nodes (1)	2	3	17.5731
6 nodes (2)	4	3	18.2378
6 nodes (3)	12	3	13.1136
7 nodes (1)	4	3	18.7682
7 nodes (2)	8	3	19.158
7 nodes (3)	6	3	25.5756
8 nodes (1)	8	4	64.08
8 nodes (2)	2	4	60.75
8 nodes (3)	6	4	57.2295
16 nodes (1)	28	2	44.9161
16 nodes (2)	36	2	51.7744
16 nodes (3)	32	2	52.7551

Table 1: Execution time for sequential implementation

	Number of conflicts	Number of colors	Time (ms)
6 nodes (1)	12	3	9.18541
6 nodes (2)	8	3	11.8861
6 nodes (3)	14	3	11.6594
7 nodes (1)	16	3	14.6977
7 nodes (2)	14	3	20.7993
7 nodes (3)	6	3	7.63725
8 nodes (1)	8	4	61.5156
8 nodes (2)	16	4	36.7697
8 nodes (3)	6	4	66.2451
16 nodes (1)	32	2	28.2978
16 nodes (2)	32	2	17.0156
16 nodes (3)	36	2	53.1832

Table 2: Execution time for parallel implementation

Number of nodes	CPU Time (ms)	CUDA Time (ms)	Speedup
6	16.30817	10.9103	0.49474
7	21.16727	14.37808	0.47219
8	60.6865	54.84347	0.10654
16	49.8152	32.8322	0.51726

Table 3: Summary table with speedup

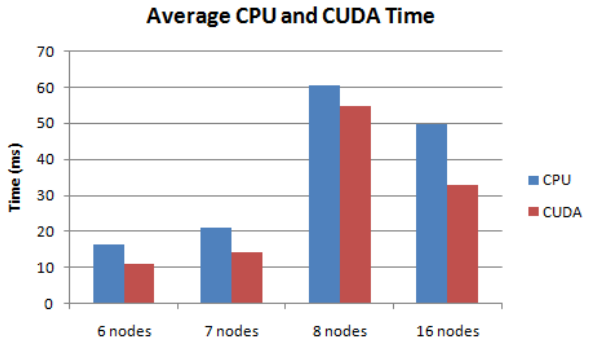


Figure 3: The figure shows the diagram of average run-time of the SA algorithm for the graph coloring problem. The input data are the same for sequential and parallel execution.

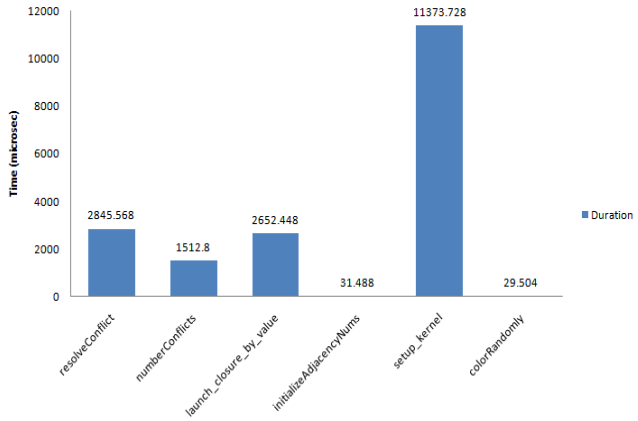


Figure 4: Shows summary time for each kernel execution for graph with 7 nodes.

Figure 3 shows how much is improvement on parallel execution. It can be seen that the best speedup is for graph with 16 nodes (0.51726). The worst speedup is for graph with 8 nodes (0.10654). That means no matter how many nodes graph has, speedup depends of number of edges. The graph with more edges will have more conflicts to resolve, so it will take more time.

Figure 5 shows average occupation of GPU in % by each function. Functions intializeAdjancencyNums and colorRandomly are called just once, so the average occupation is the smallest. The maximum of GPU occupation is taken by setup_kernel.

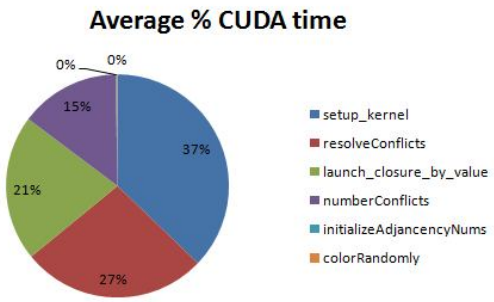


Figure 5: This figure shows occupation of GPU in % by each function.

VI. CONCLUSION

In this paper was developed implementation of an algorithm of SA algorithm for graph coloring problem, and has been carried parallelization of the code on the GPU. Implementation described above has its limitations, especially for larger graphs. For that reason measurements are done on graphs with small dimensions (6, 7, 8 and 16 nodes). The worst results are acquired for dense graphs. The best results are acquired for graphs with fewer edges. The algorithm is dependent of initial coloring. The initial coloring was achieved randomly. That means, if graph is colored with more conflicts it will take more time to resolve those conflicts.

REFERENCES

1. Jensen T.R., Toft B.: Graph coloring problems. Wiley Interscience, New York(1995)
2. David B. Kirk, Wen-mei W. Hwu , „Programming Massively Parallel Processors – A Hands-on Approach“, Elsevier Inc., 2010
3. Michael Garland, Scott Le Grand, John Nickolls, Joshua Anderson, Jim Hardwick, Scott Morton, Everett Philips, Yao Zhang, Vasily Volkov, „Parallel Computing Experiences with CUDA“, IEEE Micro, july - august 2008.
4. Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P.: "Optimization by Simulated Annealing", Science, Vol. 220, No. 4598, 1983
5. A.M.Ferreiro, J.A.Garcia, J.G.Lopez-Salas, c.Vazquez, “An efficient implementation of parallel simulated annealing algorithm in GPU’s”, Springer Science-BusinessMedia, LLC 2012
6. David Luebke, „CUDA: Scalable Parallel Programming for High-performance Scientific Computing“, 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008.
7. CURAND Library, NVidia Coroporation, Santa Clara, CA 95050, 2010
8. Thrust:A productivity-Oriented Library for CUDA, Nathan Bell, Jared Hoberock, 2011
9. Shane Ryoo, Christopher I. Rodrigues, Sara S. Baghdorkhi, Sam S. Stone, David B. Kirk, Wen-mei W. Hwu, „Optimization Principles and Application Performance Evaluation of Multihreaded GPU Using Cuda“, Proceedings of the 13th ACM SIGPLAN Symposium on Principles and practice of parallel programming, 2008.
10. Kubale, M. (Eds.): Graph colorings. American Mathematical Society, (2004).
11. Jason Sanders, Edward Kandrot, „CUDA by Example – An Introduction to General Purpose GPU Programming“, Addison-Wesley, 2011

Performance Analysis of the Traveling Salesman Problem Optimization Using Ant Colony Algorithm and OpenMP

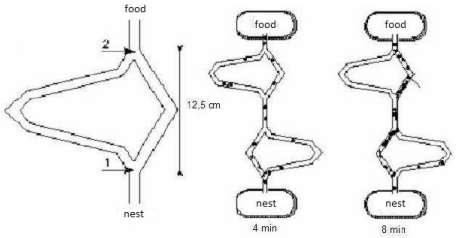
Almin Milanović, Mustafa Duranović and Novica Nosović
Faculty of Electrical Engineering University of Sarajevo, Sarajevo, Bosnia and Herzegovina
amilanovic2@etf.unsa.ba, mduranovic1@etf.unsa.ba and nnosovic@etf.unsa.ba

Abstract-Optimization using ant colony (ACO) is one of algorithms which is used for distributed control and optimization. Beside static methods these methods and algorithms are more flexible and robust in dynamical environments such as traffic on Internet and standard telephony. These problems belong to class of hard problems because of huge space of possible solutions which need to be found in reasonable time. Traveling salesman problem (TSP) belongs to hard problems. Since solving these problems need a lot of time for execution this paper presents an attempt to decrees execution time using parallelization on multicore processors. OpenMP was used as a main parallelization tool. Certain acceleration was achieved.

Ant colony optimization, artificial ants, Travelling salesman problem, Ant algorithm.

I. INTRODUCTION

Ant colony optimization is first introduced by Marco Dorigo [7]. Ant colony represents multi-agent system where the behavior of one artificially ant is inspired with the behavior of natural ants. The usage includes classic NP-hard problems and it goes back to routing in computer and telecommunications to data mining. Inspiration for ant colony optimization was given through experiment with ants. Food is given in arena with two different distance paths, as it's showed on Picture 1. On one side there is ant nest and on another food source. In every direction ants have to choose path which they will travel.



Picture 1. Experiment with ants

Experiment shows that after short transitional phase most of ants use shortest path. It is showed that by increasing path distance it also increases probability of choosing shortest one. That ant behavior is explained by stigmergy which provides feedback depends on different distance paths. Term stigmergy means indirect communication way through local modifications of the environment. It is accomplished by placing chemical compounds called pheromones. At point selection, ant brings probable decision which way to go based on the amount of pheromones on soil. In the next iteration, the chosen path is becoming more likely. In this way feedback is given. Since on the short path pheromones are placed more often, ants through a number of iterations chose shortest path.

A. TRAVELING SALESMAN PROBLEM

Traveling salesman problem represents real problem which is solved using ACO. It is defined with: Let C be a set of cities, and L a set of links that connect the elements of C. With *i* and *j* represent price (distance) between *i* and *j cities.* TSP's optimization problem of finding minimum time on Hamilton graph G=(C,L). Hamilton's path is closed tour Y which visits just once every city in C. Its length is the sum of the prices of *ij* nodes which consists from. Additionally, the distance has to be non-symmetrical.

B. ARTIFICIALLY ANTS

The number of cities to visit is n, m is total number of ants in the system, and $b_i(t)$ is number of ants in the city *i* in the time t. Ant decides which city next to go to based on function that has parameters, distance between cities and city's trail, pheromones for the path which connects them. To avoid bypassing same city more than once each ant has its taboo list in which he placed each city he visited. When ants taboo list spanning all n city, ant sets pheromones on each path he used for tour cities and resets its taboo list. Number of visits that will make each ant is restricted default. Globally every ant in the course of time t decide where it will be in time t+1, and in the next iteration step it moves there. So all ants after n iterations will go through all the cities (the instant t+n) and set a trail. After this new data is loaded to pheromones in way to ensure that the value is multiplied by evaporation factor ρ (evaporation factor is a confusing

name, in fact it says exactly otherwise, how much of the trail does not evaporate) and it is adds every trail of ants who passed that way. Formula (1) and formula (2)

$$\tau_{ij}(t+n) = \rho \cdot \tau_{ij}(t) + \Delta\tau_{ij} \quad (1)$$

Restoring pheromone trail on the path.

$$\Delta\tau_{ij} = \sum \Delta\tau_{ij}^{k} \quad (2)$$

Added value of trail

The amount of pheromone added to some path is 0 if ant hasn't passed that path. If ant passed that path, it leaves a certain amount of trail that is inversely proportional to the time that ant spend for tour, as shown formula (3)

$$\Delta\tau_{ij}^{k} = \frac{Q}{L_k} \quad (3)$$

Trail of individual ant

In the Formula 3.3 Q is default constant and Lk total time by kth ant crossed for its tour of all the cities. While deciding where to go besides using trail pheromones, the ant uses visibility. Visibility is a parameter of the inverse distance. As the city is farther, the visibility is lesser (visibility = 1/ distance). Total counting for decision is given by the formula (4).

$$p_{ij}^{k}(t) = \frac{[\tau_{ij}(t)]^{\alpha} \cdot [\eta_{ij}]^{\beta}}{\sum [\tau_{ik}(t)]^{\alpha} \cdot [\eta_{ik}]^{\beta}} \quad (4)$$

Total counting for decision

This formula (4) is valid in the event that an ant has nowhere to go, that's taboo list does not include all existing cities. η_{ij} represents visibility between i and j. α and β are the parameters that govern the relationship between importances of a trail in relation to visibility. For the greater α ant will increasingly make their decisions based on the trail of pheromones. For the greater β analogues to the trail, the ant will make the decision to give more importance visibility of cities. Result pijk(t) is the probability that the ant k move from the city i to the city j at time t.

II. ANTS ALGORITHM

When starting with algorithm the first step is to create and place cities, and trail pheromones on the paths that connects them. Then create and set up all the ants. When setting up the ants it is also needed to make their taboo list. Ants are placed in initial city and that initial city is treated as the first member of the taboo list. Then iteration begins for the movement of ants. After n iterations all ants have a full list of taboo and then the setting a trail of pheromones for each ant starts. Ants are again placed in some initial city and their taboo list get empty (ie there is only a new starting city). The cycle of n iterations is set so many times until the users interrupt the program or the program reaches the maximum allowable cycles.

Pseudo code:

When starting the program it is necessary to initialize program:

```
CycleCounter=0;
BestRoute=[maxValue];
Set n cities;
Set Trail[i,j] on default value c for all i and j from 1 to n;
Create m ants;
Loop a certain number of cycles:
While(CycleCounter < MAXNumberCycle) {
        CycleCounter++;
        DoCycle();
        CheckBestRoute();
        EvaporateTrail();
        ApendTrail();
        ResetAnts();
}
Funkction DoCycle() looks like this:
For each i to n {
        For each ant k from 1 to m {
        k.NextStep();
        }
}
Function CheckBestRoute() for flowing :
For each ant k from 1 to m{
        If (BestRoute>k.LengthRout){
        BestRoute=k.LengthRoute;
        }
}
```

Function EvaporateTrail() just multiplies all the elements of the matrix Trail by a factor of evaporation, and functions ApendTrail() going through all the ants, and for their every step adds a certain amount of matrix trail on a suitable place. ResetAnts() function erases all the ants Taboo list and returns them to their home cities. The complexity of the individual steps is different, but with a little observation and comparison it is possible to do the most complex part of the complexity of the algorithm (function DoCycle). If the maximum allowable number of repetitions of cycles denote by NC, the number of ants m, and the number of cities with n, then the complexity of the algorithm is O (n^2 xMx NC). NC is the number of times a loop can execute everything inside, n is the number of passes for loop in the function DoCycle, m is the number of ants which are used, x_n represent the probability calculations within a function NextStep for each ant, which gets *NC* x *n* x *m* x *n*.

III. EXPERIMENT DESCRIPTION

Fort this study are used next processors:

- Intel® Core™2 Duo Processor T6400 (2M Cache, 2.00 GHz, 800 MHz FSB).

- Intel® Core™ i3-330M Processor (3M Cache, 2.13 GHz)
- Intel® Core™ i5-750S Processor (8M Cache, 2.40 GHz)

More about these processors it is possible to find on site of manufacturer [6].

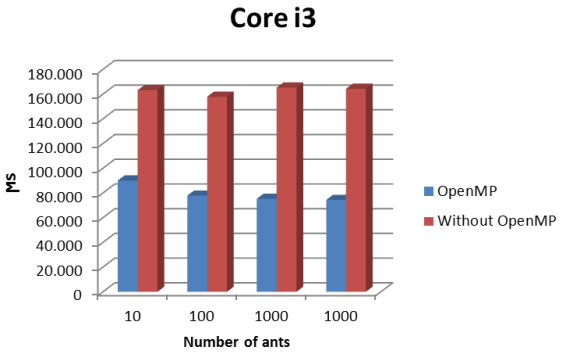
For the parallel programing is used OpenMP which is a portable standard for the programming of shared memory systems [1]. The compiler directives can be used to extend the sequential languages Fortran, C, and C++. In this study it is used OpenMP within C++. The programming model of OpenMP is based on cooperating threads running simultaneously on multiple processors or cores. These threads are used for parallel execution for each ant.

A. PERFORMANCE ANALYSIS

Piece of code that is parallelized is the movement of ants through the graph. Considering that each ant has its own list of taboo that is independent from the list of other ants we can run all the ants for parallel browsing graph. A synchronization barrier is faced when all the ants visit all the cities. Have to wait until all ants visit all the cities, then performs the resetting of ants and then restart them. Given that most of the execution time spent on passing through the towns of ants, parallelization in this section will result in significantly lower execution time of the whole algorithm. This algorithm is runned on 3 computers with parallel directives and without parallel directives. Multiplying the number of ants in the algorithm and the number of times which was launched algorithm for a test is 100 000. The results are summarized below.

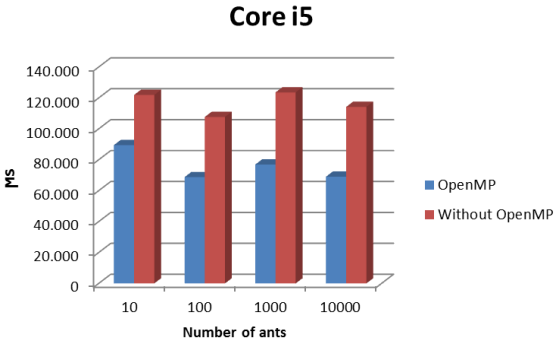
Picture 2. Results of measurements of time to execute the algorithm on a Core 2 Duo processor

Average acceleration is obtained by inclusion of parallel directives to the Core 2 Duo is approximately 37%.



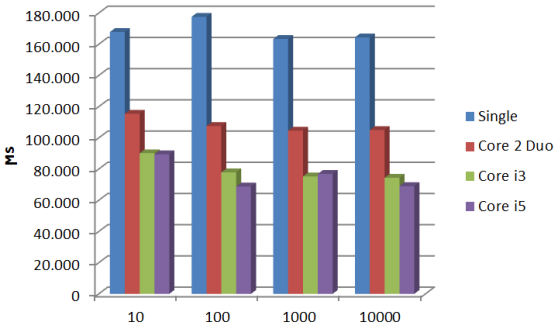
Picture 3. Results of measurements of time to execute the algorithm on a Core i3 processor

Average acceleration is obtained by inclusion of parallel directives to the Core i3 is approximately 50%.



Picture 4. Results of measurements of time to execute the algorithm on a Core i5 processor

Average acceleration is obtained by inclusion of parallel directives to the Core i3 is approximately 35%. Since all processors have speeds of 2GHz, and the time to execute on a processor Core i5 without parallel directives is for about 25% less compared to the execution of the processor Core 2 Duo and Core i3, we can conclude that this processor has some kind of parallelization although it is not explicitly stated with using OpenMP directives. The following picture shows the graph of the execution time of parallel programs on all three processors in relation to the time of execution of the algorithm without parallel directives.



Picture 5. The execution time of the program with parallel directives for 3 processors in relation to the time of execution of the algorithm without parallel directives.

IV. CONCLUSION

Using OpenMP reduces ACO algorithm execution time by 30-50%. Using some modern (evolution) algorithm and parallel programing together can be effective for hard problems such as TSP. These results are reason to try to paralelize other algorithms which would solve other hard problems. To achieve mentioned first step is to check is algorithm compatible with parallel programing. If iterations of algorithm are independent from each other that is good signal for using OpenMP within.

References:

[1] Rohit Chandra , Rohit Chandra, Leonardo Dagum, Dave Kohr, Dror Maydan, Jeff McDonald, Ramesh Menon "Parallel Programing in OpenMP" Publisher: Morgan Kaufmann; 1 edition (October 16, 2000)

[2] Thomas Rauber and Gudula Rünger "Parallel Programming: for Multicore and Cluster Systems" (Mar 10, 2010)

[3] www.openmp.org

[4] http://jason.cse.ohio-state.edu/examples/npb/index.html

[5] TSPLIB: http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95

[6] http://ark.intel.com/

[7] Dorigo, M., Gambardella, L.M "Ant Colonies for the Traveling Salesman Problem", *BioSystems*, 43:73-81, 1997.

Closed-loop system for assisted strength exercising

Tomislav Pozaić*, Matija Varga*, Dominik Džaja* i Sara Žulj*
* University of Zagreb, Faculty of Electrical Engineering and Computing / Dept. of Electronic Systems and Information Processing, Zagreb, Croatia
pozaictomislav@gmail.com

Summary - Accurate assessment of the qualitative and quantitative aspects of physical activity is considered very important in order to achieve better results in rehabilitation or training. We have developed Wireless Body Area Network for continuous and personalized health monitoring of basic physiological parameters (ECG, breathing rate and body temperature) and added the functionality of mobility measurement and closed-loop assisted strength exercising. The results has shown that assisted exercising using on-line feedback is performed closer to prescribed instructions in time and intensity while self-assessment and exercising without on-line feedback vary much more showing a higher deviation from the prescribed instructions. Our system enables sophisticated tracking of the rehabilitation/exercise performances of patients with closed-loop quantitative and qualitative results in real-time.

I. INTRODUCTION

Broad range of therapy approaches are currently practiced in clinics. These therapies differ significanlty due to type of the disease and severity of the damage the acute disease caused to the patient. Therapists and physicians are confronted with difficult task of selecting optimal at home rehabilitation programs for their patients. Impossibility to do more frequent health monitoring is particularly difficult during rehabilitation. When the rehabilitation is not successful medical specialist can not explain what exactly has happened because they usually see patients when major complications have already occured. Moreover, dramatically aging population makes it absolutely necessary to develop systems that keep people out of hospitals, but with an appropriate medical care.

From previous systematic reviews and meta-analysis there is a consensus about positive effect of exercise training on health-related quality of life, mortality and hospital admissions [1]. However, the effects of exercise within self-management programs remain unclear or gave negative feedback.

According to the impending problems in healthcare, most important challenge for every system that offers assistance in rehabilitation is to give an accurate feedback on continuous physical activity for patients and their therapists [2]. Closed-loop exercising enables efficient continuous recording of physiological parameters and real-time assessment of the quantity and quality aspects of physical activity. Efficient closed-loop assessment is considered very important in order to achieve full potential of rehabilitation or training [3].

Questionnaires, surveys and diaries of self-control have disadvantages such as inaccurate perception and recall of information by the subject, unsuitable design and delay [4]. Important limitations for wider acceptance of the existing systems for continuos monitoring are: lack of system integration of invidual sensors, interference on a wireless communication channel shared by multiple devices, nonexistent support for massive data collection and lack of closed-loop support [5, 6]. Miniature, lighweight, ultra-low power, intelligent monitoring devices enabled continuous, accurate and closed-loop monitoring of physiological parameters of all kinds of patient without influencing their usual daily activities. On the other hand, sensor integration and well-structured patients information allows professional stuff more flexible treatment and more efficient data processing.

Wireless Body Area Network (WBAN) for continuous and personalized health monitoring of basic physiological parameters (ECG, breathing rate and body temperature) was developed at the Department of Electronic Systems and Information Processing at the Faculty of Electrical Engineering and Computing, University of Zagreb and we added the funcionality of mobility measurement and assisted exercising [7]. In this paper a part of this sytem which provides closed-loop assisted strength exercising is described. The goal of this system is to enhance the effects of rehabilitation and training in patients that suffer from different types of chronic diseases. It can also support early stage detection of abnormal conditions and thus prevent its serious consequences. Additionally, real-time critical health monitoring implemented in this system can provide professionals with life-important information and reduce the risk of further complications. System integrates the knowledge of experts in kinesiology, medicine and engineering.

II. SYSTEM ARCHITECTURE

OZIMS (Osobni Zdrastveni Informacijski i Mobilni Sustav), a personal healthcare information and mobile system, is a wearable systems with tremendous potential in proactive healthcare. System integration, wireless communication and sensor development enabled lightweight, easy-to-use, unobustrive, flexible, non-invasive, personalized health systems. Our system has the ability to communicate wirelessly with a personal

server and subsequently through the Internet with a emergency services or medical and research database.

OZIMS is a combination of non-invasive sensing and contextual features such as ECG, EMG, mobility, body temperature, voice and visual features. The whole system can be divided into four layers:

- sensor layer (data collection hardware with sensors for continuous measurement of mobility and physical activity)
- communication layer (communication hardware for data stream to a PC, tablet or cloud)
- processing layer (algorithms for real-time processing of phsyological and movement data)
- display layer (graphical user interface (GUI) for display of clinically-relevant information and assisted strength exercising on PC or tablet)

A. *Sensor Layer*

Very important role for sensor applications in the field of rehabilitation are advances in development of microelectromechanical systems (MEMS). MEMS sensors are ultra-low power and lightweight sensors that have been used in motor activity and other health status monitoring systems [8].

Sensor node has three inertial sensor for trunk position measurement, mobility and assessment of physical activity (accelerometer, gyroscope, magnetometer). It also has sensor for body temperature measurement, ECG and EMG amplifier, voice interface and communication module. The node is powered by one rechargeable lithium battery Nokia BL-5B (890 mAh) It can supply power to the node for up to 13 days without recharging. One or more sensor nodes form a Sensor Layer for every patient who is being monitored during rehabilitation or training.

For data acquisition closed-loop controlled strength exercising, sensor node is using embedded three axis accelerometer (ADXL345) and three axis gyroscope (L3G4200D). The axes of the coordination system of inertial sensors are not related to the coordination system of the Earth but rather, they match with coordination system with centre in sensor case. The sensitivity of the accelerometer is set to 4 mg/LSB. Although the frequency of recorded signals depends on sensor placement, it has been shown that in 95% cases frequency of human movement is under 10 Hz and amplitude is less than 6g (1g = 9.81 m/s^2) [9]. To capture all important signal frequency components and to fulfil ultra-low power demands, the sampling frequency of a 10-bit ADC is set to 100 Hz per channel.

Microcontroller LPC1347 controls all functions of the sensor node. FreeRTOS, a real time operating system, is implemented on the microcontroller. Whole algorithm is divided into tasks and low power modes are used to reduce the consumption. Detailed sensor node block diagram is shown in Figure 1.

B. *Communication Layer*

The main protocol used for communication between nodes and PC is flexyNet protocol based on TCP/IP

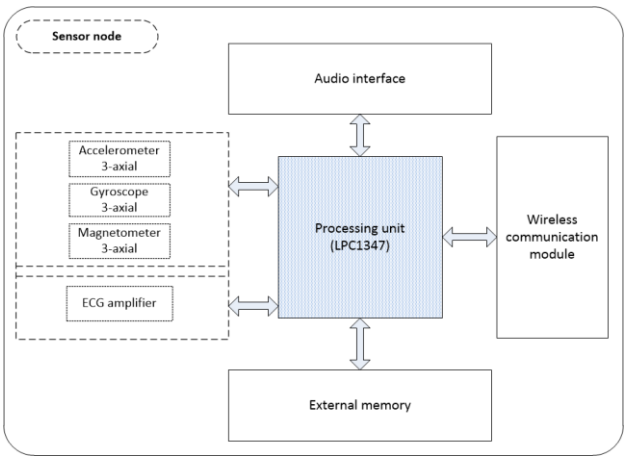


Figure 1. Sensor node block diagram

protocol [10]. Its main characteristics are high speed, reliability and broad range of different types of communication packets (23) which allow flexibility and efficiency. Central node in WBAN is communicating with the closest access point in the network (hospital, elderly home, working space, gym etc.) The node has also the option to record the data to MicroSD card in cases when the patient is out of reach of access point and GSM network.

Communication module is intended for local networks with low power consumption bellow 5 µA/byte. Module works in the ISM frequency band (2.4 GHz). To significantly reduce the power consumption, node sends only processed and carefully extracted data. In the case when power consumption is not important, it is possible to stream raw sensor's signal to the computer and than make further processing. To allow fully closed-loop exercising, two-way communication is enabled.

FlexyNet packets that are exchanged between sensor node and application are shown in Figure 2. At the beginning of exercising PC or tablet application sends packet to the sensor node that have information about patients ID, number of exercises and necessary information about every exercise (exercise ID, pause duration, dominant axis and number of repetition in each series). At the end of every exercise, the node sends

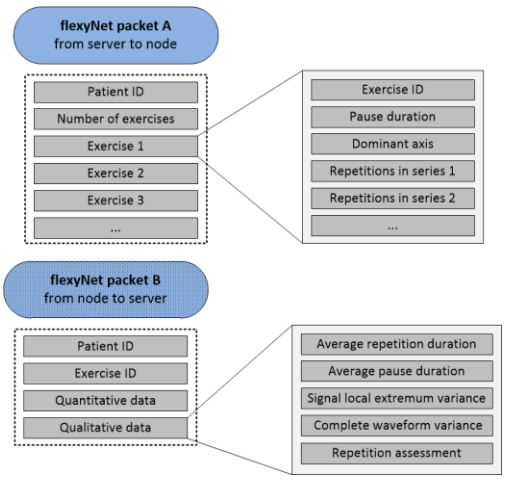


Figure 2. Communication packets between application and sensor node

packet to application with information about patient and exercise that is performed (exercise ID, quantitative index, repetition duratio, pause duratio, amplitude ratio, fluency of movements and described angles).

C. Processing Layer

The algorithm for processing of data collected from sensors was implemented on the microcontroller using cascaded integrator-comb filtering, derivating, signal thresholding and peak detecting. The algorithm is performed in real-time and no search-back is used. The result of processing is a feature vector with characteristic values for each type exercise that application user has to perform. Feature vector is a 5-dimensional vector of numeric features of exercise (minimum and maximum amplitude of signal, average duration of one repetition, average duration of pause and described angle during exercising). Feature vector is stored in database and it is used to describe qualitative and quantitative parameters of every strength exercise. Data that was gathered during processing is then transferred to the Display Layer.

Algorithm used in processing of signals obtained during exercising can be divided into three major blocks:

- Time Management,
- Quantity Management,
- Quality Management.

For every patient and for each exercise personalized parameters are entered into the exercise plan. Personalized parameters are stored in the database and depend on patient's age, therapy, health conditions etc. Once the patient has entered the assisted exercise program and accepted the exercise plan, at the beginning of each exercise personalized parameters are sent to patient's central sensor node using flexyNet protocol. Time Management block is used for controlling the time aspects of exercising. It is using timeouts to determine whether the patient is exercising as instructed, i.e. according to recommended timing. Thus, outputs of the Time Managment block participate in quality estimation of physical activity. Quantity Management block is used for repetition counting. Its main part is peak detection algorithm which detects each single repetition. Quality Management according to the parameters previously stored in database, then determines the quality of every repetition and it gives the final average score of whole physical activity during specific therapy.

D. Display Layer

The results of processing are sent to PC application where they are displayed in an intuitive and effective application. The application can, on demand, also analyze raw signal data in real-time from any combination of sensors, acting as a central sensor unit. Along with the display functions, application also provides key information that apply to patient's health and wellness such as: qualitative and quantitative score of physical activity, efficiency and intensity of physical activity, calories burnt, activity schedule and therapy information. It is developed in Microsoft Visual C# 2010 Express as an object oriented user interface so additional upgrades or integrations are easy to perform.

Application's Graphical User Interface (GUI) is particularly adapted for professional and non-professional users (athletes, patients). On user's PC, depending on the status of registered user, GUI is shown as a separate node which is communicating over the Central Server with the database and other nodes. Providing information about physical activity and therapy on any PC and whenever needed, it is very easy to further enhance rehabilitation and training output.

III. METHODS

A. Anthropometric measurement

In this experiment anthropometric measurements were conducted for better qualitative and quantitative signal analysis and to achieve the desired level of diagnostic certainty and resolution. Anthropometric measurements are important for better definition of sensor placement on the body, more accurate description and performance of given physical activity and more credible and repeatable statistic score of result of exercising. For every examinee, body mass index was calculated. Optimal position for placing the nodes was defined by analysis of previous papers [11]. Sensor node was placed on the wrist of dominant arm. Results of anthropometric measurement of examinees in this project are shown in Table I. Each individual value is shown along with mean and standard deviation rounded to one decimal place.

B. Experiment setup

One experiment was set for the investigation of user friendly aspects, intuitiveness, accuracy, assessment and usefulness of the closed-loop strength exercising application is implemented. A group of ten healthy examinees was formed, comprising of 8 male and 2

TABLE I. ANTHROPOMETRIC MEASUREMENTS

Examinee	**Male**		
	weight/kg	***hight/cm***	***age/years***
1	72	174	23
2	80	182	26
3	70	175	23
4	87	185	23
5	82	180	23
6	75	183	23
7	86	184	24
8	86	181	22
average	79,6	180,5	23,4
SD	6,7	4,0	1,2
Examinee	**Female**		
	weight/kg	***hight/cm***	***age/years***
1	64	172	26
2	62	168	21
average	63,0	170,0	23,5
SD	1,4	2,8	3,5

female examinees. Before exercising, every examinee signed the *Informed Consent for Physical Activity*.

Experiment was divided into two steps. In the first step, examinees were exercising following written instructions and without the assistance of our application. Signals were recorded on MicroSD card so further processing and feature extraction in OZIMS application would be possible. Every examinee got a plan of exercising in writing with detailed description of every exercise (number of repetition, number of series, recommended duration of each repetition, recommended duration of each pause, initial and ending extremity and/or body position). They had the assignment to count repetitions of prescribed movements and to estimate the time, effort and efficiency they have spent exercising. The aim was to be as close as possible to the default repetition and pause duration and to do exact number of prescribed repetitions and series. Examinees filled the form which was used for their self-assessment and which was compared with the recorded signals.

In the second step examinees had auditory and visual assistance during exercising and all information about every exercise have been displayed in application in real-time. Signal from the sensor node was streamed to application where it was further processed and where feature extraction was performed so examinee had real-time estimation of physical activity during exercising. Quantitative and qualitative parameters of recorded signals from first and second step were compared. After finishing the prescribed exercising program, the examinees overall satisfaction with application was tested by filling a short questionnaire. Results are shown in the next chapter of the paper.

In both phases of the experiment, all examinees were performing two different dumbbell exercises (Figure 3) which were chosen based on professional advice of medical staff [12]. Every exercise had three series of repetition with pause of 15 seconds after each series. In the first series the examinees had to perform six repetitions, in second they had to perform five repetitions and in the third four repetitions. The duration of each single repetition was set to 2 seconds. Exercises were performed with dumbbells having a mass of 1 kg. The target group of muscles were arm muscles.

We hypothesize that assisted exercising using on-line feedback through OZIMS application will be performed closer to prescribed instructions in time and intensity while exercising performance results inter and intra-examinees without on-line feedback as well as the results of the following self-assessment will vary much more showing a higher deviation from the prescribed instructions.

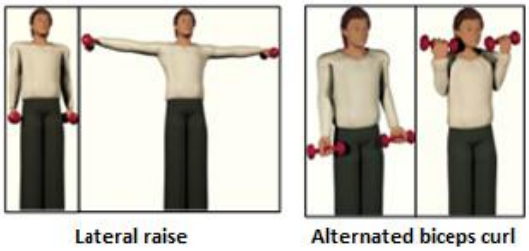


Figure 3. Dumbbell exercises

C. Signal processing

In both experiment stages signals where processed within a desktop application. In the first, self-assessment stage of this experiment, signals were recorded on the MicroSD card and later fed to desktop application with the aid of data generator form. The form was used to simulate real-time signal. In the second stage signals were processed in real-time within the desktop application. Signal processing is applied on one axis off acceleration signal with the highest amplitudes (dominant axis).

The first stage of signal processing was filtering and decimation. The most convenient solution to this problem was designing *cascaded integrator-comb* filter. We have implemented 8 sections of CIC decimator with the decimation factor of 5, thus sampling frequency at the output was 20 Hz.

The second stage of signal processing involved edge detection. Each exercise has predefined parameters that are loaded from the database. One of the parameters are derivation thresholds, as well as signal thresholds at which derivations should occur. Two pairs of derivation and signal thresholds are used, one for the repetition start and one for the repetition end. To eliminate false positive edge detections an approach similar to debouncing was used. If initial edge was detected, the algorithm starts searching for local extreme until the final edge is detected. The initial and the final edge denote single repetition duration. Another local extreme is searched between the final edge and the initial edge of the next repetition. Local extrema are used to qualitatively describe the dominant axis waveform. One repetition is characterized by:

- the two local extrema,
- repetition duration,
- 2 pairs of derivation and signal thresholds.

Depending on the workload assigned to the patient and his physical status, different tolerances are applied on the signal and derivation thresholds. This feature makes an algorithm less rigid and more suitable for different types of users. For example, if one person is at the beginning of his rehabilitation and can not follow the

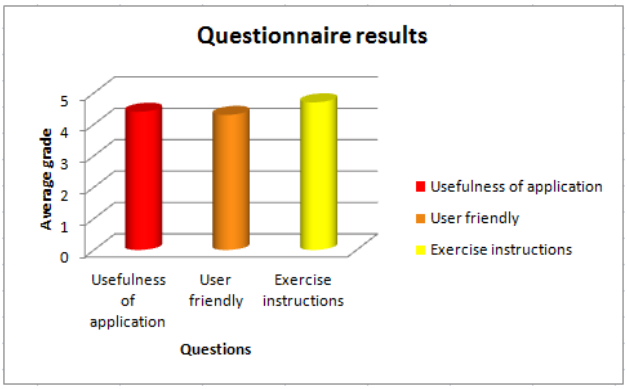


Figure 4. Average grades on questions about OZIMS system

prescribed pace of exercising, his fitness level will be saved to database as low. As a consequence, high tolerances will be applied in the second stage of signal processing. At the beginning of one's rehabilitation, even finishing the whole exercise plan is satisfactory rather than following strict rules of exercising. On the other hand, if a person is at the end of his/her rehabilitation, his/her fitness level are expected to have reached a medium or high level, hence the tolerances will be narrowed. Repetitions that are not within the tolerances will be marked as a poor performance.

IV. RESULTS

A detailed questionnaire regarding exercising with the assistance of the developed application consisted of 3 major questions:

- usefulness of the OZIMS system for the assisted strength exercising,
- user friendly and intuitive aspects of the OZIMS system,
- satisfaction with the provided instructions for exercising.

Responses to each question were on 6-point scale where a score 0 indicated that examinee could not finish the exercise, 1 indicated the least satisfaction, while 5 indicated the total satisfaction with the application. The information obtained from these questionnaires is shown in Figure 4 as an average value of all scores rounded to one decimal place. Results of the questionnaire have shown average score of 4.4 for the usefulness of the OZIMS systems for the assisted strength exercising, average score of 4.3 for the user friendly and intuitive aspects of the OZIMS system and 4.7 for the satisfaction with the provided instructions for exercising. Number of repetitions for first exercise in self-assessment is estimated to (16.2 ± 3.1), number of repetitions for second exercise is estimated to (15.5 ± 1.6), duration of each repetition in self-assessment is estimated to (2.0 ± 0.8) seconds and duration of each pause is estimated to (13.5 ± 2.1) seconds.

Using the developed algorithm in application, we processed signals and compared average repetition duration of each type of exercise with and without the support of application for assisted exercising. Repetition duration in first phase of experiment for first exercise is (1.6 ± 0.9) seconds, while for second exercise is (1.5 ± 1.0) seconds. In the second phase of experiment, average first exercise repetition duration is (1.9 ± 0.3) seconds, while average second exercise repetition duration is (1.9 ± 0.2) seconds.

Quality of each repetition is also calculated within C#. It is measured on scale from 0 to 10, where 0 designates the case when the quality mark could not be given. Grade 10 is given if both local extrema are within tolerances. Average quality of each repetition in first phase of experiment is (6.0 ± 1.6), while average quality of each repetition in second phase of experiment is (8.6 ± 0.5).

V. CONCLUSION

The results of self-assessment have shown that examinee's quantitative assessment of exercising vary from prescribed exercising. Self-assessment depends on various factors such as motivation, concentration and trained level of examinee, understanding of given instructions, successfulness in taking starting position and performing the exercise to the instructed ending position. Furthermore, examinees in the first phase of experiment showed greater deviation of repetition and qualitative performance of exercises than in the second phase.

With the results of questionnaires, we have shown that our system is well accepted within the group of examinees and since it ensures quantitative and qualitative assistance during strength exercising it might have the positive effect on rehabilitation and training in larger groups as well.

Streaming information between sensor node and desktop or mobile application, real-time feature extraction and context classification using OZIMS system that integrates commodity hardware and attached application for assisted exercising enables the real-time classification of performances with feedback without the need for other infrastructure, invasive methods and obstruction of patient's daily life. Since in the rehabilitation is not only important to simply exercise, but to exercise exactly as prescribed and as much as prescribed, it is obvious that OZIMS application can further enhance the process of rehabilitation. Streaming processed data from the sensor node to Central Server where they are easy accessible to professional and end users for further review, we introduced closed loop for continuous monitoring of physical activity of patients in real-time and enabled sophisticated tracking the rehabilitation/exercise performance to patients and their therapists.

ACKNOWLEDGMENT

The authors would like to acknowledge the contribution of L. Celić in the design and development of Wireless Body Area Network and flexyNet protocol.

LITERATURE

[1] G. Ninot et al., "Cost-saving effect of supervised exercise associated to COPD self-management education program", Respiratory Medicine, vol. 3, pp. 377-385, 2011.

[2] H. Viswanathan, C. Baozhi, D. Pompili, "Research challenges in computation, communication, and context awareness for ubiquitous healthcare", IEEE Communication Magazine, vol. 50, pp. 92-99, 2012.

[3] B. Najafi, K. Aminian, A. Paraschiv-Ionescu, F. Loew, C. J. Bula, P. Robert, "Ambulatory system for human motion analysis using a kinematic sensor: Monitoring of Daily physical activity in the elderly", IEEE Transaction on Biomedical Engineering, vol. 50, pp. 711-723, 2003.

[4] F. Pitta, T. Troosters, V. S. Probst, M. A. Spruit, M. Decramer, R. Gosselnik, "Quantifying physical activity in daily life with questionnaires and motion sensors in COPD", European Respiratory Journal, vol 27, pp. 1040-1055, 2006.

[5] E. Jovanov, A. Milenkovic, C. Otto, P. de Groen, "A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation", Journal of NeuroEngineering and Rehabilitation, vol. 2, pp. 6-16, 2005.

[6] M. Sung et al., "Wearable feedback systems for rehabilitation", Journal of NeuroEngineering and Rehabilitation, vol. 2, pp. 17-29, 2005.

[7] L. Celic, M. Varga, T. Pozaic, D. Dzaja, S. Zulj, R. Magjarevic, "WBAN for physical activity monitoring in health care and wellness", World Congress on Medical Physics and Biomedical Engineering, Beijing, China, 2012.

[8] S. Patel, H. Park, P. Bonato, L. Chan, M. Rodgers, "A review of wearable sensors and systems with application in rehabilitation", Journal of NeuroEngineering and Rehabilitation, vol. 9, pp. 21-38, 2012.

[9] A. G. Bonomi, "Physical activity recognition using a wearable accelerometer: New perspectives for energy expenditure assessment and health promotion", Eindhoven: Datawyse / Universitaire Pers Maastricht, 2010.

[10] L. Celic, D. Trogrlic, I. Paladin, "Integration of measurement devices supporting diabetic patients into a remote care system", IFMBE Proc., vol. 37, pp. 39-42, 2011.

[11] S. E. Crouter, K. G. Clowers, D. R. Bassett, "A nove method for using accelerometer data to predict energy expenditure", Journal of Applied Physiology, vol. 100, pp. 1324-1331, 2006.

[12] P. Manocchia, "Anatomy of Exercise: a trainer's guide to your workout", 1st ed., Mozaik knjiga, 2011.

Najbolji Plan za posao!

Nove poslovne tarife!

Sony Xperia™ Z
već od 498kn
u tarifi Plan za posao

SONY
make.believe

Sony Xperia™ J
već od 8kn
u tarifi Plan za posao

Izaberite svoj **Plan za posao**, nove tarife koje jedine u cijeni mjesečne naknade, pored neograničenih razgovora, interneta i SMS-a uključuju:

- **primanje i slanje „velikih“ mailova i kada ste van ureda, uz 4G mobilni internet**
- **dodatne minute za bezbrižne međunarodne i roaming razgovore**

Izaberite svoj **Plan za posao** i super telefone na **hrvatskitelekom.hr**, **u T-Centrima** i na **0800 9100**.

Živjeti zajedno

Hrvatski Telekom

sponsors

golden sponsors

silver sponsor

bronze sponsors

sponsors

MIPRO
Kružna 8/II, HR-51000 Rijeka, Croatia

phone: +385 51 423 984
mipro@mipro.hr • www.mipro.hr

ISBN 978-953-233-074-8